# Overview of the z/OS Load Balancing Advisor: Making External IP Load Balancers Sysplex Aware

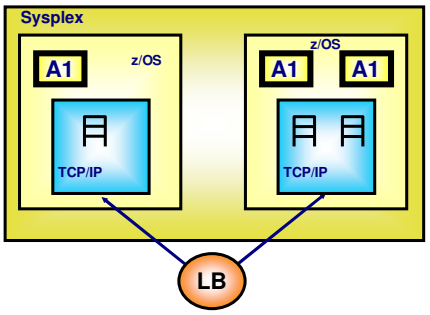
Enterprise Networking and Transformation Solutions, Raleigh



## Agenda

- Sysplex IP workload balancing overview
- Service/Application State Protocol (SASP)
- Overview of z/OS Load Balancing Advisor (LBA)

## Workload balancing: a question of performance, availability, and scalability - multi-instance applications (data sharing)




**Application characteristics:**

- Multiple instances of the server are able to provide the exact same services to clients (will typically require data sharing)
- No state preserved at server between two connections (application protocol has to include support for such behavior or store state data in shared storage)

**Benefits of intelligent load balancing:**

- ***Performance*** - improving response time
- ***Availability*** - If one instance goes down, connections with it break, but new connections can be established with remaining instance(s)
- ***Scalability*** - more server instances can be added on demand (horizontal growth)

**Connection load balancing technologies:**

Between z/OS images:

- a DNS
- b NAT - CISCO CSM, CISCO CSS, many others
- c Dispatchers - ND, CISCO CSM, Sysplex Distributor
- d Contents-based - CISCO CSM, CISCO CSS, etc.

Inside single z/OS TCP/IP stack:

- e Port sharing

**Examples:**

- Web server
- TN3270 server
- Some CICS applications
- FTP server
- DB2
- MQ
- WAS
- LDAP
- RYO

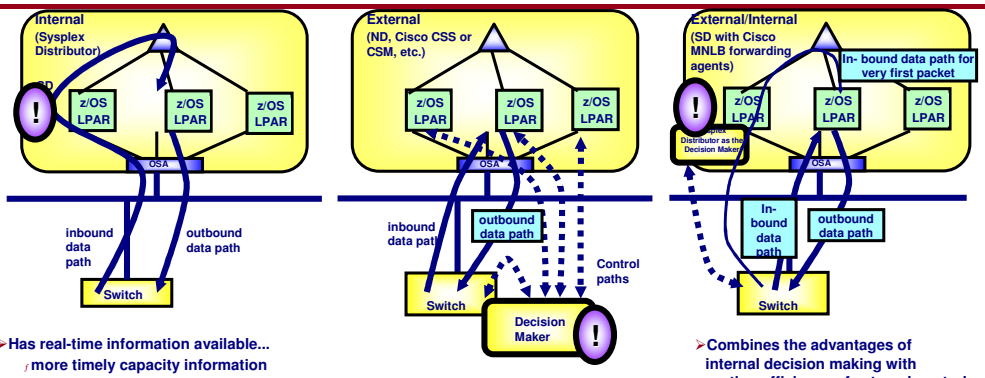


The load balancing technologies are the generalized ones. There are other solutions that are application-specific, such as the web servers use of WLM multiple address space support.

Application characteristics
Application must be multi-instance capable

No state or affinity
Benefits of intelligent load balancing
As opposed to round-robin approach

## Sysplex internal vs. external workload balancing - which technology is best for me?




- Has real-time information available...
  - more timely capacity information
  - QoS from Service Policy Agent
  - application-independent server availability
- No problems with shared OSA adapters; no intermediate routers
- Additional processing within z/OS system that is acting as the Sysplex Distributor
  - Inbound traffic funneled through single point that is then routed to destination system

- Specialized routing hardware may be more cost-effective
- May be configured for no single point of traffic flow
- So far no general Sysplex-capacity feedback technology (some specialized attempts have been made)
- Requires application-specific health probes ("application ping")
- Problems with shared OSA adapters or other intermediate routers

- Combines the advantages of internal decision making with routing efficiency of external control point
- Shared OSA adapters require use of Generic Routing Encapsulation by the switch.
- Preferred technology especially when workload includes large amounts of inbound data flows.

! Denotes decision point



Different ways to perform load balancing in Sysplex

Each has different benefits and drawbacks.
Which one you choose to use depends on which of the advantages and disadvantages are most important to you

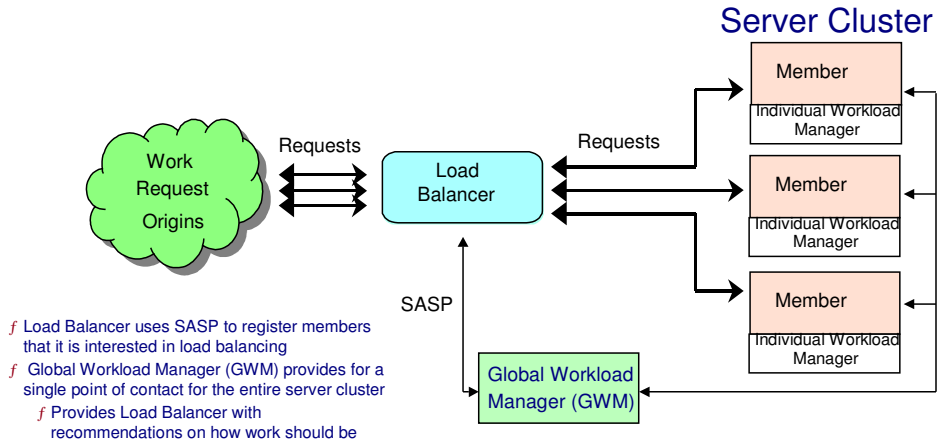
# Problem Statement

- Let's assume that you have selected an external IP load balancing solution for your z/OS Sysplex environment
  - Some possible reasons:
    - Prefer to have a single load balancing solution across multiple platforms in your environment
    - Administration of the load balancing functions belongs to network administration domain (not z/OS administrators)
    - Requirements for Content based load balancing
      - Need to perform load balancing/routing decisions based on data content (inspection of url, session IDs, cookies, etc.)
      - This is often combined with additional functions
        - SSL offloading functions (Need to decrypt data prior to inspection)
        - Web caching functions
- But what if the external load balancing solution has no awareness of the sysplex environment?
  - Is the sysplex treated just like any other server cluster?
  - Is it aware of the current/changing workload conditions on the various systems in the cluster?
  - Is it aware of the health and status of applications and/or systems?
- Making the external load balancing solution "sysplex aware" can help answer many of these questions
  - The z/OS Load Balancing Advisor is a key component that allows any external load balancing solution to become "sysplex aware".

# z/OS Load Balancing Advisor Solution

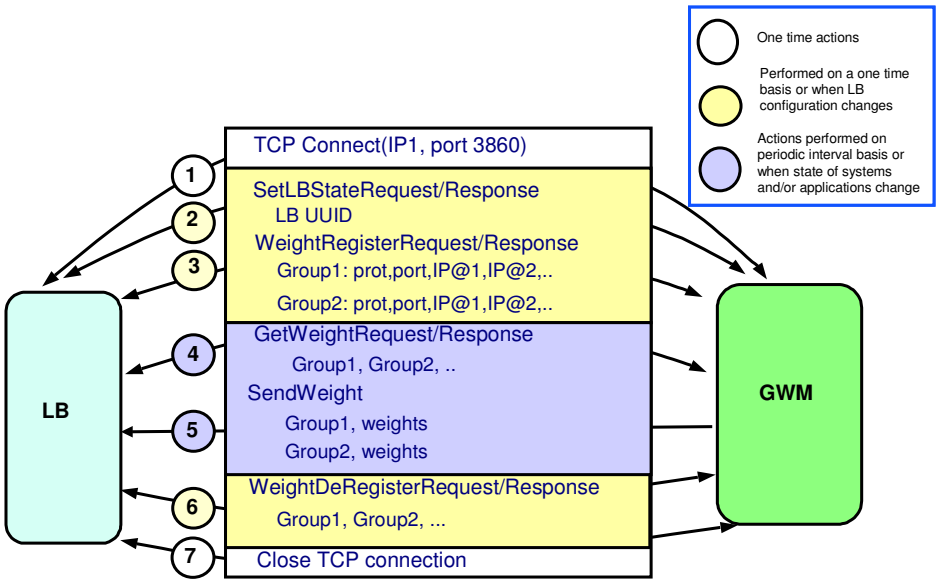


# Server Application State Protocol (SASP) Architecture




- Load Balancer uses SASP to register members that it is interested in load balancing
- Global Workload Manager (GWM) provides for a single point of contact for the entire server cluster
  - Provides Load Balancer with recommendations on how work should be balanced across servers
  - Based on current capacity and server availability
- SASP: Open protocol
  - currently being pursued as an individual Internet Draft RFC submitted to the IETF

# SASP
# Overview of Flows



# SASP Registration

- Load Balancer registers Groups of clustered servers it is interested in load balancing
  - Each group designates an application cluster to be load balanced
- Each group consists of a list of members (i.e. target servers)
  - System-level cluster: A list of target Systems identified by IP address (in lieu of individual application servers)
    - Recommendations returned in this scenario are also at a System-level
    - No specific target application information returned in this case
  - Application-level cluster: A list of applications comprising the “load balancing” group
    - Identified by protocol (TCP/UDP), IP address of the target system they reside on, and the port the application is using.
    - SASP allows for target servers in a load balancing group to use different ports (and even different protocols TCP/UDP)
      - Probably not applicable for real application workloads
- Support for both IPv4 and IPv6 protocols

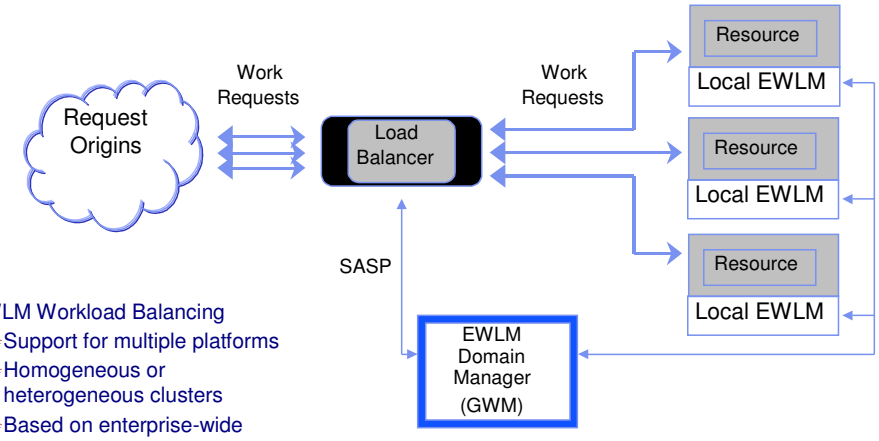
# SASP Update Frequency

- SASP supports both a "push" and a "pull" model for updating the load balancer with workload recommendations
  - Support of either by the load balancer is implementation dependent
  - Load balancer tells GWM which model it wants to use
  - "Pull" model
    - GWM "suggests" a polling interval to the load balancer
      - z/OS Load Balancing Advisor uses the configurable *update_interval* value for this purpose
    - Load balancer has the option to ignore this value
    - Load balancer requests updates each polling interval
  - "Push" model
    - GWM sends updated information to the load balancer on an interval basis
      - z/OS Load Balancing Advisor uses the configurable *update_interval* value for this purpose
    - GWM may send data more frequently than the interval period
  - Load balancer determines whether it wants information about all members it registered or only changed information about its registered members

# Products Supporting SASP

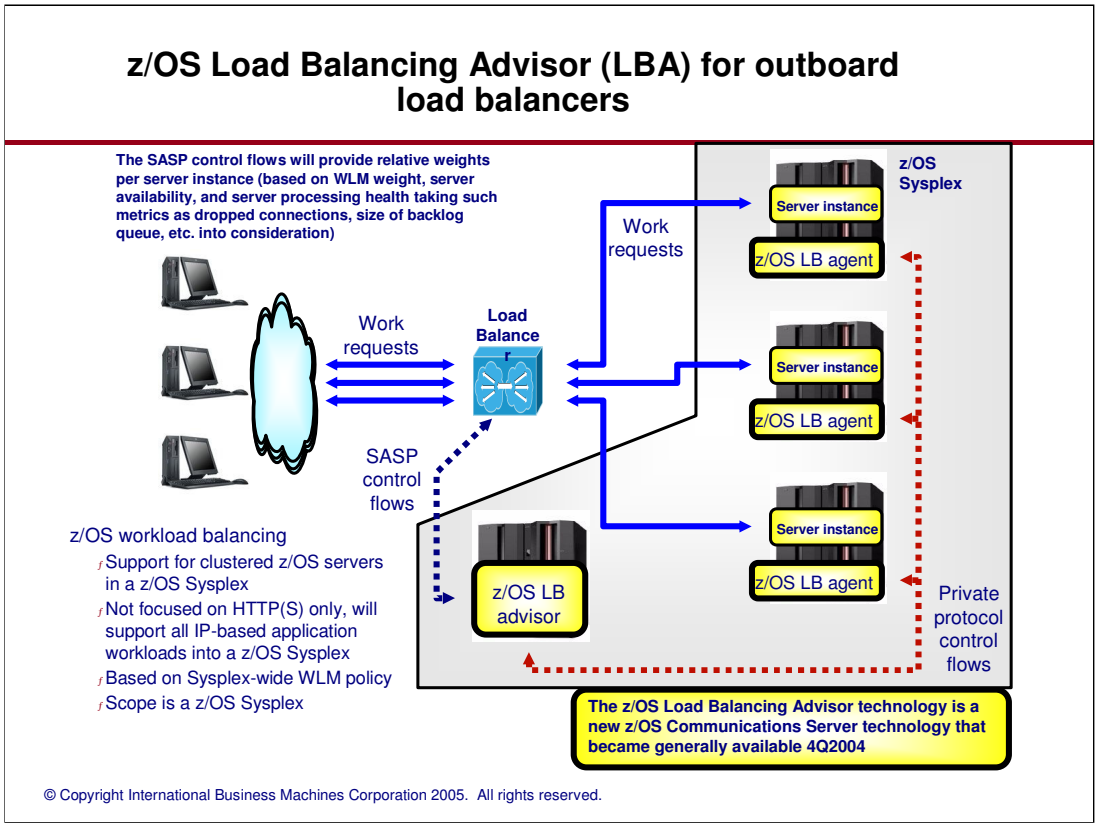
- SASP Global Workload Managers (GWMs)
  - EWLM (Enterprise Workload Manager)
    - Part of IBM Virtualization Engine 1.0
      - Supported Platforms:
        - IBM AIX 5L™ Version 5.2
        - Microsoft® Windows® 2000 Advanced Server, 2000 Server, 2003 Enterprise Edition, 2003 Standard Edition
        - Sun Microsystems Solaris 8 (SPARC Platform Edition), 9 (SPARC Platform Edition)
        - z/OS V1R6 (part of IBM Virtualization Engine Enterprise Workload Manager for z/OS V1.1.0 )
  - z/OS Load Balancing Advisor (became available 4Q2004)
    - Part of z/OS Communications Server (z/OS V1R4 and higher)
      - V1R4 (APAR PQ90032) - V1R5 and V1R6 (APAR PQ96293)
        - Documentation of function available at:
          http://www-1.ibm.com/support/docview.wss?uid=swg27005585
      - Will be part of the base z/OS V1R7 Communications Server
      - Can be used within scope of EWLM z/OS solution or in a z/OS WLM environment
- Load Balancers
  - CISCO CSM level 4.1 (2.5)
  - Other vendors possibly in the future

# SASP - EWLM Workload Balancing




EWLM Workload Balancing

- Support for multiple platforms
- Homogeneous or heterogeneous clusters
- Based on enterprise-wide EWLM policy




Standard protocols used between LBA and load balancer
SASP is being pursued as an open standard in the IETF

The initial draft has already been published and the work adopted by the reliable server pool working group
The LBA uses a private protocol to obtain weights and recommendations for server instances from LB agents
A LB agent runs on each target stack in the Sysplex

Weights are based on WLM recommendations, as well as various QoS metrics

When the weight of a particular server instance changes, the LBA notifies the Load Balancer using SASP flows
This can be a push or pull model, depending on the Load Balancer requests

# Solution - z/OS Implementation
# Load Balancer Config

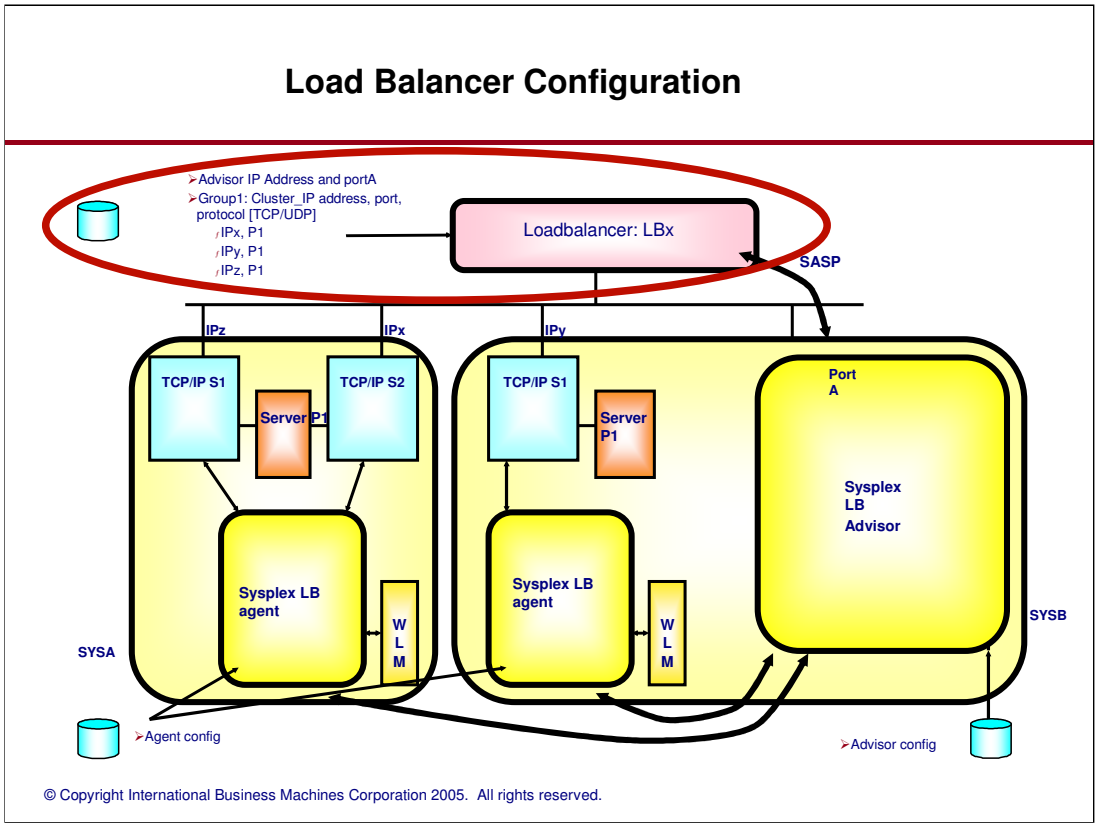
- Group definition rules/guidelines
  - Multiple Groups may be defined representing different server clusters
  - All Members of a Group must belong to the same sysplex, hence,
    - Members of the same Group must all be managed by the z/OS Load Balancing Advisor, or EWLM. A Group may not be managed by both.
  - May not contain mixtures of application members and system members
  - Clients connect to the cluster IP address of the group
- Member definition rules/guidelines
  - IP addresses should be static VIPA addresses for availability reasons
  - Should NOT contain the following types of addresses
    - Distributed DVIPAs
    - Addresses which are not unique within the sysplex
    - Addresses which would not be reachable from the Load Balancer, including
      - Loopback addresses
      - "Unavailable" IPv6 addresses
    - "Deprecated" IPv6 addresses

## The weights

- The weights are composed of two main elements:
  - ***WLM weight***
    - –The WLM weight capacity as we know from other WLM-based load balancing solutions, such as Sysplex Distributor (System or Server-specific)
      - ✓A numeric value between 0 and 64
  - ***Communications Server weight***
    - –This weight is calculated based on the availability of the actual server instances (are they up and ready to accept workload) and how well TCP/IP and the individual server instances process the workload that is sent to them.
      - ✓Expressed as a numeric percentage value between 0 and 100
    - –Purpose of calculations is to:
      - ✓Prevent stalled server from being sent more work (accepting no new connections and new connections are being dropped due to backlog queue full condition)
      - ✓Proactively react to server that is getting overloaded (accepting new connections, but size of backlog queue increases over time approaching the max backlog queue size)
- The final weight is calculated by combining the WLM and the CS weights into a single metric
  - Final weight = WLM weight * CS weight / 100
- Due to current external load balancer behavior when a weight of zero is returned, the z/OS LBA currently will never return a zero weight - the lowest weight it will return is a weight of 1
  - Weights that are returned to the load balancer are normalized to values between 1 and 64
    - –If all server instances have the same final weight (example 32), then a 1 will be returned for all server instances



Weights are a combination of WLM recommendations and QoS information maintained by the TCP/IP stack

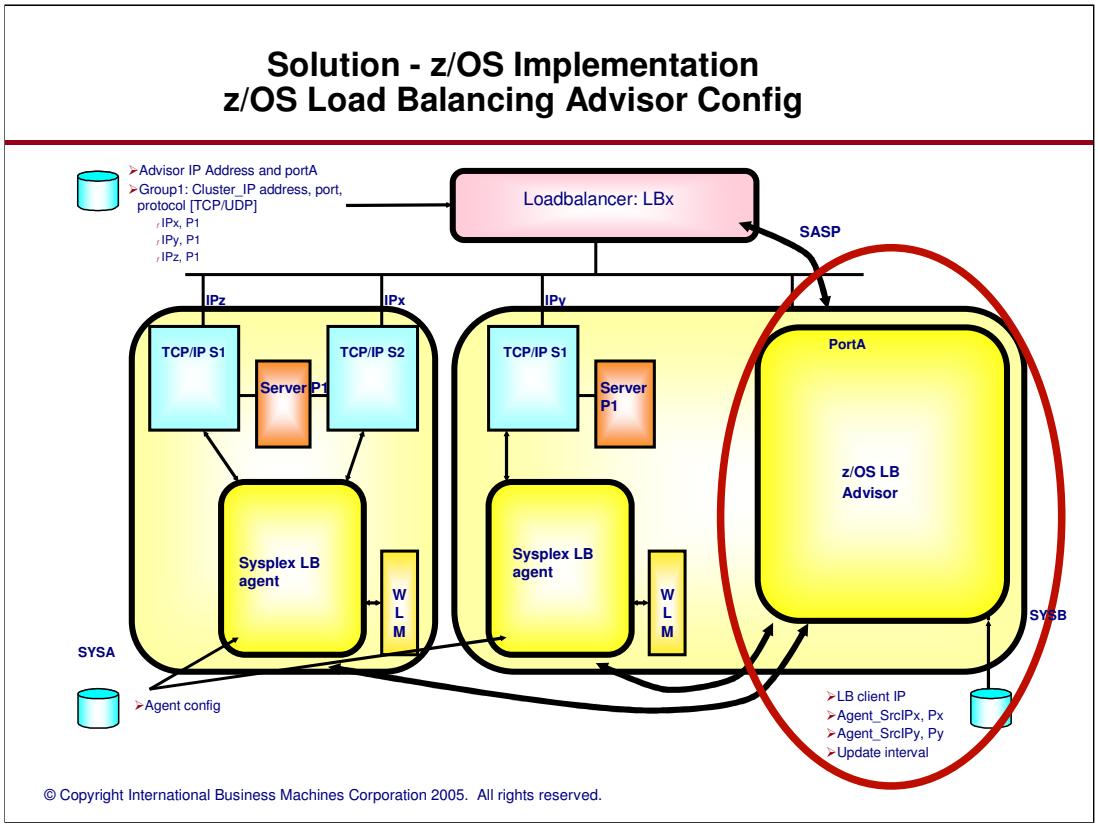



Load Balancer configures...
IP address and port (PortA) of Advisor

One or more groups of homogeneous applications or systems (e.g. Group1)

The cluster IP address clients use to connect to an application (e.g. Cluster_IP address)

The protocol which the group will use (TCP or UDP)
For each group
IP address, and port of each application in the group (e.g. IPx P1, IPy P1, IPz P1)




Advisor configuration includes...

For each Load Balancer
IP address of SASP client (LB client IP)
For each Agent
IP address and port (e.g. Agent_SrcIPx, Px)
How often data should be sent from the Agents to the Advisor (update interval). Update interval may also indicate how often to send data to the load balancer (implementation dependent)

Which type of WLM recommendations to use (System vs. Server-specific)

Other optional data (not shown, see later foils)

# z/OS Load Balancing Advisor

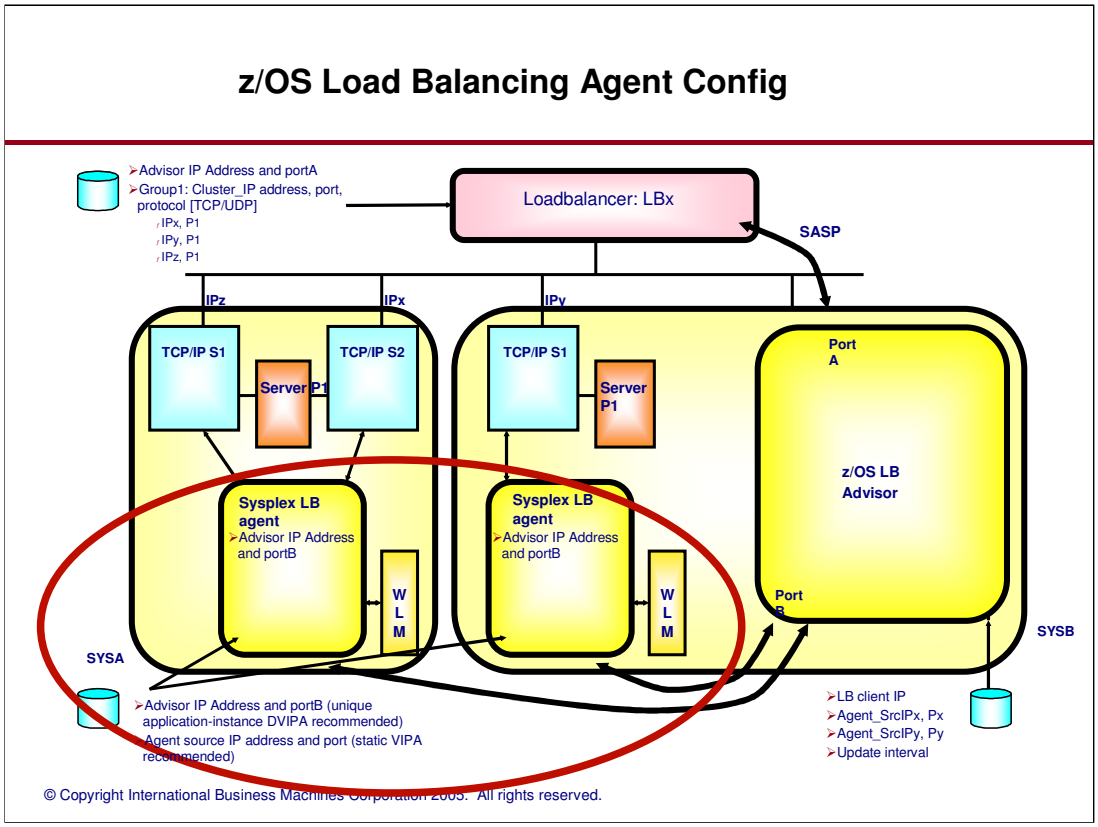
- z/OS LB Advisor
  - New, stand-alone application
    - Started via an MVS Started Task (accepts MVS operator commands for display and modification purposes)
  - Executes on any system within the sysplex
    - Provides Load Balancing advice for any TCP/UDP server applications within the sysplex
    - Acts as a TCP server application supporting SASP (port 3860 by default, but can be customized)
    - Supports multiple LBs concurrently, only one active instance allowed per sysplex
    - Does not require Sysplex Distributor to be configured
  - Communicates with local Load Balancing Agents
    - Uses TCP connections, acts as TCP server (on separate port from SASP)
    - Obtains server topology information and workload balancing recommendations from each target system and for each target application
  - Configuration
    - Must identify all eligible Load Balancing Agents (by source IP address and source port)
    - Must identify all eligible Load Balancers by source IP address
    - IP address and port it should listen to (Application Specific Dynamic DVIPA strongly encouraged)
    - Other parameters (debug levels, polling interval, WLM recommendations, etc.)

# WLM recommendations

- System vs. Server-specific WLM recommendation types
  - Configurable as a global default *(wlm* statement)
    - Values for *wlm* statement are *basewlm* or *serverwlm*
    - Default value is *basewlm*, meaning, use System WLM recommendations
    - Chosen as the default for compatibility with pre-V1R7 systems
    - *serverwlm* value means use Server-specific WLM recommendations
    - May be overridden on a port number basis *(port_list* statement)
  - System WLM recommendations
    - Reflect the displaceable capacity of the z/OS system relative to other z/OS systems in the sysplex
  - Server-specific WLM recommendations
    - Reflect how well the application is meeting its WLM goals, and...
    - How much displaceable capacity is available on the target system at the importance level of the application
    - Only available with z/OS V1R7
      - Pre-V1R7 solution supports System level WLM weights only

# System vs. Server-specific WLM recommendations

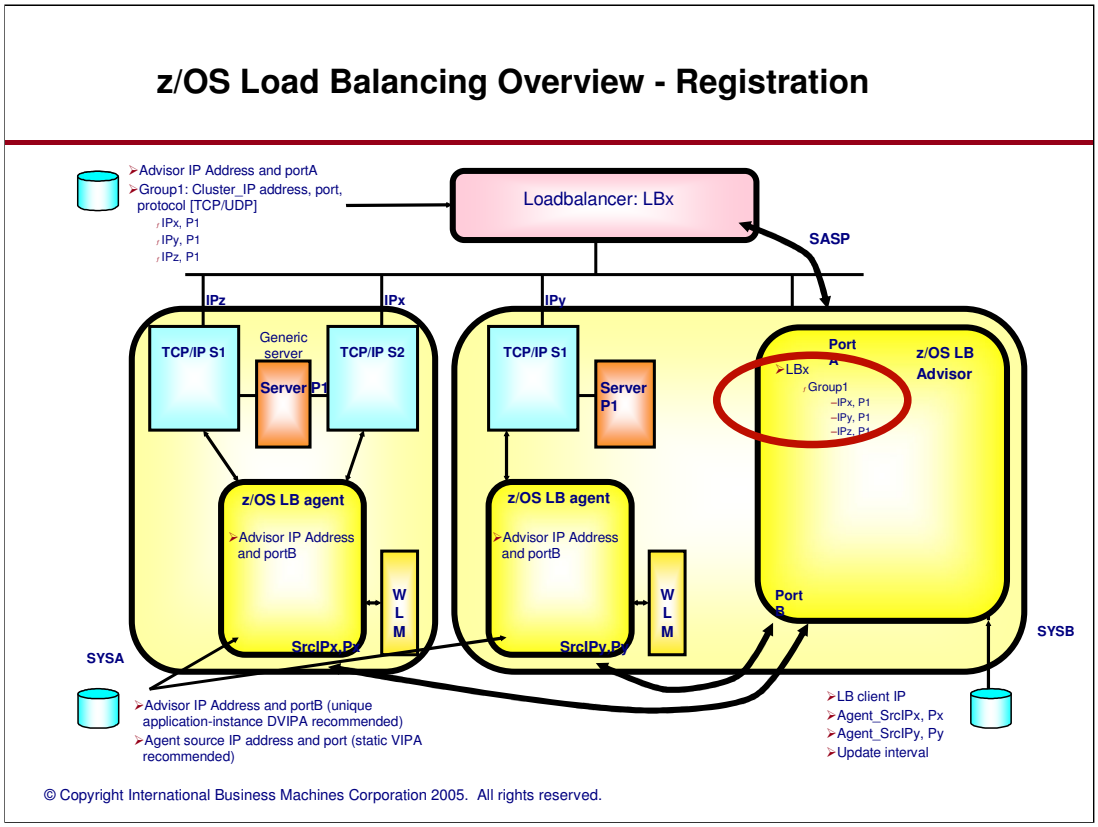
- Choosing WLM recommendation types
  - Server-specific *(serverwlm)* recommendations are recommended for most applications
    - Provide more granular, more accurate load balancing information
  - Some applications may be better suited for System WLM recommendations *(basewlm)*
    - Those which serve as an access point to applications which run in a separate address space (and therefore, a different service class)
    - INETD
    - FTPD
      - If FTP servers run in the same service class as the FTP daemon, use Server-specific WLM recommendations (*serverwlm*)
      - However, if FTP servers run in a different service class from the FTP daemon, then use System WLM recommendations (*basewlm*)




Agent configuration includes...
IP address and port the Agent will bind to (source IP address and port)
Static VIPA is recommended
IP address and port used to reach the Advisor
Unique application-instance DVIPA recommended

# z/OS Load Balancing Agent

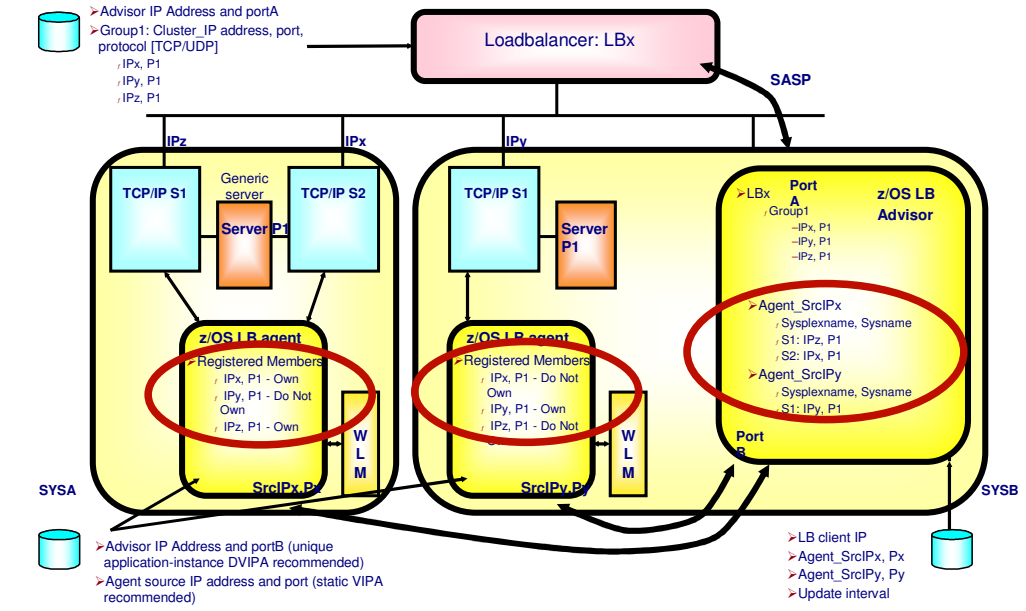
- z/OS LB Agent
  - New, stand-alone application
    - Started via an MVS Started Task
    - Accepts MVS operator commands for display and modification purposes
  - Executes on every target system in the sysplex
    - Or at least on every system in the sysplex that is a target of a load balanced request
    - Provides Load Balancing advice for specified TCP/UDP server applications on local system
    - Only one active instance allowed per MVS system
  - Supports multiple TCP/IP stacks and all known server types: stack-affinity, generic, bind-specific, shareport groups etc.
    - Computes weights based on WLM, server availability, server health (dropped connections due to backlog queue full or dropped datagrams due to UDP queue limit exceeded)
    - Server-specific WLM weights (V1R7) or system WLM weights
  - Simple Configuration
    - Specify IP address and port for Load Balancing Advisor
    - Specify Source IP address/port to be used in connecting to Advisor
      - –Static VIPA recommended for single stack systems (allows for failures in physical interfaces)
      - –For CINET, unique application-instance DVIPA recommended
      - –The same source port can be used by all Agents (simplifies configuration)
    - Optionally specify the debug level




Each group and its corresponding members are cached in the Advisor

Data is stored on a per-Load Balancer basis

## z/OS Implementation Agent Connection

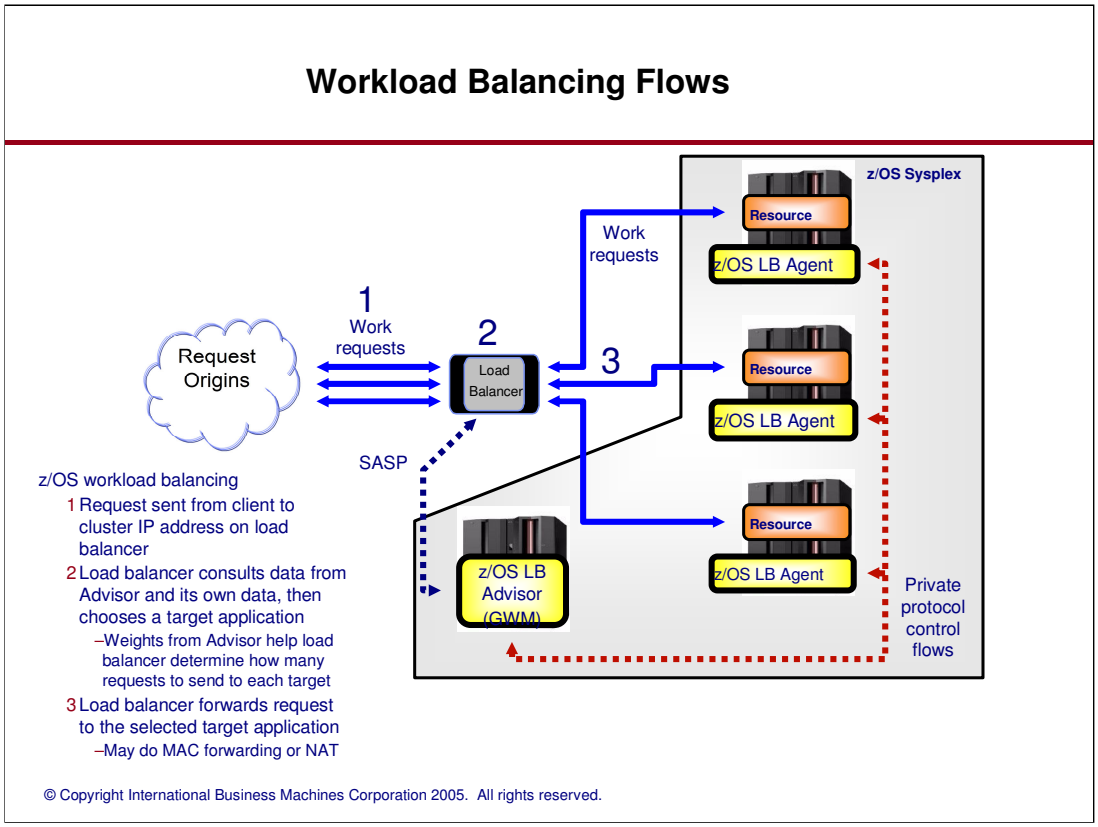





Groups registered by load balancers are sent to the Agents

Each Agent reports back on which members it owns, and provides weight data, which is cached

Advisor aggregates availability and weight information for each member and reports that data to the appropriate load balancer




The following is repeated periodically as long as the Advisor and Agent application instances are active. Agents collect information from the sysplex about members the load balancer has registered, including availability information, and capacity of the target application's system, as well as the ability of the target application to handle more workload.

Advisor consolidates information for all Agents and assigns weights to the members.

Advisor sends the availability information and weights to the load balancer.

Client sends workload request to a cluster IP address which reaches the load balancer.

Load balancer finds a Group defined for that cluster IP address, port and protocol
Load balancer consults the availability and weight information for each target application (member) within

# Trademarks and notices

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

  - AIX®
  - AnyNet®
  - AS/400®
  - Candle®
  - CICS®
  - CICSPlex®
  - CICS/ESA®
  - DB2®
  - DB2 Connect™
  - DPI®
  - DRDA®
  - e business(logo)®
  - ESCON®
  - eServer™
  - ECKD™
  - FFST™
  - GDDM®
  - GDPS®
  - HiperSockets™
  - IBM®
  - Infoprint®
  - IMS™
  - IP PrintWay™
  - iSeries™
  - Language Environment®
  - MQSeries®
  - MVS™
  - MVS/ESA™
  - NetView®
  - OS/2®
  - OS/390®
  - Parallel Sysplex®
  - PrintWay™
  - PR/SM™
  - pSeries®
  - RACF®
  - Redbooks™
  - Redbooks (logo)™
  - S/390®
  - System/390®
  - ThinkPad®
  - Tivoli®
  - Tivoli (logo)®
  - VM/ESA®
  - VSE/ESA™
  - VTAM®
  - WebSphere®
  - xSeries®
  - z/Architecture™
  - z/OS®
  - z/VM®
  - zSeries®

- Cisco, Cisco Systems, the Cisco Systems logo, Catalyst, and Cisco IOS are registered trademarks or trademarks of Cisco Systems, Inc. and/or its affiliates in the United States and certain other countries.