



RAMP deployment options

Copyright © 2010 VMT
All rights reserved

This document contains Proprietary and Confidential information of VMT, and is protected by copyright, trade secret and other laws. Its receipt or possession does not convey any rights to reproduce, disclose its contents, or to manufacture, use or sell anything it may describe. Reproduction, disclosure or use without specific written authorization of VMT is strictly prohibited.

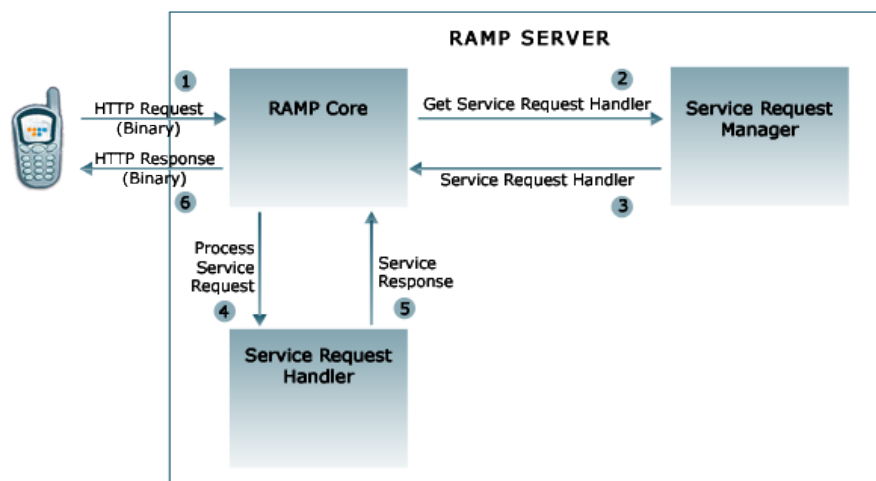
Table of Contents

RampDeploymentOptions.....	1
Overview.....	1
A closer look at capacity.....	2
RAMP server approach.....	2
Scale-up approach.....	2
Scale-out approach.....	2
Managing environment factors.....	3
Monitoring and transparency.....	4
Conclusion.....	4

RampDeploymentOptions

Overview

This paper will outline the options that are available to deploy the RAMP server into a live production environment. The RAMP server can be viewed as the gateway between the various mobile devices (and, in the future, desktop) and the back-end enterprise services. The interaction between the RAMP platform and the back-end enterprise services can be illustrated as follows:



Over the last several years end customers have become less and less tolerant of service downtime. Service downtime can lead to a loss of revenue and customer confidence. It is therefore critical that any data center service be highly available and able to increase its capacity (i.e. can scale) as load increases.

There are fundamentally two approaches that have emerged in the industry to increase capacity as load increases, scale-up (vertical scaling) and scale-out (horizontal scaling).

A scale-up solution achieves additional capacity by having fewer bigger machines that are given more processing power as load increases. An example of this would be having a single IBM System z based server and then adding CPUs and memory to the machine as load increases. A scale-out solution achieves additional capacity by having many smaller servers and adding more servers as the load increases. An example of a scale-out solution would be to have two or more Dell rack mounted servers and to then add additional rack mounted servers as load increases. A scale-out solution is usually implemented in a cluster within which each server (or node) shares data with all the other nodes.

A system should also provide high availability in addition to being able to increase capacity as load increases. High availability is the ability for a system to continue to provide a service even in the scenario where certain system components have failed. In a scale-up approach this is provided by the underlying hardware. In a scale-out approach it is provided by each server being able to take over the workload of another failed server in the cluster.

The RAMP server is able to provide both approaches depending on the preference of the client's operations team. In addition to this the RAMP server is a Java EE (Java enterprise edition) server application and is therefore supported by a wide host of hardware vendors and operating systems.

A closer look at capacity

Before we take a closer look at how the RAMP server provides a highly scalable solution it is necessary to define what we mean by capacity. Capacity is a term that has a slightly different meaning to many people and consists of several components.

In order to define capacity we need to define performance and throughput.

Performance measures how fast a system is able to process a single transaction (or workload). End customers are only concerned with performance since a single slow transaction, from an end customer's point of view, means a slow system.

Throughput describes the number of transactions a system can process in a given time span.

Finally, the maximum throughput a system can sustain, for a given workload, while maintaining a desired response time for each individual transaction is its capacity.

Capacity is affected by various environment factors namely network latency, network bandwidth, network reliability, integration point latency and of course the available hardware.

The RAMP server maximizes capacity given the available hardware resources and environment.

RAMP server approach

The RAMP server has been designed to be able to take advantage of both a scale-up and a scale-out approach to capacity and high availability.

Scale-up approach

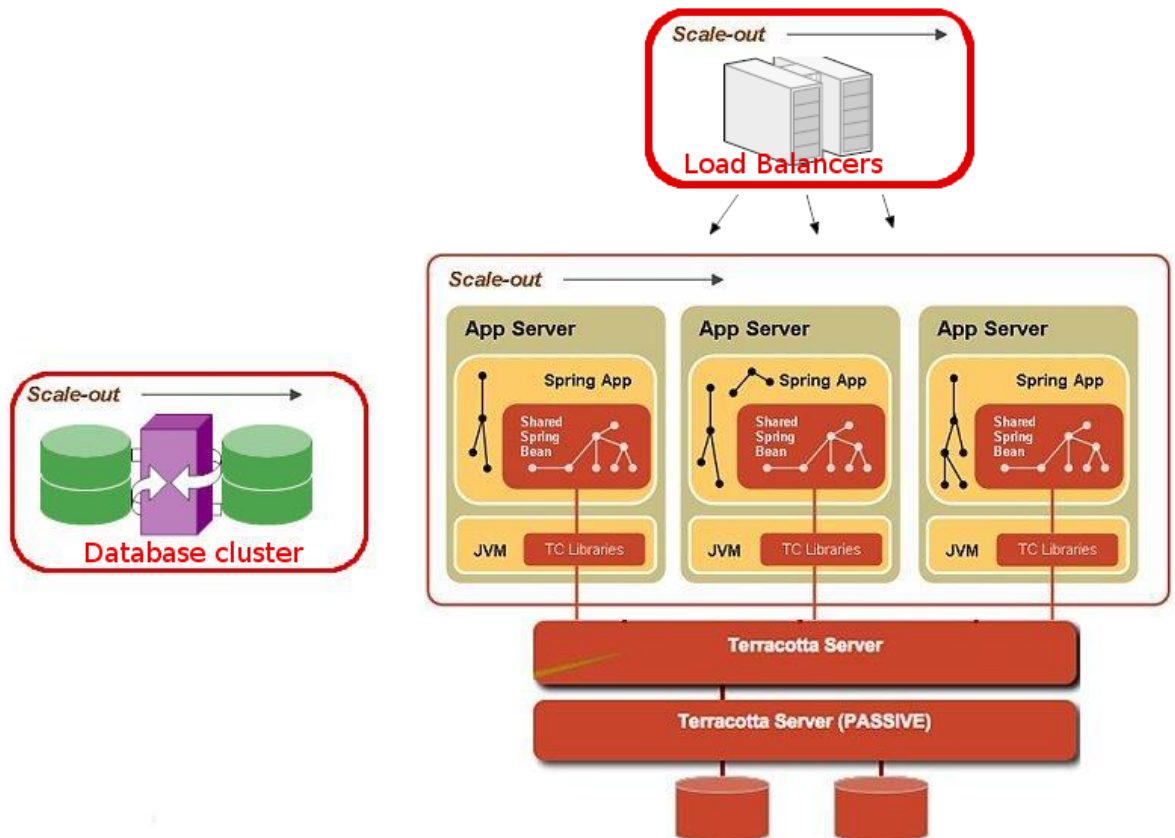
The RAMP server has been designed to be able to run efficiently within a multi-processor hardware setup. This is critical when a scale-up approach is followed where it is entirely possible to have a single large server with anywhere between 1-64 CPUs. Developing parallel server applications that can take advantage of multi-processor servers is a highly skilled process and the RAMP server was designed to run in such an environment.

High availability is provided by the hardware vendor that provides redundancy and performance guarantees for the underlying hardware components.

Scale-out approach

The ramp server has also been designed to be able to run in a scale-out approach. A set of RAMP servers can be fronted by a set of load balancers so that when the load increases it is simply a case of adding additional server nodes to increase capacity. The RAMP server scale-out implementation can scale to dozens of machines. An example of a deployed system can be illustrated as follows:

RAMP deployment options



The load balancing servers are the first point of entry to the deployed system for a connecting client. They are responsible for delegating requests to the application servers. The load balancers will monitor the load on the application server nodes and if one or more nodes are under too high a load they will redirect work to nodes that are under a lesser work load. The load balancers can also detect if an application node becomes offline and redirect work to other nodes. The load balancers should have sticky sessions enabled.

The state clustering servers ensure that the application servers have consistent access to client session information so that should one application server fail, another application server can seamlessly take over its work load.

Managing environment factors

The RAMP server and the overall RAMP platform mitigates the impact of various environment factors on the performance of the system. Environment factors that have been taken into consideration include:

- Network bandwidth
- Network latency
- Network reliability

The RAMP server communicates with the RAMP client via a binary protocol that greatly reduces the bandwidth requirements. The reduced bandwidth requirements also leads to better request latency and request reliability due to smaller packets being sent over the wire. The RAMP client and RAMP server also place additional transaction reliability on top of the network connection.

Another method to mitigate these factors is to have a data center pairing with the network operators and then to have traffic intended for the service to be routed via that pairing instead of through the general public pipe.

Monitoring and transparency

Regardless of the approach chosen (scale-up or scale-out) it is critical that a solution provides transparency and exposes monitoring metrics to the operations team. The RAMP platform provides extensive monitoring metrics such as:

- CPU usage
- Memory usage
- Network metrics
- Java EE JVM application metrics (Garbage collection, active threads, available memory etc.)
- RAMP server metrics (Requests served, integration point availability, integration point latency etc.)

This allows the operations team to have a comprehensive view of the system in order to identify and mitigate potential problems. The RAMP platform monitoring is also able to automatically monitor metrics and notify the operations team(via email or sms) should certain metrics exceed specified limits. The monitoring platform also provides a dashboard from where current and historic monitoring data can be viewed.

Conclusion

The RAMP platform in its totality was designed to be an enterprise class extension of a company's existing data center services. As such it provides all the functionality expected from a modern enterprise platform.