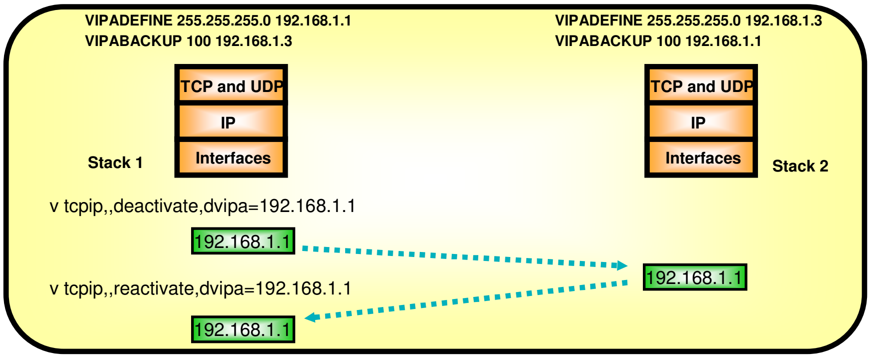
Improved Sysplex Operations

IBM Software Group, Enterprise Networking and Transformation Solutions

Content

- Review improved Sysplex operations:
- Improved Operations for Sysplex Functions (z/OS V1R7)
- Improved Utilization of XCF resources (z/OS V1R6 and z/OS V1R7)
- Improvements for external load balancing solutions:
  The z/OS Load Balancing Advisor
- z/OS V1R5 enhancements (See Appendix)

Improved operations: operator-initiated movement of individual stack-managed dynamic VIPA addresses



```
VIPADEFINE 255.255.255.0 192.168.1.1
VIPABACKUP 100 192.168.1.3
```

Stack 1

TCP and UDP

IP

Interfaces

```
VIPADEFINE 255.255.255.0 192.168.1.3
VIPABACKUP 100 192.168.1.1
```

Stack 2

TCP and UDP

IP

Interfaces

```
v tcpip,,deactivate,dvipa=192.168.1.1
```

192.168.1.1 → 192.168.1.1

```
v tcpip,,reactivate,dvipa=192.168.1.1
```

192.168.1.1 ← 192.168.1.1

- **Deactivate**
  - DVIPA is deactivated and a configured backup stack will takeover the DVIPA
  - Backup DVIPA can be deactivated also removing eligibility as a backup
- **Reactivate**
  - Original owner can regain ownership
  - Can also reactivate a backup DVIPA that's been deactivated
  - Prior to these commands, Vary obey files were needed to cause a DVIPA takeover
  - These commands can not be used for DVIPAs created from VIPARANGE with bind, ioctl(), or the Modvipa utility

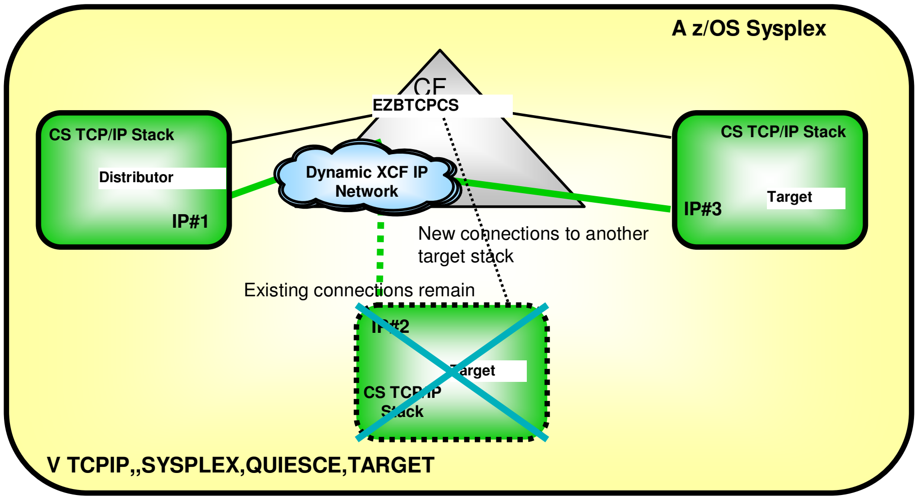
Determining current status of DVIPAs

Example of a Netstat VIPADCFG/-F report after a DVIPA is deactivated via:

VARY TCPIP,,SYSPLEX,DEACT,DVIPA=197.11.221.1

```
netstat vipadcfg
MVS TCP/IP NETSTAT CS V1R7       TCPIP Name: TCPCS           19:52:38
Dynamic VIPA Information:

  VIPA Define:
    IpAddr/PrefixLen: 197.11.221.2/24
      Moveable: Immediate  SrvMgr: No

Deactivated Dynamic VIPA Information:

  VIPA Define:
    IpAddr/PrefixLen: 197.11.221.1/24
      Moveable: Immediate  SrvMgr: Yes

  VIPA Distribute:
    Dest:        197.11.221.1..20
      DestXCF:   ALL
        SysPt:   No   TimAff: No    Flg: BaseWLM
    Dest:        197.11.221.1..21
      DestXCF:   ALL
        SysPt:   No   TimAff: No    Flg: BaseWLM
```

Improved operations: operator-initiated quiesce/resume of target systems/applications



V TCPIP,,SYSPLEX,QUIESCE,TARGET

Improved operations: operator-initiated quiesce and resume of individual server applications or full target systems

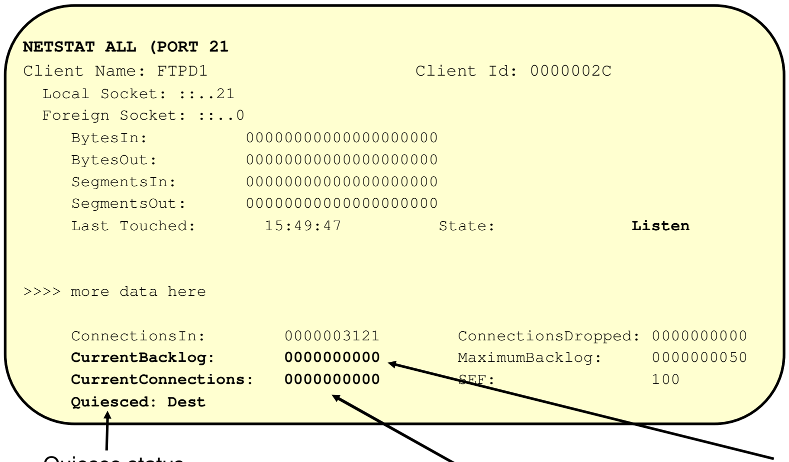
- **Ability to quiesce a system or an application prior to shutdown**
  - Planned maintenance scenarios of system or application
    - Allows existing systems or applications to drain work queue prior to shutdown
  - Relieve temporary constraints of resources on target system
  - Temporary - Does not affect Sysplex Distributor's permanent configuration
  - Supports applications in SHAREPORT group (must be targets for SD)
  - Issued on target system being affected
  - Only way to achieve similar capability earlier was via temporary configuration changes based on OBEYFILE commands

- **VARY TCPIP,,SYSPLEX,QUIESCE,options**
  - TARGET - Quiesces all applications on target stack.
  - PORT=xxx - Quiesce all applications bound to the specified port on this stack
    - JOBNAME=jobname - Allows quiesce of a single application in SHAREPORT group
    - ASID=asid - Further qualify job being quiesced (such as when dealing with duplicate jobnames)
  - No new TCP connections sent to the quiesced target (stack or application)
    - For all Distributed DVIPAs that the entity is a target for
  - Existing TCP connections are maintained (or in other words, the process is non-disruptive)

- **VARY TCPIP,,SYSPLEX,RESUME,options**
  - TARGET|PORT|JOBNAME|ASID
  - Allows identified target stacks and/or applications to once again be targets for distribution

connections in

When is it safe to recycle a quiesced application without impact to end users?



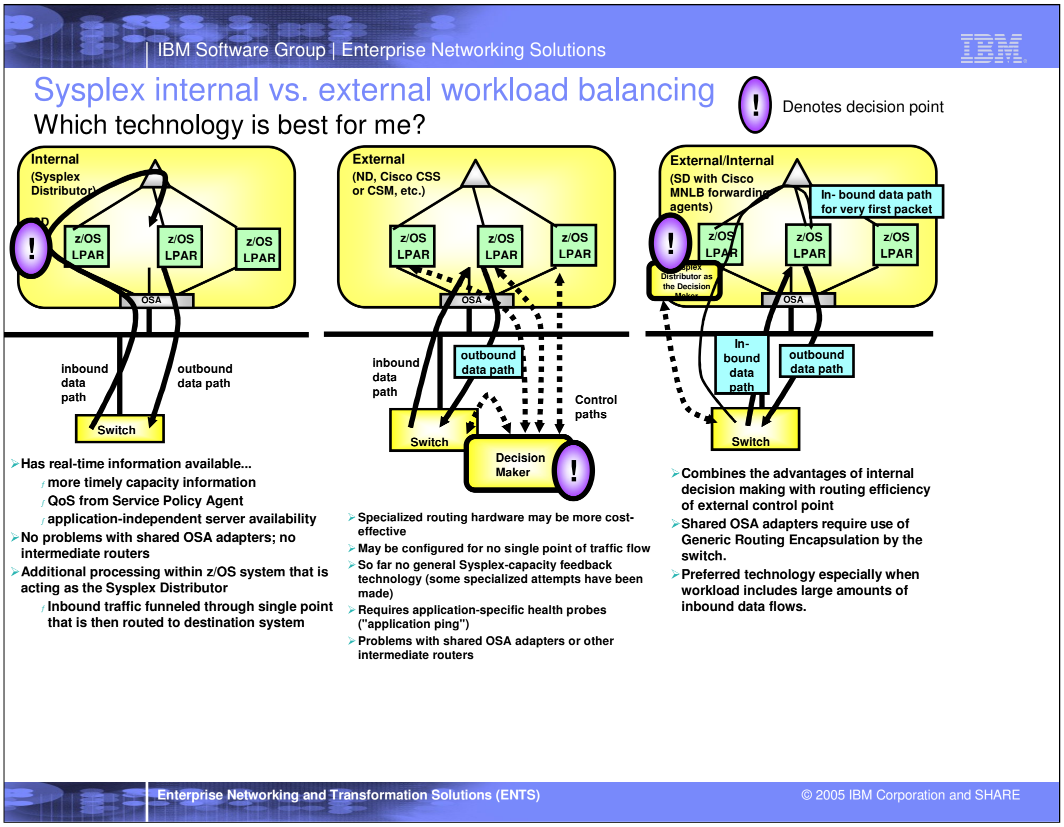
```
NETSTAT ALL (PORT 21
Client Name: FTPD1                    Client Id: 0000002C
  Local Socket: ::..21
  Foreign Socket: ::..0
     BytesIn:          00000000000000000000
     BytesOut:         00000000000000000000
     SegmentsIn:       00000000000000000000
     SegmentsOut:      00000000000000000000
     Last Touched:       15:49:47         State:              Listen


>>>> more data here

     ConnectionsIn:        0000003121       ConnectionsDropped: 0000000000
     CurrentBacklog:       0000000000       MaximumBacklog:     0000000050
     CurrentConnections:   0000000000       SEF:                100
     Quiesced: Dest
```

Quiesce status

No connections being processed and no connections in the backlog

Improvements for external load balancing solutions:

The z/OS Load Balancing Advisor



IBM Software Group | Enterprise Networking Solutions

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.

Enterprise Networking and Transformation Solutions (ENTS)

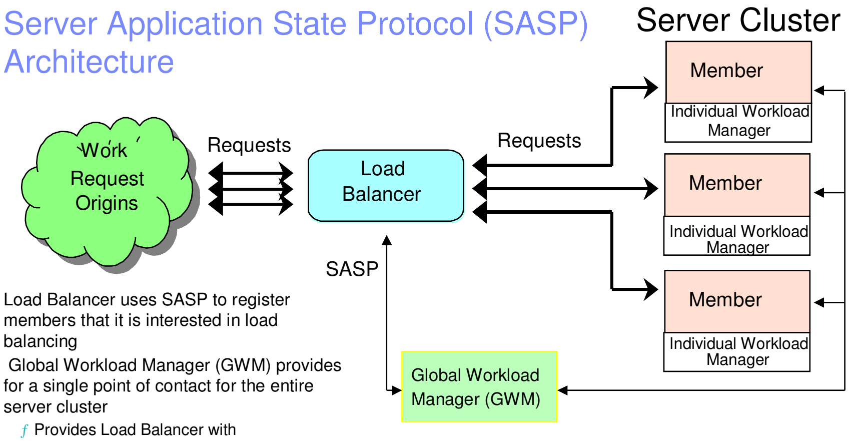
Different ways to perform load balancing in Sysplex

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".

Server Application State Protocol (SASP) Architecture


Server Cluster
Work Request Origins
Requests
Load Balancer
Requests
Member
Individual Workload Manager
Member
Individual Workload Manager
Member
Individual Workload Manager
SASP
Global Workload Manager (GWM)

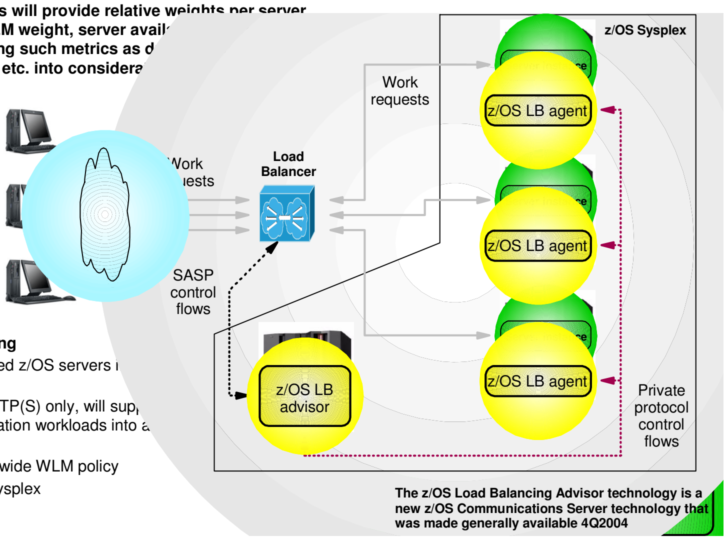
- Load Balancer uses SASP to register members that it is interested in load balancing
- Global Workload Manager (GWM) provides for a single point of contact for the entire server cluster
  - Provides Load Balancer with recommendations on how work should be balanced across servers
  - Based on current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

Products Supporting SASP

- IBM products supporting SASP
  - 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)
  - Nortel Alteon
  - Other vendors possibly in the future

z/OS Load Balancing Advisor (LBA) for outboard load balancers

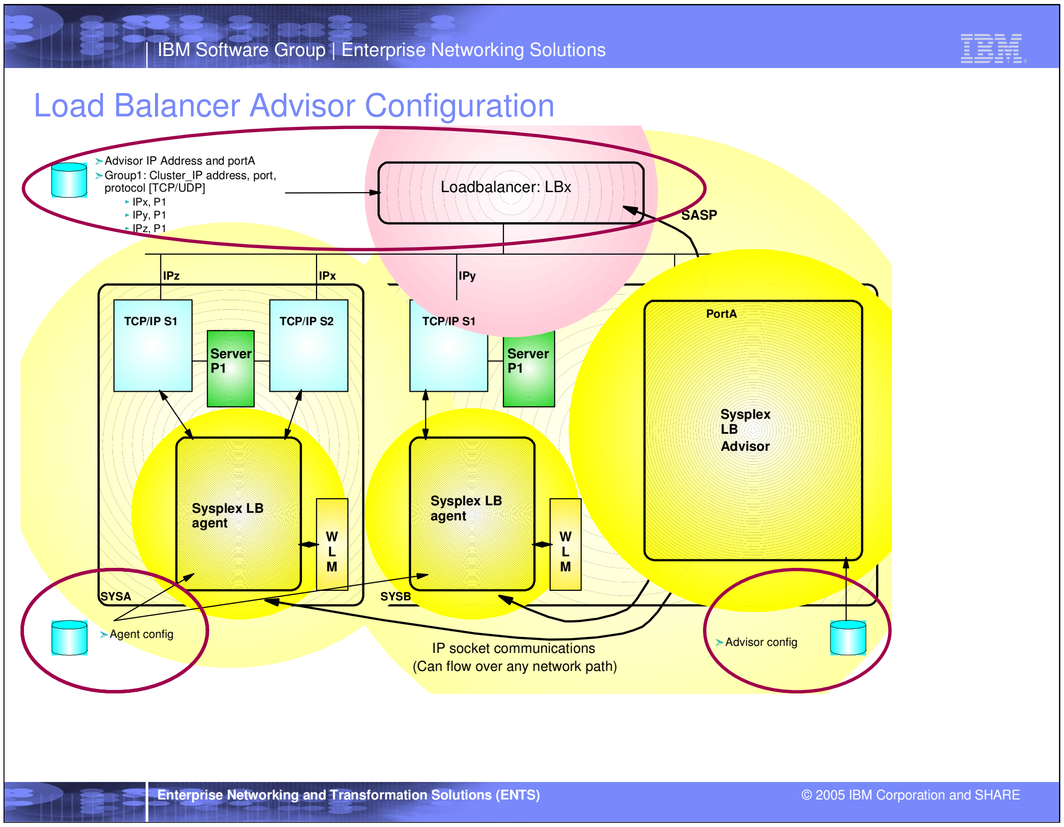
The SASP control flows will provide relative weights per server instance (based on WLM weight, server avail processing health taking such metrics as d size of backlog queue, etc. into considera



The z/OS Load Balancing Advisor technology is a new z/OS Communications Server technology that was made generally available 4Q2004

z/OS workload balancing

- Support for clustered z/OS servers i z/OS Sysplex
- Not focused on HTTP(S) only, will sup all IP-based application workloads into a z/OS Sysplex
- Based on Sysplex-wide WLM policy
- Scope is a z/OS Sysplex



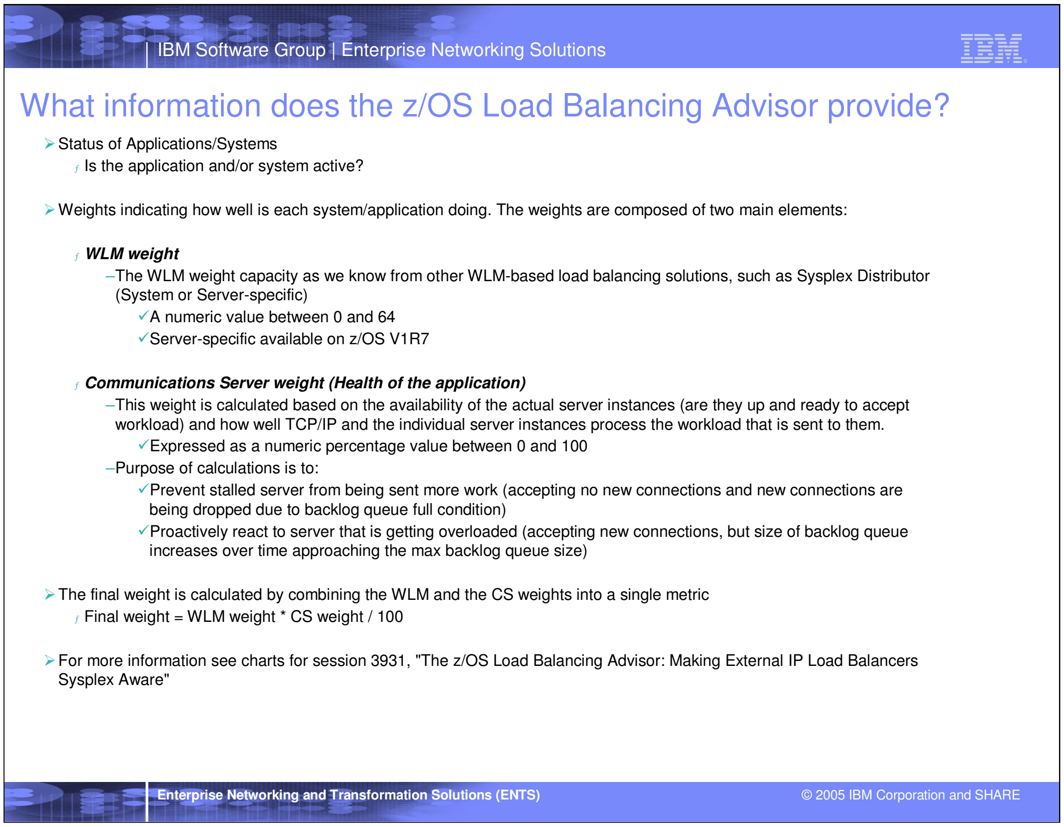
Load Balancer Advisor Configuration

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)



What information does the z/OS Load Balancing Advisor provide?

- Status of Applications/Systems
  - Is the application and/or system active?

- Weights indicating how well is each system/application doing. 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
      - ✓Server-specific available on z/OS V1R7

  - ***Communications Server weight (Health of the application)***
    - 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

- For more information see charts for session 3931, "The z/OS Load Balancing Advisor: Making External IP Load Balancers Sysplex Aware"

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

Some strategies for workload request balancing into a z/OS Sysplex

A **DNS/WLM as workload balancing should not be used any longer**

B **Where HTTP workload is to be balanced based on content of HTTP requests, an outboard load balancer that supports contents inspection must be deployed**

- If HTTPS workload is to be included, the load balancing node must be accompanied by an SSL/TLS offload technology
- Can be combined with a cache appliance for improved performance

C **UDP workload balancing must be deployed using an outboard load balancer - SD does not support UDP balancing**

D **Remaining TCP connection balancing can be deployed using either SD or an outboard load balancer:**

- SD has more real-time information available than outboard load balancers - even with outboard load balancers using the SASP protocol
- Who is to apply management control over the workload balancing function will be a major factor in deciding which solution to use
- When using an outboard load balancer the use of SASP is recommended as that will help improve the load balancing decisions (makes the outboard load balancers *"sysplex aware"*)

For More Information....

URL	Content
http://www.ibm.com/servers/eserver/zseries	IBM eServer zSeries Mainframe Servers
http://www.ibm.com/servers/eserver/zseries/networking	Networking: IBM zSeries Servers
http://www.ibm.com/servers/eserver/zseries/networking/technology.html	IBM Enterprise Servers: Networking Technologies
http://www.ibm.com/software/network/commserver	Communications Server product overview
http://www.ibm.com/software/network/commserver/zos/	z/OS Communications Server
http://www.ibm.com/software/network/commserver/z_lin/	Communications Server for Linux on zSeries
http://www.ibm.com/software/network/ccl	Communication Controller for Linux on zSeries
http://www.ibm.com/software/network/commserver/library	Communications Server products - white papers, product documentation, etc.
http://www.redbooks.ibm.com	ITSO redbooks
http://www.ibm.com/software/network/commserver/support	Communications Server technical Support
http://www.ibm.com/support/techdocs/	Technical support documentation (techdocs, flashes, presentations, white papers, etc.)
http://www.rfc-editor.org/rfcsearch.html	Request For Comments (RFC)

Appendix:
Sysplex Enhancements
z/OS V1R5

IP Sysplex enhancements in z/OS V1R5

- Increase ports on VIPADISTRIBUTE from 4 to 64 (PTFed back to z/OS V1R2 - APAR PQ65205)

- Dynamic port definition for VIPADISTRIBUTE dynamic VIPA when server binds to dynamic VIPA

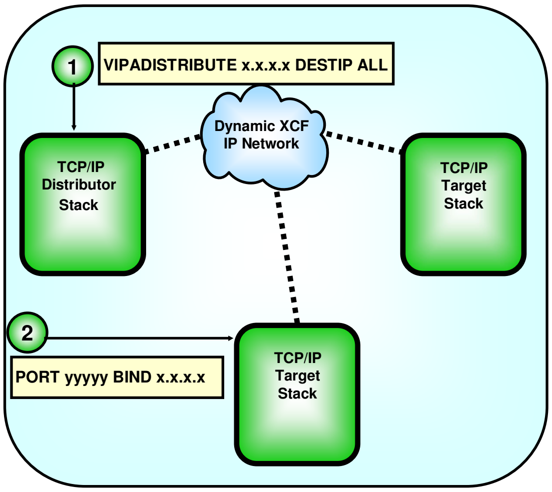
- Increase limit of DVIPAs per stack from 256 to 1024

- New round-robin distribution method in Sysplex Distributor (PTFed back to z/OS V1R4 - APAR PQ76866)
  - Alternative to WLM-based distribution
  - Useful where availability is more important than capacity

- Sysplex Distributor affinity
  - Configurable timer-based stickyness per source IP address, server DVIPA and port

- Support DVIPA activation based on VIPABACKUP before VIPADEFINE ever processed

Dynamic Port Assignment (z/OS V1R5)



Allows applications on any port to become automatic candidates for Sysplex Distributor load balancing when they bind to a Distributed DVIPA

1 If PORT is omitted from VIPADISTRIBUTE definition, the Sysplex Distributor reacts dynamically to servers binding to the distributed DVIPA and creating a listening socket, by adding a port to the list of ports for which connection workload balancing will occur.

2 Servers either explicitly bind to distributive DVIPA or configure "PORT BIND" definition if binding to INADDR_ANY.

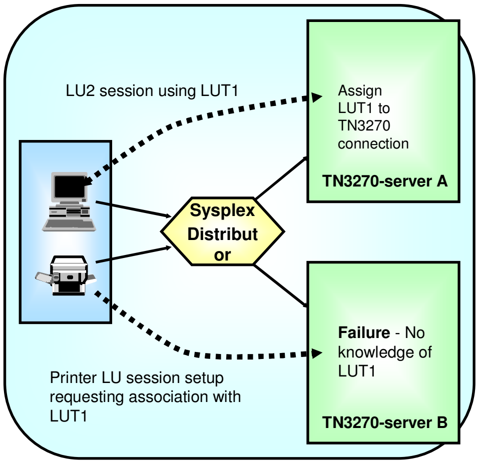
Benefits:

- Lifts restriction on number of ports that can be distributed for a DVIPA.
- Minimizes updates to the VIPADISTRIBUTE statement
- ***But*** applications being load balanced may be less visible/controlled!

Application Server Affinity (z/OS V1R5)

- Optional keyword TIMEDAFFINITY on VIPADISTRIBUTE statement determines whether or not a connection from a client (as identified by source IP address) to a particular server instance of several serviced by Sysplex Distributor shall establish an affinity for future connections from the same client (IP address) to the same Distributed DVIPA and port(s).

- The affinity duration maintained between connections differs from application to application (configurable)
  - The affinity timer begins after the last connection from that client has terminated
    - As long as new TCP connections from that client arrive within this interval, the affinity persists
- Beware of Gateway type of applications!
  - All connections from the Gateway application will have affinity to a single server!

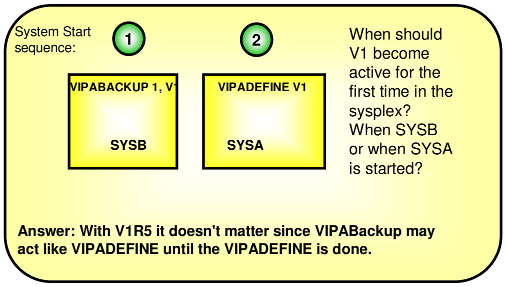

LU2 session using LUT1
Assign LUT1 to TN3270 connection
TN3270-server A
Sysplex Distributor
Failure - No knowledge of LUT1
TN3270-server B
Printer LU session setup requesting association with LUT1

Allowing for DVIPA activation when owning system is not started

- Support DVIPA activation based on VIPABACKUP before VIPADEFINE ever processed

- Allows independent startup order for DVIPA owning systems and backup systems
  - Backup system will retain ownership of the DVIPA until the primary owning system is started
  - Ownership reverts to "VIPADEFINE" stack when that stack is activated

- Requires changes to the VIPABACKUP statements
  - MOVEABLE IMMEDIATE, address mask, etc. should be specified
  - Also requires that VIPADISTRIBUTE statement is present (after the VIPABACKUP statement) on the Backup TCP/IP stacks



When should V1 become active for the first time in the sysplex? When SYSB or when SYSA is started?

Answer: With V1R5 it doesn't matter since VIPABackup may act like VIPADEFINE until the VIPADEFINE is done.

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:

- AIX7
- AnyNet7
- AS/4007
- Candle7
- CICS7
- CICSPlex7
- CICS/ESA7
- DB27
- DB2 Connect™
- DPI7
- DRDA7
- e business(logo)7
- ESCON7
- eServer™
- ECKD™
- FFST™
- GDDM7
- GDPS7
- HiperSockets™
- IBM7
- Infoprint7
- IMS™
- IP PrintWay™
- iSeries™
- Language Environment7
- MQSeries7
- MVS™
- MVS/ESA™
- NetView7
- OS/27
- OS/3907
- Parallel Sysplex7
- PrintWay™
- PR/SM™
- pSeries7
- RACF7
- Redbooks™
- Redbooks (logo)™
- S/3907
- System/3907
- ThinkPad7
- Tivoli7
- Tivoli (logo)7
- VM/ESA7
- VSE/ESA™
- VTAM7
- WebSphere7
- xSeries7
- z/Architecture™
- z/OS7
- z/VM7
- zSeries7

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.