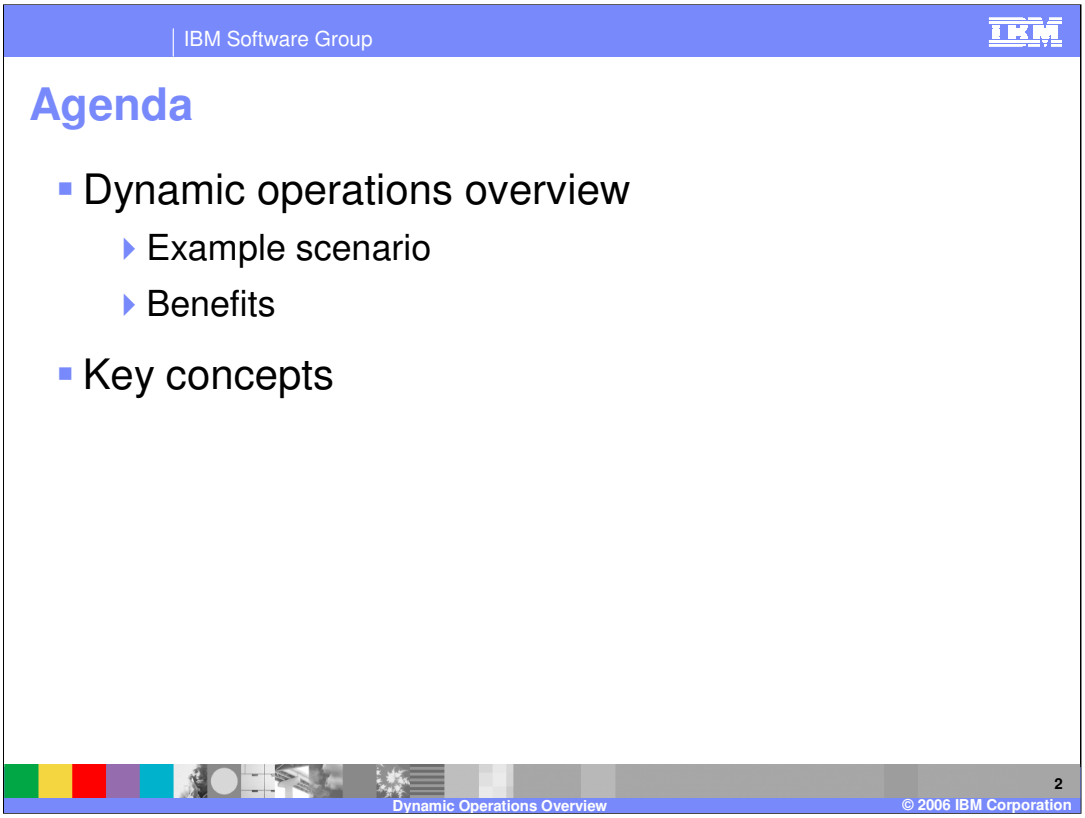
This presentation will give an overview of the dynamic operations features in WebSphere Extended Deployment for z/OS Version 6.0.1.

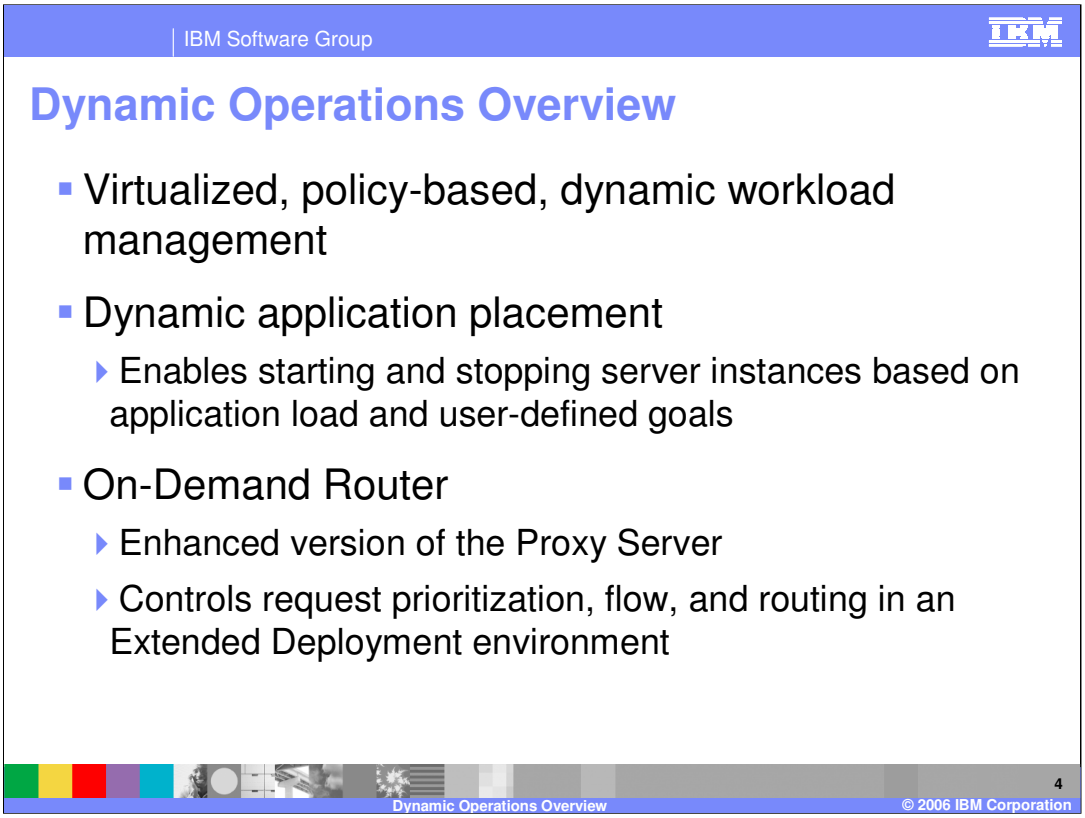


This presentation will begin by illustrating the main ideas behind dynamic operations using a simple example scenario that highlights the benefits of a dynamic WebSphere Extended Deployment environment, and introduce some of the key concepts involved in creating a dynamic operations-based WebSphere Extended Deployment environment.

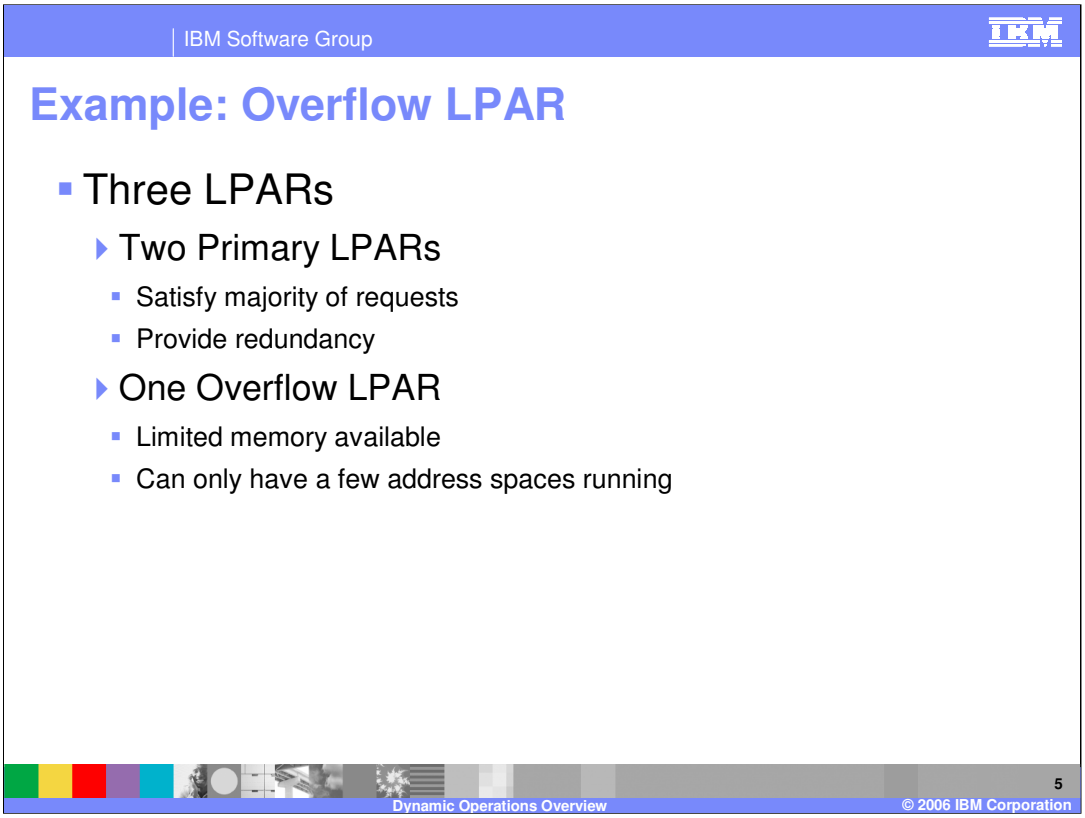
Section

Overview

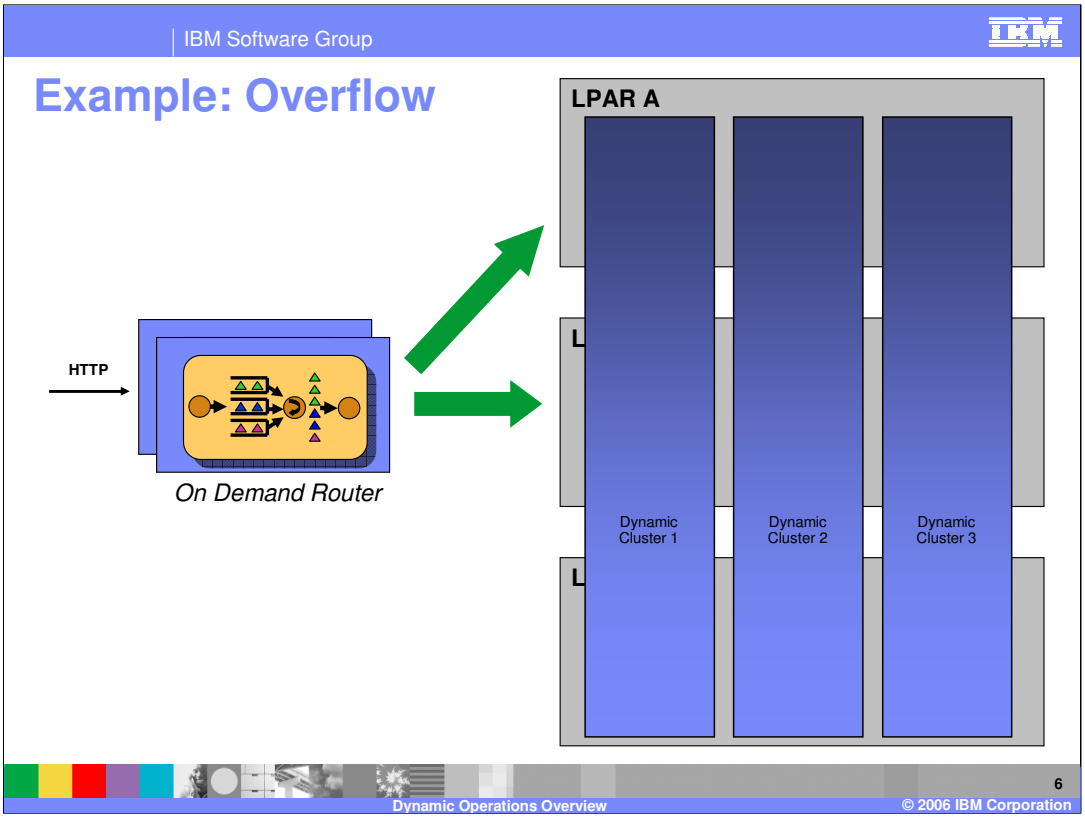
This section will give an overview of dynamic operations.



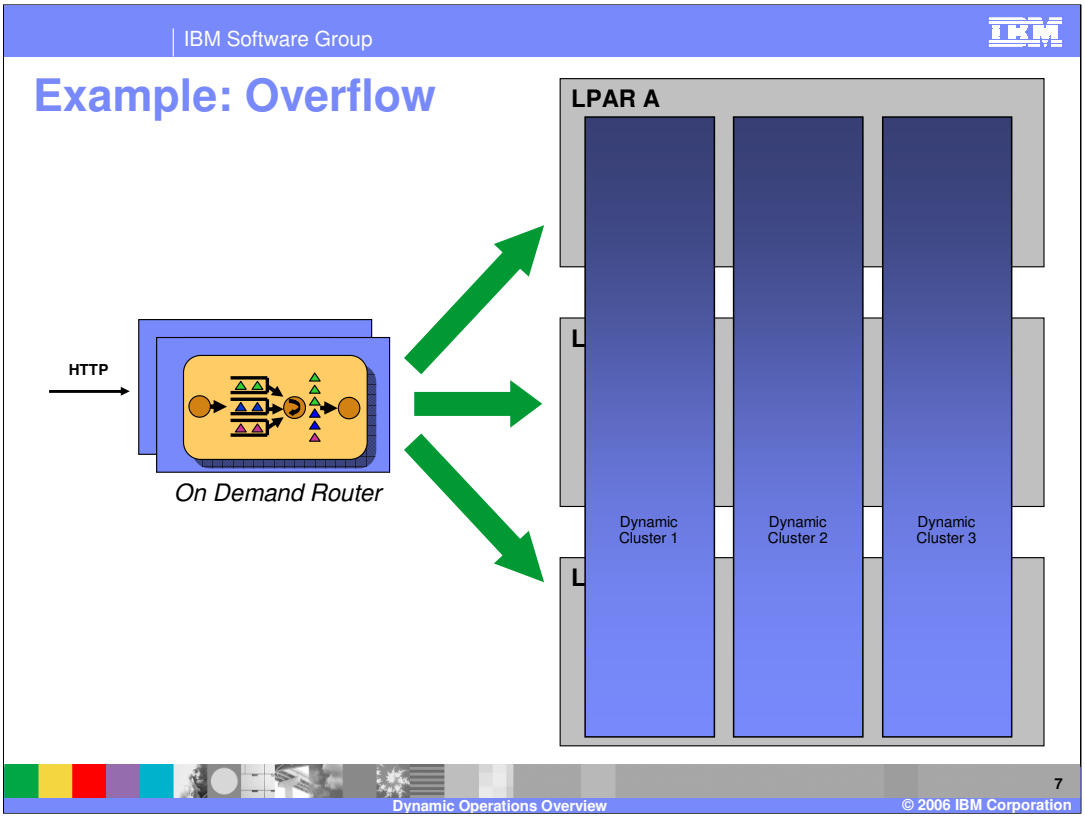
The dynamic operations features of WebSphere Extended Deployment give you the capability to build a dynamic, virtualized, goal-oriented environment for workload management. The two major features that enable these capabilities are dynamic application placement and the On-Demand Router. Dynamic application placement enables starting and stopping additional server instances to accommodate changes in load, balancing processing power among your applications to best meet your defined performance goals. The On-Demand Router is an intelligent HTTP proxy server that manages request prioritization, flow control and dynamic routing of requests to your application servers.



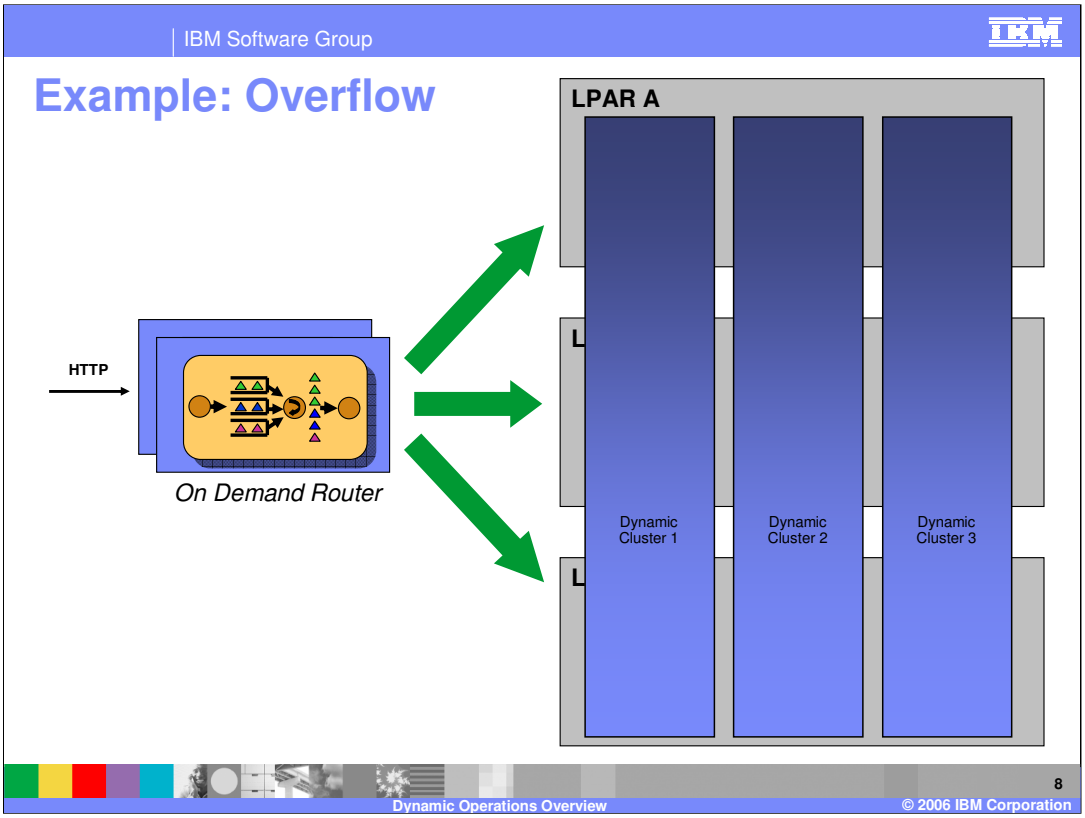
In this example, your organization has many departments and they each have several applications they run on their departments WebSphere server in a central complex. For administrative purposes these departments do not share servers but rather maintain their own through a central IT group. As the overall administrator, you have three LPARs configured. “LPAR A”, “LPAR B” and “LPAR C”. A and B are the primary LPARs for satisfying requests for all departments. On occasion, one or more servers exceed the capacity available on LPARs A and B. To handle this overflow you have configured “LPAR C” as a spare. Therefore “LPAR C” can only have a few servers (controller/servant region pairs) running at any one time.



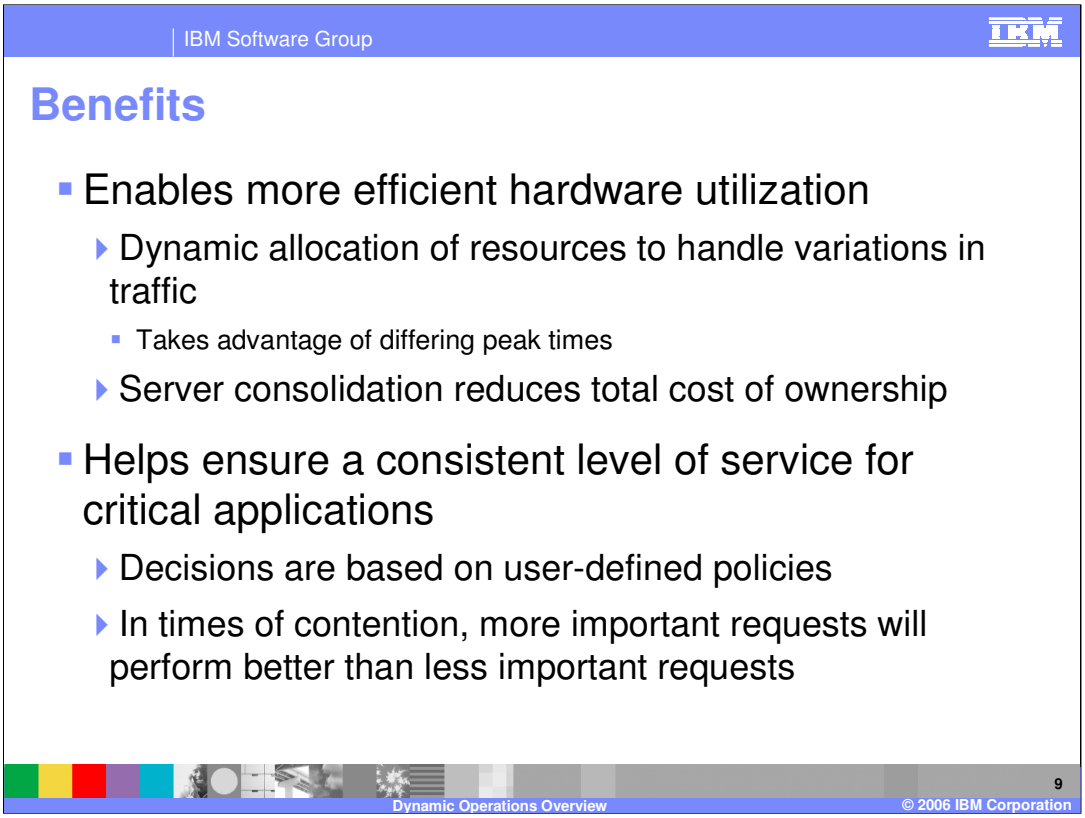
This picture represents nominal configuration. When the On Demand Router startup, it detect that LPAR C does not have as much memory or CP allocated to it as the other two LPARs. Here LPAR A and LPAR B have servers running for all three departments and these server are handling the entire offered workload. As the workload for Department Two begins to increase and the running servers can no longer satisfy the pre-configured goals, the On Demand Router determines that LPAR C has spare bandwidth available.



Here the Automatic Placement Facility in the On Demand Router has started a server for Department Two on LPAR C and is routing work to the new server in addition to the two existing servers. Later on the offered load for the Department Two server decreases and the offered load for Department One increases. The On Demand Router will detect the change and reconfigure the dynamic clusters.



Here, the third server for Department Two has been stopped and a new server for Department One has been started on LPAR C. This dynamically accommodates the decrease in the offered load for Department Two and the increase in the offered load for Department One. Now assume that the offered load for Department Three increases and the two Department Three servers can not satisfy the configured goals. Now the On Demand Router examines the LPARs and determines that LPAR C does not have enough memory resource to start another server, so it does not start another server. The dynamic behavior of the On Demand Router is able to maximize the resources available.



By taking advantage of differing peak times in application load, hardware can be utilized much more efficiently in a dynamic operations environment, resulting in lower overall hardware costs. A dynamic operations environment also helps ensure a consistent quality of service for your applications. WebSphere Extended Deployment can allocate hardware resources to help ensure that applications meet their defined goals. This allows the most important requests to perform better than less important requests when there is contention for resources.

Section

Key Concepts

This section will cover the key concepts and components of a dynamic operations environment.

Node Groups

- A Node Group defines a shareable pool of hardware resources (nodes)
  - These nodes have the same capabilities
    - Network, database access, and others.
- Node Groups are a boundary for clusters
  - That is, all members of a cluster must be within the same Node Group
- Node Groups can overlap
  - New in Extended Deployment 6.0
  - In Extended Deployment 5.1, Node Group membership was exclusive

A Node Group is a shareable pool of hardware resources, or nodes. All members of a cluster must be contained in the same Node Group. It is important that all members of a Node Group have the same capabilities, such as database drivers or access to network resources, so that they have the ability to run the same set of applications. In WebSphere Extended Deployment version 6, a node can be a member of more than one Node Group, unlike version 5.1, in which node group membership was exclusive.

Dynamic Clusters

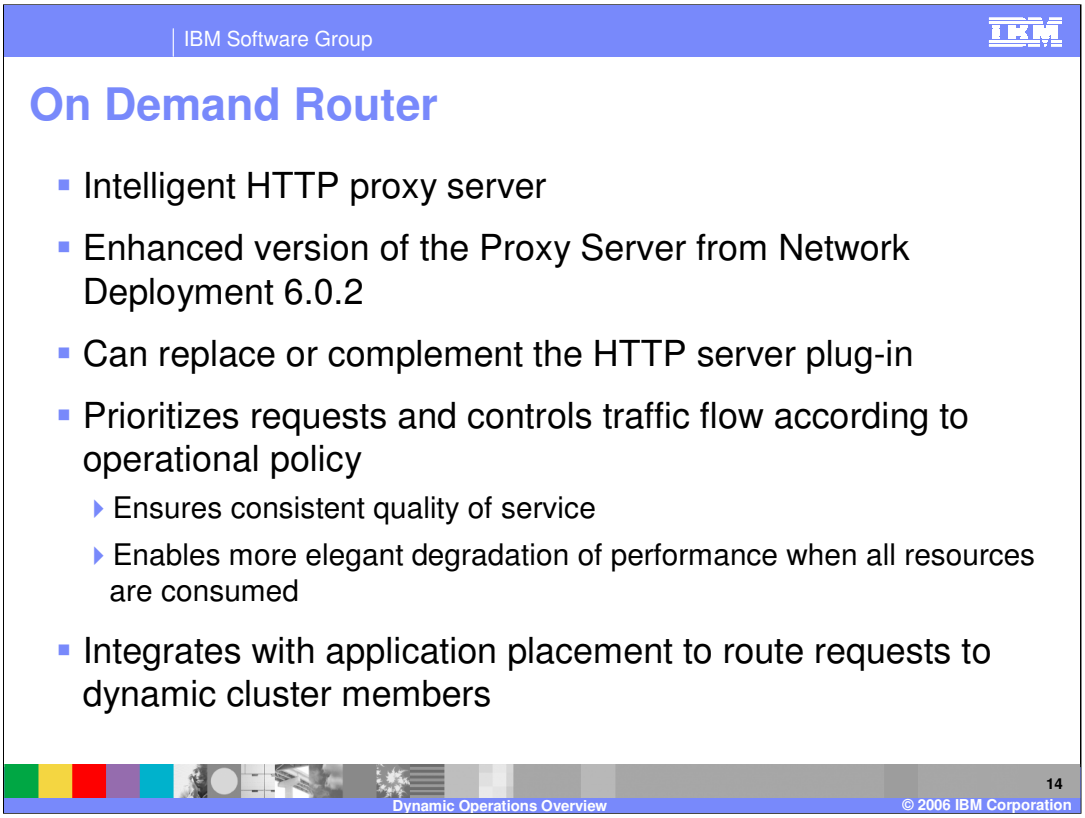
- A dynamic group of servers to which applications can be deployed
- Similar to a static Cluster, but can be resized dynamically within a bounding Node Group
- Each Dynamic Cluster has a template that defines settings for member servers
  - Modifying this template affects all servers in the Dynamic Cluster
  - Static Clusters require changing each individual server to achieve the same effect

A Dynamic Cluster is similar to the familiar concept of a ‘cluster’ from WebSphere Application Server, but can be resized dynamically within a Node Group. As demand for applications running on a Dynamic Cluster increases or decreases, instances of that Dynamic Cluster can be started or stopped on nodes within the bounding Node Group to accommodate the changes in load. Each node in the bounding Node Group has a configured instance of the Dynamic Cluster that is ready to be started dynamically when needed. These server instances are configured based on a server template that defines the configuration for all of the cluster members. This template is used as a single point of configuration for all members of the Dynamic Cluster.

Dynamic Cluster Enhancements (new in 6.0)

- Dynamic Clusters can now be 'vertically stacked'
  - More than one member on the same node
  - Extended Deployment 5.1 was limited to one member per node
  - 'Stacking Number'
    - User-defined number that defines how many application server instances are required to exercise the full power of a given node
- Lazy Start
  - A Dynamic Cluster can be configured to have no active instances when the application is idle for a period of time
    - Allows application to be stopped if memory is needed by other applications
    - Useful for rarely used applications where users can afford to wait for startup
  - In Extended Deployment 5.1 at least one instance was always active

In version 6, Dynamic Clusters have gained some new configuration options. 'Vertical stacking' allows more than one instance of a Dynamic Cluster to run on the same node if multiple Java™ Virtual Machines are required to fully utilize the processing power of your machine. This setting may come in handy when using hardware with many processors or when your application is heavily synchronized. Lazy start is a setting that enables you to configure a Dynamic Cluster that will have no active instances after a defined period of inactivity. This setting enables you to keep rarely-used applications installed and ready for use without consuming any server resources. If a request is made for an application with no running instances, the user will have to wait for the application server to start.

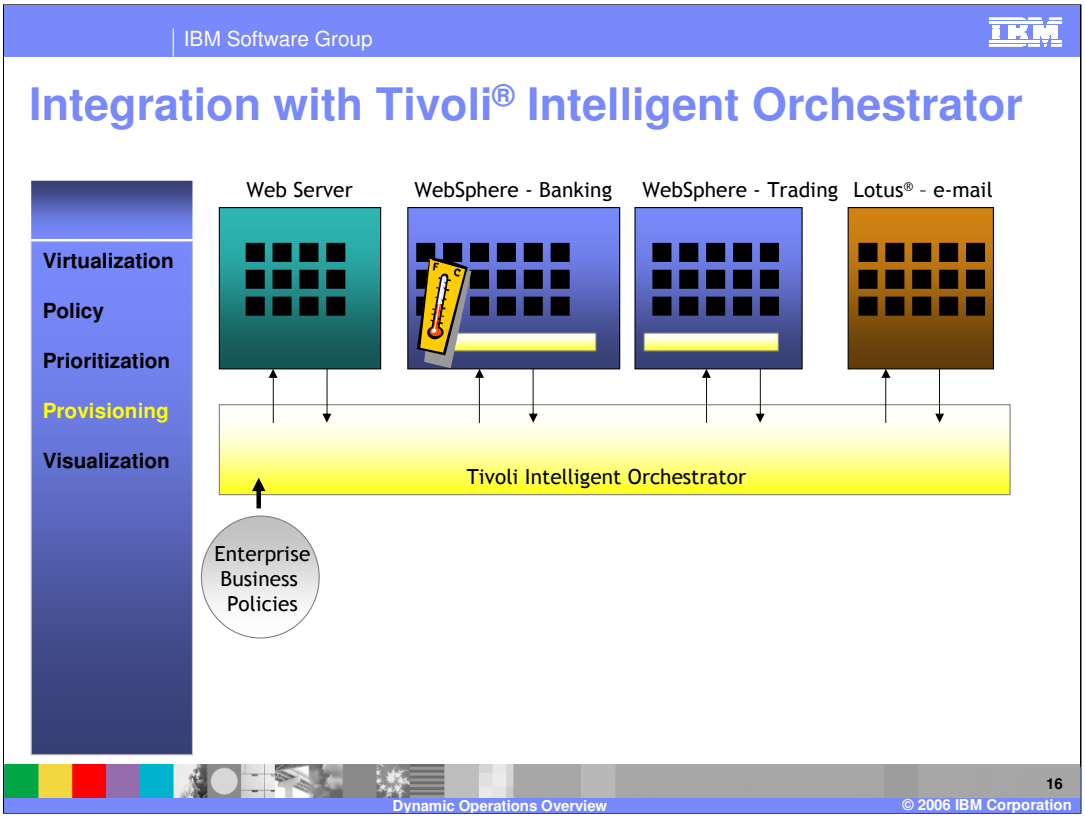


The On-Demand Router is an intelligent HTTP proxy server that is provided with WebSphere Extended Deployment. It is the point of entry into a WebSphere Extended Deployment environment, and is responsible for request prioritization, flow control, and distributing requests to application servers. It can momentarily queue requests for less important applications in order to allow requests for more important applications to be handled more quickly. It is aware of the current location of dynamic cluster instances, so that requests can be routed to the correct endpoint. The On Demand Router can also dynamically adjust the amount of traffic sent to each individual server instance based on processor utilization and response times. These and other advanced features distinguish the On Demand Router from the HTTP server plug-in, and give the On Demand Router the ability to ensure a more consistent quality of service for your enterprise applications. It can be used in place of, or in concert with the HTTP server plug-in, depending on your needs.

Application placement controller

- Decides how many instances of each Dynamic Cluster should run, and where they should run
- Determines the available capacity (memory, processor) of each node
  - Aware of the capacity in use by other processes and subtracts from available capacity (new in Extended Deployment 6.0)
- One Application Placement Controller per cell
  - Extended Deployment 5.1 had one Application Placement Controller per node group
  - This change accommodates overlapping node groups

The Application Placement Controller is the component that decides how many instances of each Dynamic Cluster should be running to most effectively handle the current amount of traffic. It determines the available processor and memory capacity of each node, including resources that are in use by other processes. It uses this information to determine the optimal placement of each application to best meet your defined performance goals. Each cell has one Application Placement Controller, which is a highly available singleton service that runs inside one of the Node Agents within the cell.



WebSphere Extended Deployment can be integrated with Tivoli Intelligent Orchestrator for server provisioning in a larger, heterogeneous environment. Tivoli Intelligent Orchestrator is a product that provides the capability to dynamically allocate hardware resources across products within an enterprise. Business policies dictate the allocation of enterprise-wide server resources, which can be reallocated based on need. For example, if a WebSphere Extended Deployment cell has exhausted all of the resources available to it, servers that were previously part of an underutilized environment can be reprovisioned as WebSphere Extended Deployment servers. It can then be added into the WebSphere Extended Deployment cell to begin hosting Dynamic Cluster instances. WebSphere Extended Deployment provides the required hooks for operating within a Tivoli Intelligent Orchestrator environment. This capability now works with overlapping node groups, which was not possible in WebSphere Extended Deployment V5.1.

Summary

- WebSphere Extended Deployment creates a dynamic, virtualized, goal-based environment for application hosting
  - The environment can adapt to varying traffic levels and allocate server resources as necessary
- Dynamic Clusters are similar to Clusters, but can be dynamically resized within a Node Group
  - The Application Placement Controller decides when this should take place

In summary, WebSphere Extended Deployment enables you to create a dynamic, virtualized, goal-based environment for hosting your enterprise applications. This environment can adapt to varying traffic levels and allocate server resources as necessary to help meet the performance goals of your applications. Applications are installed to Dynamic Clusters, which can be dynamically resized within a virtual pool of resources, called a Node Group. The Application Placement Controller is the component that is responsible for making placement decisions based on current load levels and user-defined performance goals.

Summary (cont.)

- The On Demand Router is a new component provided by WebSphere Extended Deployment
  - On Demand Router is an intelligent HTTP proxy server
  - Performs request classification, flow control, and dynamic workload management

Finally, the On-Demand Router is the point of entry for HTTP requests into a dynamic operations environment. It performs request classification based on user-defined rules, ensures that more important requests flow through to the back end more quickly than less important requests, and dynamically routes requests to dynamic cluster members.

Trademarks, copyrights, and disclaimers

The following terms are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both:

IBM	CICS	IMS	MQSeries	Tivoli
IBM(logo)	Cloudscape	Informix	OS/390	WebSphere
e(logo)business	DB2	iSeries	OS/400	xSeries
AIX	DB2 Universal Database	Lotus	pSeries	zSeries

Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are registered trademarks of Microsoft Corporation in the United States, other countries, or both.

Intel, ActionMedia, LANDesk, MMX, Pentium and ProShare are trademarks of Intel Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Linux is a registered trademark of Linus Torvalds.

Other company, product and service names may be trademarks or service marks of others.

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This document could include technical inaccuracies or typographical errors. IBM may make improvements and/or changes in the product(s) and/or program(s) described herein at any time without notice. Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business. Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program, that does not infringe IBM's intellectual property rights, may be used instead.

Information is provided "AS IS" without warranty of any kind. THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IBM EXPRESSLY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT. IBM shall have no responsibility to update this information. IBM products are warranted, if at all, according to the terms and conditions of the agreements (e.g., IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. IBM makes no representations or warranties, express or implied, regarding non-IBM products and services.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.

Note to U.S. Government Users - Documentation related to restricted rights-Use, duplication or disclosure is subject to restrictions set forth in GSA ADP Schedule Contract and IBM Corp.