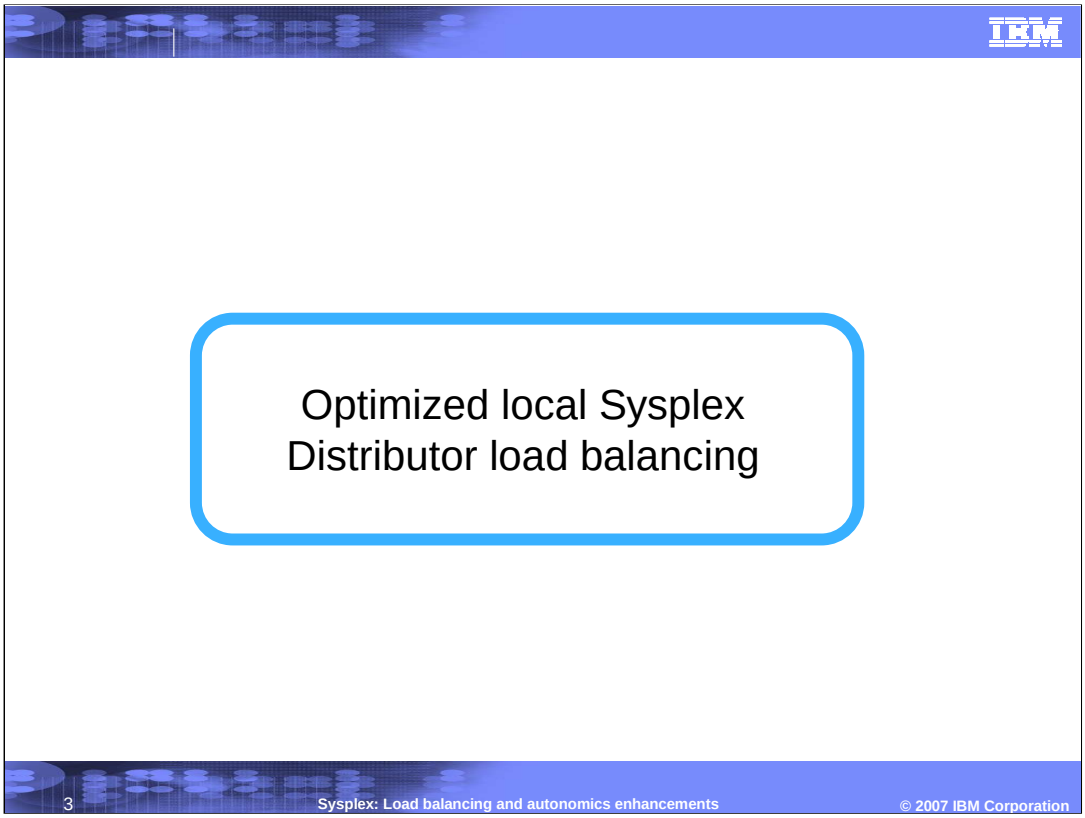
IBM eServer™

Sysplex: Load balancing and autonomics enhancements

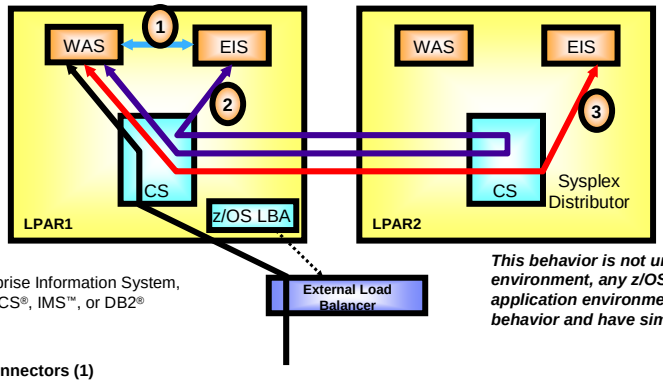
ebusiness on demand software

Agenda - Load Balancing and Autonomics Enhancements

1. **Optimized local Sysplex Distributor load balancing**
2. **Support for WLM reporting of abnormal conditions**
3. **Sysplex autonomics adds monitoring of selected network interfaces**


Optimized local Sysplex
Distributor load balancing
3
Sysplex: Load balancing and autonomics enhancements
© 2007 IBM Corporation

Local vs. remote connector support in today's z/OS® environment



Today, multi-tier subsystems and applications need to make a trade-off between availability and performance objectives.

EIS: Enterprise Information System, such as CICS®, IMS™, or DB2®

This behavior is not unique to a WAS environment, any z/OS Sysplex-resident multi-tier application environment may exhibit similar behavior and have similar issues.

- **Local connectors (1)**
  - Optimized high-speed path (based on local services, such as cross-memory services and RRS)
  - Concern - what happens if local target is not available
    - No automatic switch to alternate target on another LPAR
    - WAS transactions may complete fast causing WLM to prefer that LPAR for increased workload (storm-drain issue)
- **Remote connectors (2 and 3)**
  - Uses TCP/IP for communication
  - Sysplex Distributor (or other load balancer) selects a target among any available targets in the Sysplex
  - If target is local and Sysplex Distributor is remote, communication path is not efficient (2)
  - It is not today possible to favor a local target even if one exists and has capacity

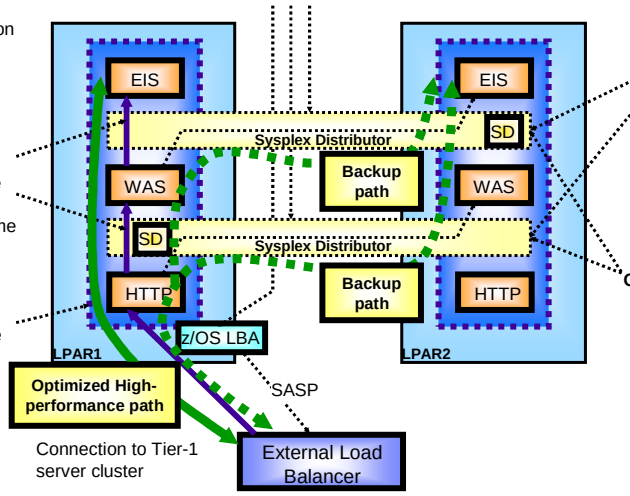
Availability?

End-to-end performance?

Improved multi-tier application support by Sysplex Distributor - optimized for local performance without losing availability

Application endpoint awareness via enhanced Sysplex sockets API processing
- Avoid authentication overhead
- Avoid data conversions

1 WLM LPAR and server-specific performance weights
2 TCP/IP stack server-specific health weights



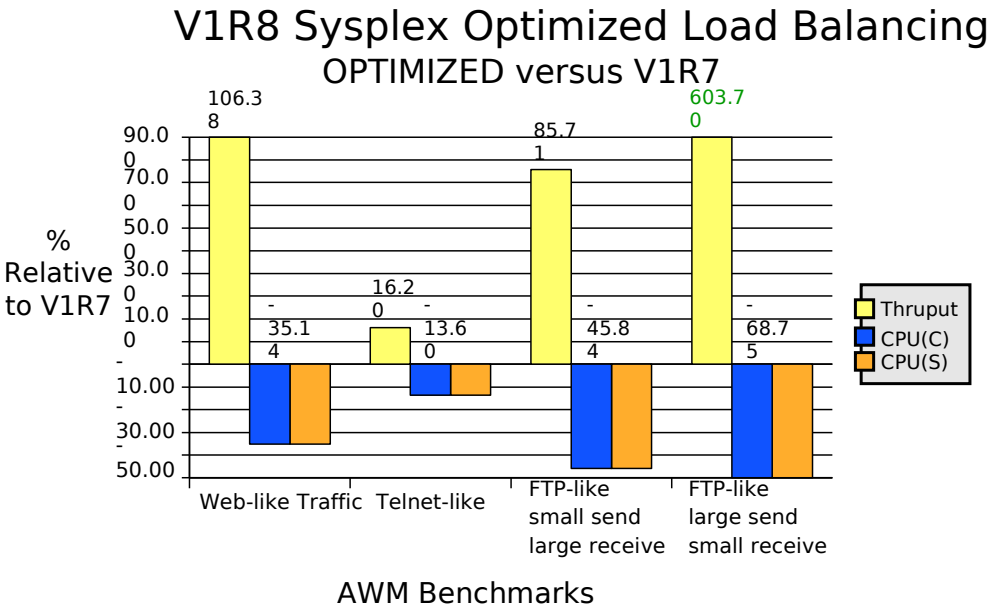
Fast direct local sockets path inside the same "tower" (inside the same TCP/IP stack)

Server instances within same "tower" are **preferred targets**

Level of **local favoritism** can be configured
- Always choose local target if target is available and healthy
- Control level of WLM weight impact on target selection

Optimized traffic flow:
- "Distributed" logic in target stack avoids cross-LPAR flows to SD for connection setup when local target is chosen - **configured**
- Avoids traffic routing via SD-owning LPAR to local targets - **automatic**

Benchmark comparison



Benchmark Comparison Notes

NOTES

- Application Workload Modeler (AWM). A tool for end-to-end performance benchmarking and tuning to measure networking costs associated with workloads. In this instance:
  - It is being used to customize the sizes and patterns of typical Web, Telnet, and FTP workloads.
  - The Application Workload Modeler client and Application Workload Modeler server are running on the same TCP/IP stack.
  - This information is valid for comparison purposes only as this work is running on a dedicated system.
- Types of Traffic being modeled:
  - Web-like traffic - Each transaction:
    - Establish Connection
    - Client Sends 64 bytes
    - Client Receives 8K bytes
    - Close Connection
  - Telnet-like
    - Establish Connection
    - Each transaction:
      - Client sends 200 bytes
      - Client Receives 800 bytes
  - FTP-like (small send, large receive)
    - Establish Connection
    - Each transaction:
      - Client Sends 1 byte,
      - Receives 20M
  - FTP-like (large send, small receive)
    - Establish Connection
    - Each Transaction:
      - Client sends 20M
      - Client receives 1 byte

Load balancing optimization - configuration control

- **A new keyword, OPTLOCAL, is introduced on the VIPADISTRIBUTE configuration statement to cause the client to bypass sending the connection request to the distributing stack.**
  - Three sets of values on the OPTLOCAL statement influence the conditions in which the connection will remain local.
    - A value of **0** indicates that the connection should always remain local.
    - A value of **1** indicates that the connection should remain local unless the server's WLM weight is zero.
    - Values of **2-16** are used as multipliers to increase the local servers WLM weight to favor the local stack.
  - Regardless of the value specified, the connection will always be sent to the distributor if any of the following are true:
    - No server application is available on the local stack
    - Server Efficiency Fraction (SEF) value on the local stack is less than 75
    - The health indicator for the local stack is less than 75
    - The abnormal transactions count for the local stack is greater than 250

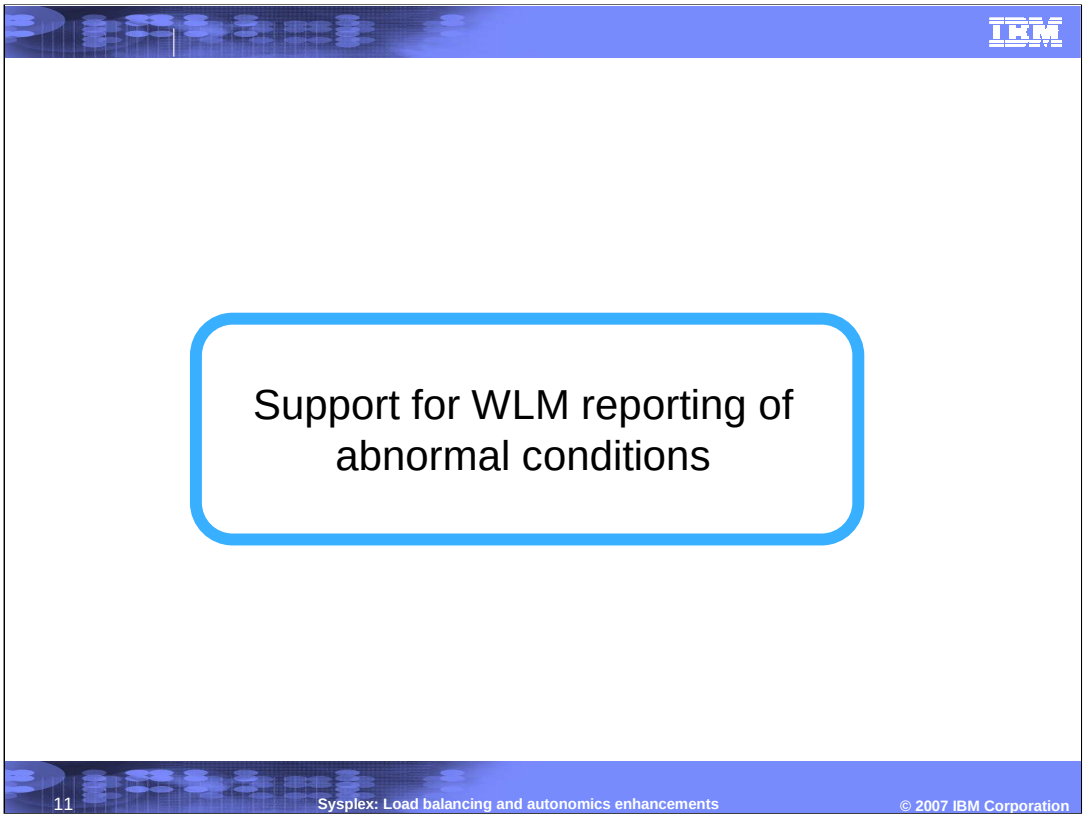
OPTLOCAL

NOTES

```
|--+-------------------+-------------------------------------|
   +-| VIPADISTribute |-+
   +-| VIPAROUTE |------+
   '-| VIPASMparms |----'

VIPADISTribute:
|--VIPADISTribute--| Options |--+-ipv4_addr-----+--------------->
                                '-ipv6_intfname-'

>--+---------------+--DESTIP--+-ALL----------+-----------------|
   |       .------. |          | .----------. |
   |       V      | |          | V          | |
   '-PORT----num-+-'          '---dynxcfip-+-'

VIPAROUTE:
|--+------------------+--dynxcfip--target_ipaddr---------------|
   | .-DEFINE-.       |
   '-+--------+-DELEte-'

Options (These can be specified in any order):

   .-DEFINE-.                  .-TIMEDAFFinity 0-------.
|--+--------+--+--------------+--+----------------------+------>
   '-DELEte-'  '-SYSPLEXPorts-'  '-TIMEDAFFinity seconds-'

   .-DISTMethod BASEWLM---------. .-NOOPTLOCAL-----------.
>--+---------------------------+-+----------------------+-------------|
   '-DISTMethod--+-ROUNDROBIN-+-' |          .-1------.  |
                 '-SERVERWLM--'   '-OPTLOCAL +--------+--'
                                             'value   '
```

Things to think about

- **The data path optimization on a V1R8 Target stack will be performed even if the Distributing stack is downlevel.**
- **The load balancing optimization will be available if both a Target stack and the Distributing stack are running at a V1R8 level (even if other Target stacks are downlevel).**

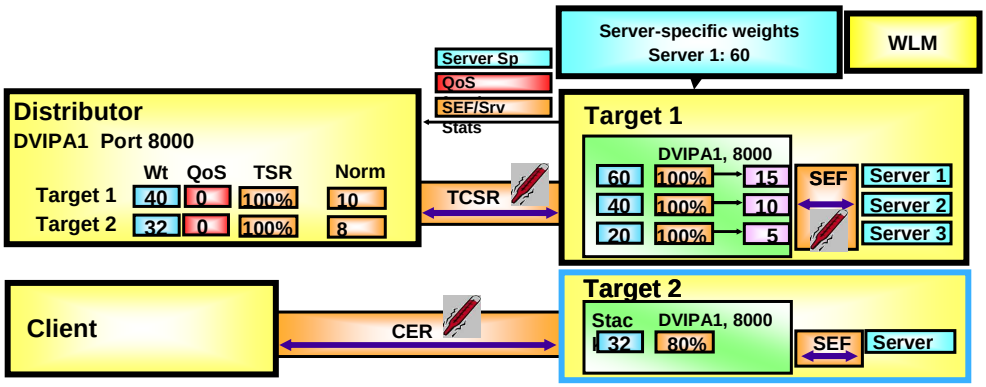

Support for WLM reporting of abnormal conditions
11
Sysplex: Load balancing and autonomics enhancements
© 2007 IBM Corporation

Background information - workload distribution

- **System weight - WLM provides a system weight for each target based on:**
  - Comparison of available capacity
  - Comparison of displaceable capacity (lower importance work that can be displaced)
- **System weights do not reflect**
  - How well the server is performing
- **Server-Specific weights were added in z/OS V1R7 - WLM provides a weight for each server based on:**
  - How well each server is meeting the goals of its service class.
  - Comparison of displaceable capacity on each system based on the importance of the server's work
- **You decide if Sysplex Distributor or the z/OS Load Balancing Advisor should use WLM system weights or server-specific weights per load-balanced application cluster:**
  - System weights: BASEWLM
  - Server-specific weights: SERVERWLM

Background information: Sysplex Distributor adjustment to the WLM weights based on server responsiveness (from an IP perspective)

- **The Sysplex Distributor can reduce the WLM weights by using:**
  - **TSR**: Target Server Responsiveness fraction, which is a compound health-metric per target server (range from 0 (bad) to 100 (good)):
    - **TCSR**: Target Connectivity Success Rate. Connectivity between the distributing stack and the target stack - are the new connection requests reaching the target? (0 is bad, 100 is good)
    - **CER**: Connection Establishment Rate. Network connectivity between Server and client - are new connections being established? (0 is bad, 100 is good)
    - **SEF**: Server accept Efficiency Fraction. Target Server accept efficiency - is the server accepting new work? (0 is bad, 100 is good)
    - **QoS**: QoS fractions. Taking retransmits and packet loss into consideration. (0 is good, 100 is bad)



Background notes

NOTES

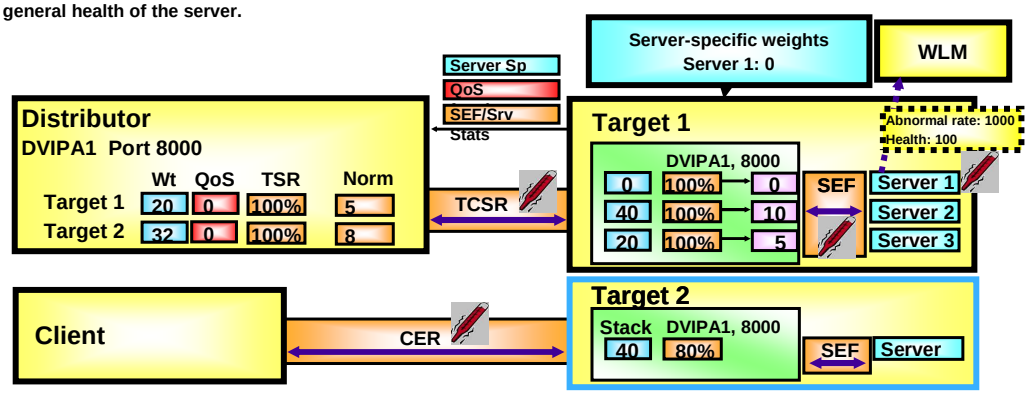
- WLM weights
  - When determining a **System weight**, WLM assigns a relative weight to each system in the sysplex with the highest weight going to the system with the most available CPU capacity. The weights range between 0 & 64. If all systems in the sysplex are running at or near 100% utilization, WLM will assign the highest weights to the systems with the largest amounts of lower importance work. In this way, new connection requests will be distributed to the systems with the highest displaceable capacity.
    - This method does not reflect how well the server application is actually meeting the goals of its service class
    - If all systems are using close to 100% of capacity, then the WLM weight is based on a comparison of displaceable capacity - the amount of lower importance work on each system, but if the service class of the server is of low importance then it may not be able to displace this work.
  - When determining a **Server-specific weight**, WLM assigns a relative weight to each server based on how each server is meeting the goals of its service class. The weights range between 0 & 64. If all systems in the sysplex are running at or near 100% utilization, WLM will assign the highest weights to the servers running on systems with the largest amounts of work that can be displaced by that server (based on the importance of its service class)
- Calculation of TSR
  - SEF - This value is based on whether the server is processing new connections
    - New connections are being established - Connection Establishment Rate (CER)
    - The server is accepting the new connections
  - TCSR - The distributor determines this value from the number SYNs it has sent to the target and the statistics returned from the target
  - TSR - Based on SEF value (which includes CER) and TCSR value

Storm drain problems

- **WLM is not aware of all problems experienced by load balancing target applications or application subsystems:**
  - The server application needs a resource such as a database, but the resource is unavailable
  - The server application is failing most of the transactions routed to it because of internal processing problems
  - The server application acts as a transaction router for other back-end applications on other system(s), but the path to the back-end application is unavailable
- **In each of these scenarios, the server appears to be completing the transactions quickly (using little CPU capacity) when they are actually being failed**
- **This is one of the storm drain problems**
  - The server is favored by WLM since it is using very little CPU capacity
  - As workloads increase, the server is favored more and more over other servers
  - All this work goes "down the drain"

WLM abnormal transaction rate and server-perceived health

- **WLM provides an interface which allows a server application (or subsystem) to pass additional information to WLM about its overall well-being:**
  - ***Abnormal transaction completion Rate***
    - Applications such as the CICS Transaction Server for z/OS, that act as Subsystem Work Managers, can report an abnormal transaction completion rate to WLM (abnormal completions per 1000 transactions). The value is between 0 and 1000 with 0 meaning no abnormal completions.
  - ***General health of the application***
    - Applications can report their general health to WLM. The value is between 0 and 100 with 100 meaning that a server has no general health problems (100% healthy).
- **WLM will reduce the reported server-specific weight based on abnormal completion rate and the perceived general health of the server.**



Netstat reports can show all the elements that influence the final normalized weights used by Sysplex Distributor

- **Netstat VDPT detail sample:**

```
MVS TCP/IP NETSTAT CS V1R8       TCPIP Name: TCPCS
Dynamic VIPA Destination Port Table:
Dest:       201.2.10.11..8000
  DestXCF:   193.9.200.1
  TotalConn: 0000000050  Rdy: 001  WLM: 00  TSR: 100
  Flg: ServerWLM Local
  TCSR: 100 CER: 100 SEF: 100
  Abnorm: 1000      Health: 100
  ActConn:      00000042
  QoSPlcAct:    *DEFAULT*
    W/Q:  0
  QosPlcAct:    Gold-Service
    W/Q:  0
Dest:       201.2.10.11..8000
  DestXCF:   193.9.200.2
  TotalConn: 0000000050  Rdy: 001  WLM:  4  TSR: 100
  Flg: ServerWLM  Local
  TCSR: 100 CER: 100 SEF: 100
  Abnorm: 0000      Health: 100
  ActConn:      00000042
  QoSPlcAct:    *DEFAULT*
    W/Q: 15
  QoSPlcAct:    Gold-Service
    W/Q: 15
```

The normalized WLM weight (between 0 and 16).

Target Server Responsiveness fraction - the compound health metric from TCP/IP's perspective (0 bad - 100 good).

Local: Indicates that the target stack specified by the DestXCF Addr value is currently processing outbound connections for this destination and port pair locally.

Target Connectivity Success Rate (TCSR), Connection Establishment Rate (CER), and Server Efficiency Fraction (SEF) - (0 bad - 100 good).

Abnormal completion rate and server-perceived health as reported through WLM.

ActConn is current number of active connections.

QoS fraction per policy. (0 good - 100 bad).

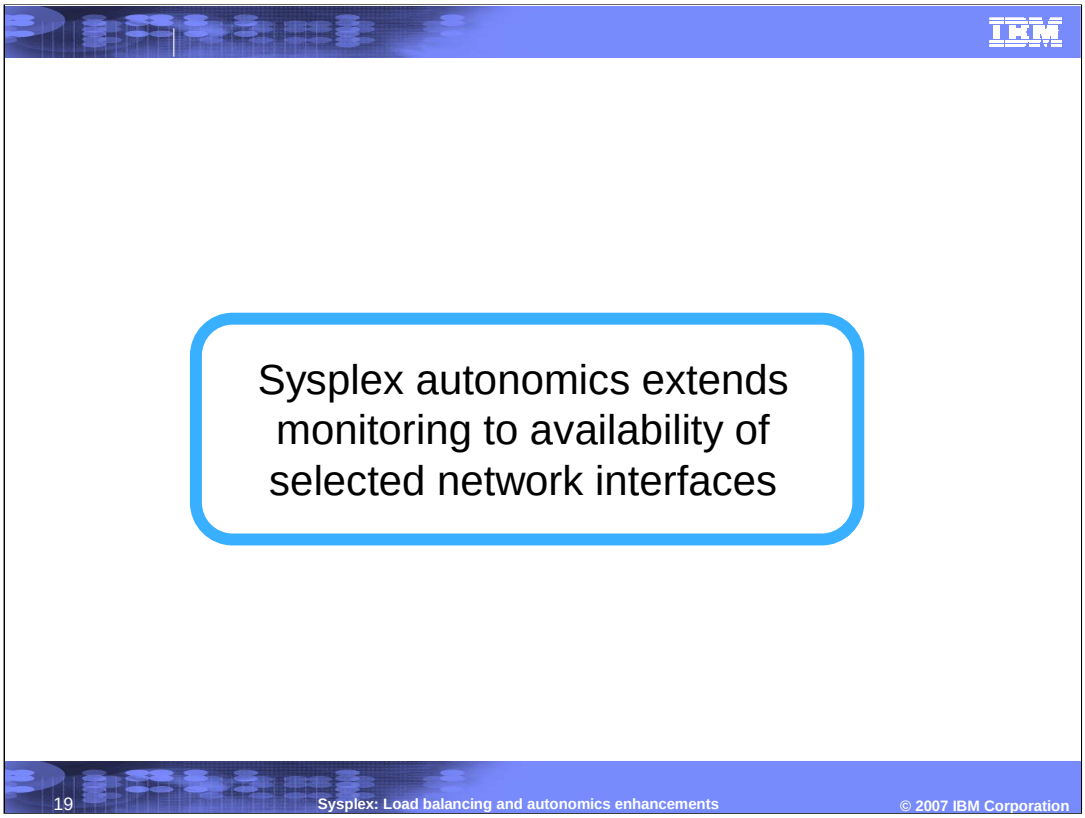
Things to think about

- **Server application health values will be used by WLM to modify the reported Server-specific weights if the target system that the application is running on is V1R8 or later.**

- **If the Sysplex distributor is a release earlier than V1R8**
  - It receives any weights that may have been modified by WLM
  - But will not be able to display server application health values in the Sysplex Netstat VDPT DETAIL display

- **If the Sysplex distributor is V1R8**
  - Default values of 0 abnormal completions and health of 100 will be shown if the target stack is not at least V1R8.

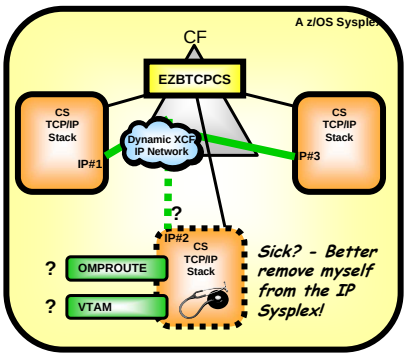
- **The Load Balancing Advisor & Load Balancing Agent**
  - Forwards server availability and WLM weights to external load balancers
  - Server application health values will only displayed if both the Load Balancing Advisor and the Server's Load Balancing Agent are V1R8.


IBM
Sysplex autonomics extends monitoring to availability of selected network interfaces
19
Sysplex: Load balancing and autonomics enhancements
© 2007 IBM Corporation

Background information: TCP/IP Sysplex autonomics - proactive monitoring with automated recovery actions at a stack level

- **TCP/IP Sysplex recovery functions to protect against major hardware and software failures are triggered when a TCP/IP stack leaves the TCP/IP XCF messaging group, which prior to z/OS V1R6 only would occur when a stack terminated:**
  - If the leaving stack was a DVIPA owner, a backup stack will take over the DVIPA along with any associated Sysplex Distributor responsibilities - and new workload will continue to be processed by the Sysplex
  - If the leaving stack was a target stack for distributed workload, the distributing stack will remove it from its list of candidate target stacks - stop sending more connections to it
- **If a TCP/IP stack doesn't terminate, but enters an "unresponsive" condition, recovery functions are not triggered**
  - If the unresponsive stack is a Sysplex Distributor stack, no new connections to the distributed application will be processed and routing of inbound data through the distributing stack to target stacks for existing connections will cease
  - If the unresponsive stack is a target stack, the distributing stack will continue to send new connections to it and since WLM may see the target stack as lightly loaded, that stack may even be seen as a preferred stack for new workload - sending even more workload down the drain
- **There are a few known error conditions that can cause TCP/IP to become unresponsive or appear to be hanging - without actually terminating:**
  - The downstream network lost visibility of the distributing stack due to an OMPROUTE outage or malfunction and the network routers do not know how to reach the destination DVIPA addresses
  - VTAM is malfunctioning, data link control services are not working properly, and IP packets cannot be received or sent
  - TCP/IP is in a critical storage constraint situation
  - XCF IP network connectivity (Dynamic XCF) between the distributing stack and the target stacks is not functioning
  - Abends/errors in the TCP/IP Sysplex code components
  - All downstream network interfaces are malfunctioning and the stack is unreachable from interfaces other than the intra-Sysplex links

Background information: TCP/IP Sysplex autonomics phase I - overview (at a z/OS V1R6 level)



The assumption is that if a TCP/IP stack determines it can no longer perform its Sysplex functions correctly, it is better for it to leave the TCP/IP XCF group and by doing so, signal the other TCP/IP stacks in the Sysplex that they are to initiate whatever recovery actions have been defined, such as moving dynamic VIPA addresses or removing application instances from distributed application groups.

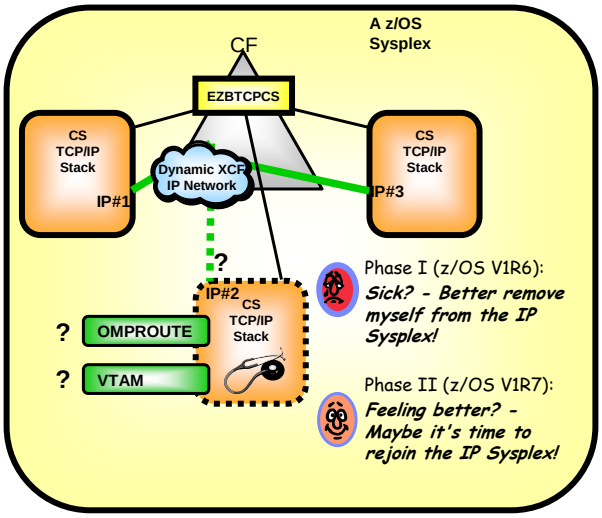
- **Autonomic functions to reduce single point of failure for distributed applications in a Sysplex**
  - Monitor CS health indicators
    - Storage usage - CSM, TCPIP Private & ECSA
  - Monitor dependent networking functions
    - OMPROUTE availability
    - VTAM availability
    - XCF links available
  - Monitor Communications Server component-specific functions
    - Selected vital internal components
- **Monitors determine if this TCPIP stack will remove itself from the Sysplex (disconnect from the TCP/IP XCF messaging group) and allow a healthy backup to take ownership of the Sysplex duties (own DVIPAs, distribute workload, etc.)**
- **Monitoring is always done, but configuration controls in the TCPIP Profile determine if the TCPIP stack will remove itself from the Sysplex automatically**
  - Messages will always be sent to the MVS™ console if a condition that would cause the stack to leave the Sysplex is encountered
  - Operator commands can be used to instruct the stack to leave the Sysplex

Messages are always issued to the console when these conditions are detected regardless of SYSPLEXMONITOR Recovery specification
Messages are eventual action (deleted when the action is taken or problem is resolved)

New operator command is provided to allow TCPIP to leave the sysplex (ie. EZBTCPCS xcf group)
Vary TCPIP,,SYSPLEX,LEAVEGROUP

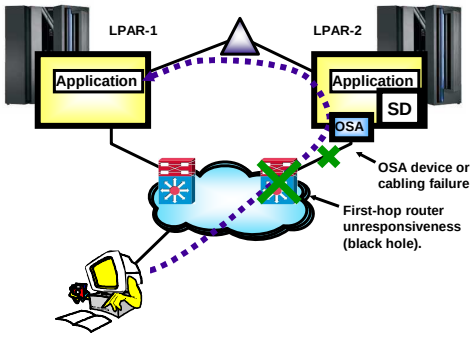
To have TCPIP rejoin the sysplex group, a Vary Obey of the TCPIP profile with sysplex configuration statements is needed.
Severe problems may require a TCPIP stack restart

Background Information: TCP/IP Sysplex autonomics phase II - new functions added in z/OS V1R7



- **z/OS V1R7 added the following functions to the TCP/IP Sysplex autonomics:**
  - Retain the current Sysplex configuration data in an inactive state when a stack leaves the Sysplex
  - Reactivate the currently inactive Sysplex configuration when a stack rejoins the Sysplex
  - New options for rejoining the Sysplex:
    - Via an operator command
    - Automatically when the error condition that caused the stack to leave the Sysplex has been cleared

TCP/IP Sysplex autonomics phase III - new functions added in z/OS V1R8: Compensating for "less-than-perfect" network design !!



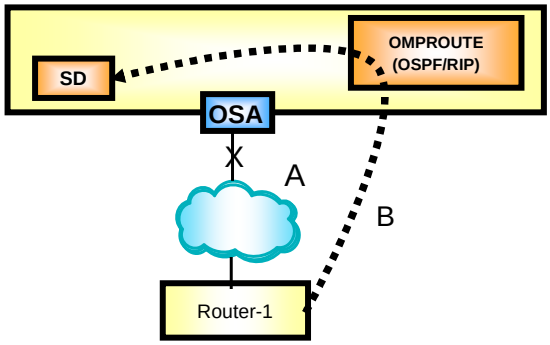
Network design with only a single downstream network interface from a z/OS LPAR in a z/OS Sysplex is obviously not a recommended configuration!

- **Assume that DynamicXCF is not an OSPF interface or that we have disabled routing through z/OS LPARs in general (NODATAGRAMFWD):**
  - Assume also that the downstream nodes cannot reach the SD node (LPAR-2):
    - OSA device or cabling failure
    - First hop router (downstream) problems
  - All Sysplex health monitors indicate a healthy environment
    - Dynamic XCF connectivity is working
    - No storage issues
    - VTAM is operational
    - OMPROUTE is operational
    - Target Server Responsiveness Fraction indicates no SD environment health problems
  - But since there is no route from the client into the SD node, the SD functions appear unavailable (or are indeed unavailable from a user perspective!)

Network interface monitoring component of Sysplex autonomics

- **The option of enabling the TCP/IP stack to monitor the status (active or inactive) of selected network interfaces.**
  - Optionally monitor the presence of dynamic routes over the interfaces.
- **A new parameter is provided to identify which links/interfaces are critical for inbound network connectivity for a DRVIPA/DVIPA owning TCP/IP stack.**
  - These are the ones that will be monitored.
- **The Sysplex Autonomics function will:**
  - If all monitored interfaces become inactive for a configured time interval:
    - Eventual action message (EZD1209E) is issued on the MVS console.
    - Time interval as defined in TIMERSECS on the SYSPLEXMONITOR configuration statement.
  - If monitoring of dynamic routes is specified, and no dynamic routes over monitored interfaces were found for a configured time interval:
    - Eventual action message (EZD1210E) is issued on the MVS console.
  - Additionally, if GLOBALCONFIG SYSPLEXMONITOR RECOVERY option is active it will initiate a recovery action (leave sysplex group).

Two conditions monitored to determine interface availability



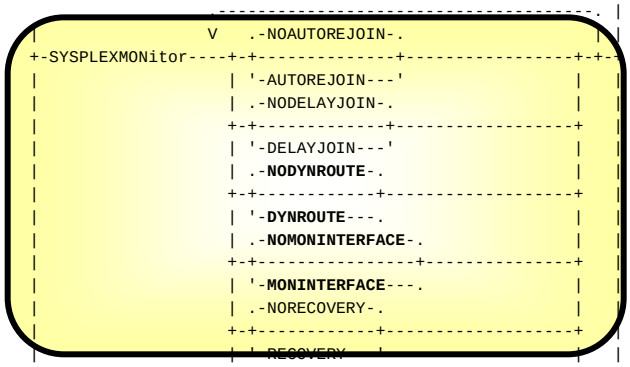
A **Is the monitored interface up?**
- Device layer knows that based on status information from OSA

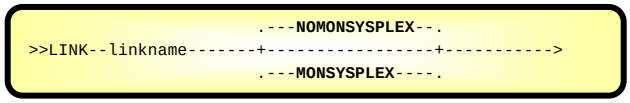
B **Is OMPROUTE receiving route information over that monitored interface from Router-1 (first-hop router)?**
- OMPROUTE provides this route information to the TCP/IP stack
- DELAYJOIN processing when a TCP/IP stack is starting up has also been enhanced with logic to not report OMPROUTE availability until routes have been learned through one or more network interfaces.

How to enable network interface monitoring

- **New options on the GLOBALCONFIG statement for the SYSPLEXMONITOR section:**
  - [NO]MONINTERFACE
    - Monitor network interfaces status (those interfaces that have been configured with the MONSYSPLEX option)
  - [NO]DYNROUTE
    - Additionally monitor for dynamic routes over those monitored network interfaces
- **New option on LINK and INTERFACE statements:**
  - [NO]MONSYSPLEX
    - Monitor this network interface as part of overall Sysplex Autonomics

Only network interfaces with the MONSYSPLEX option will be monitored by Sysplex autonomics.





Things to think about

- **Users should disable monitoring an interface before stopping and deleting a device or interface.**
  - This is to avoid an unexpected recovery action from taking place.

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:

CICS IMS MVS VTAM z/OS

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 or changes in the products or programs 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 (for example, 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.