



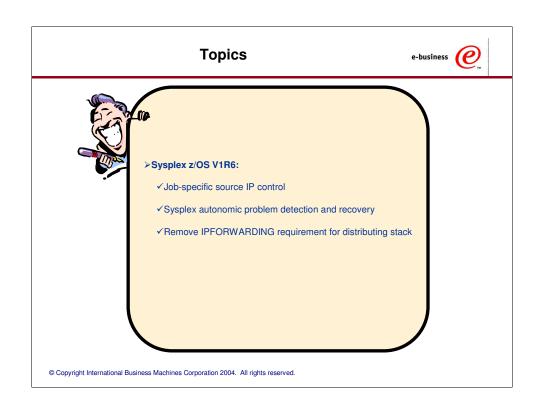
Communications Server z/OS V1R6 Technical Update

Sysplex for z/OS V1R6

© Copyright International Business Machines Corporation 2004. All rights reserved.



© Copyright International Business Machines Corporation 20(m. Business Machines Mach





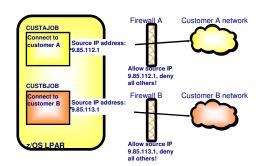
Job-specific source IP addressing



- 1 If TCPSTACKSOURCEVIPA is configured in z/OS V1R4 or later, it applies to all outbound TCP connections
 - $_{\it f}$ Same address for all outbound TCP connections if enabled
 - fNo impact on UDP or RAW sockets
- 2 TCPSTACKSOURCEVIPA only works if no bind() is issued before the connect()
 - f Even if the bind() is to inaddr_any (IP address 0.0.0.0)
 - $_{\it f}$ Job-specific source IP addressing will work for applications that issue an explicit bind() to inaddr_any/unspecified address
 - •If the application issues an explicit bind() to a specific local IP address, Job-specific source IP address configurations will have no impact
- 3 Source IP address control requirements from customers and the SHARE user group:
 - f Single Sysplex IP address, inbound and outbound •TCPSTACKSOURCEVIPA in z/OS V1R4
 - fSingle IP address for an application
 - •Job-specific source IP address in z/OS V1R6

Job-specific source IP address control added in V1R6 for easier firewall filter rule administration





© Copyright International Business Machines Corporation 2004. All rights reserved.

Extending configuration control over which local IP address to use for outbound connections from z/OS

- ✓Outbound connections can use same IP addresses as inbound connections to same application without application change:
 - Fasier for accounting and management Fasier for security (firewall admin)
 Fermits source IP address selection controls for applications even when application doesn't provide for this programmatically (most don't, but some do!)
- ✓Introducing Job-specific Source IP Addressing
 - , A new TCPIP.Profile statement SRCIP/ENDSRCIP allows the selection of a source IP address for outbound TCP connections by job name , Overrides TCPSTACKSOURCEVIPA and SOURCEVIPA specifications

Job-specific source IP address configuration V1R6 SRCIP/ENDSRCIP statements >>SRCIP ---+------+-- ENDSRCIP --+----JOBNAME-- jobname --+-- ipv4_address ----+--+ +-- ipv6_address -----+ +-- ipv6_interface_name+ Example JOBNAME USER15 9.43.242.5 JOBNAME USER* 9.43.242.4 JOBNAME USER15 2EC0::092B:F203 ETHER1 JOBNAME JOB* JOBNAME * 9.43.242.3 ENDSRCIP © Copyright International Business Machines Corporation 2004. All rights reserved.

SRCIP/ENDSRIP



- >JOBNAME Indicates to create a job specific source IP address designation with designated IP address or interface name.
- >jobname The name of the job for which the designated interface should be used as the source IP address when the job initiates outbound TCP connections. The job name may end in an asterisk (*), and the job name of any job executing later which begins with the same characters before the asterisk will match this designation. If several different designations exist, then the actual source IP address used will be determined by the most complete match either an exact match, or the most characters before the asterisk in the job specific source IP address designation. If the jobname "*" is specified and it is a VIPA source IP address, then this provides the same function as the TCPSTACKSOURCEVIPA parameter on the IPCONFIG and IPCONFIG6 statements.
- ▶ipv4_address A dotted-decimal notation for an IPv4 address. The IPv4 address is specified as it is in the TCP profile for the static VIPA, Dynamic VIPA (defined by VIPADEFINE DVIPA or previously activated via bind() or IOCTL SIOCSVIPA within a VIPARANGE) or any real IPv4 IP address. The specified IP address does not need to be defined prior to the processing of the SRCIP statement but it must be defined before the first TCP connect request is issued by the associated job, otherwise the connect request will fail.
- ▶ipv6_address A standard IPv6 notation for an IPv6 address. The IPv6 address is specified as it is in the TCP profile for the static VIPA, Dynamic VIPA (defined by VIPADEFINE DVIPA or previously activated via bind() or IOCTL SIOCSVIPA6 within a VIPARANGE) or any real IPv6 IP address. The specified IP address does not need to be defined prior to the processing of the SRCIP statement but it must be defined before the first TCP connect request is issued by the associated job, otherwise the connect request will fail.
- ➤ ipv6_interface_name An IPv6 interface name. The IPv6 interface name is specified as it is in the TCP profile for the static VIPA, Dynamic VIPA (defined by VIPADEFINE or VIPARANGE) or any real IPv6 interface. The specified interface name does not need to be defined prior to the processing of the SRCIP statement but it must be defined before the first TCP connect request is issued by the associated job, otherwise the connect request will fail.

© Copyright International Business Machines Corporation 2004. All rights reserved.

О Т

Е

S

What is the job name?



- >Applications submitted as batch jobs use the batch job name.
- >TSO users have their user ID as job name.
- >The job name associated with applications started from the MVS operator console using the START command will be determined as follows:
 - , If the START command is issued with the name of a member in a cataloged procedure library (for example, S APP1), the jobname will be the member name (for example, APP1).
 - , If the member name on the START command is qualified by a started task identifier (for example, S APP1.ABC), the jobname will be the started task identifier (for example, ABC). The started task identifier is not visible to all MVS components, but TCP/IP will use it to match a jobname specified in the SRCIP block.
 - , The JOBNAME parameter can also be used on the START command to identify the jobname (for example, S APP1, JOBNAME=XYZ).
- > Applications run from the z/OS UNIX shell normally have a job name that is the logged on user ID plus a one-character suffix. Because this job name may not be predictable, you can use a job name ending in * to ensure these applications are governed by the SRCIP statement. Since different applications may have the same job name, and since job name is the way jobs are identified for designating source IP address for outbound TCP connections, the user will have to be careful in designating job names for applications running from the z/OS shell.
 - Authorized users can run applications from the z/OS shell and use the _BPX_JOBNAME environment variable to set the job name. In this case, the value specified for the environment variable is the job name.
 - $_{\rm f}$ z/OS UNIX applications started by INETD typically use the jobname of the INETD server plus a one-character suffix.

SRCIP/ENDSRIP



- >For successful creation of a job specific source IP address designation, the source IP address does not have to exist on the TCP/IP stack when the SRCIP/ENDSRCIP is processed.
- > When a new SRCIP/ENDSRCIP statement is specified in a configuration data set on a VARY TCPIP,,OBEYFILE command, the new designation will completely replace the existing designation. If a SRCIP/ENDSRCIP statement contains no entries (e.g. no JOBNAME defined), then all existing job specific source IP address designations that were from the previous SRCIP/ENDSRCIP statement are removed.
- Creating and deleting job specific source IP address designations will not in any way affect existing connections.
- ➤ Parsing rules of SRCIP/ENDSRCIP statements:

0

Т

Е

S

- $_f$ Only one SRCIP/ENDSRCIP statement block should appear in a configuration data set. Any subsequent statement blocks will be ignored and an informational message will be issued.
- , If the stack is not IPv6-enabled and an IPv6 IP address or IPv6 interface name is specified, the JOBNAME entry will be ignored and an existing warning message will be issued. (EZZ0695I JOBNAME STATEMENT ON LINE lineno NOT VALID IPv6 SUPPORT IS NOT ENABLED)
- If a error is encountered (e.g. invalid IPv4 address, or invalid IPv6 address, or only jobname is coded, but without any IPv4, or IPv6 address or IPv6 interfacename associated it), a common config error message will be issued and the whole SRCIP/ENDSRCIP block will be ignored.
- f If duplicate jobnames are specified in the same statement, and the duplicate entries specify an IP address in the same IP address family (e.g., both entries specify IPv4 or both specify IPv6 IP addresses) the first JOBNAME entry will be in effect. Any duplicate JOBNAME entries will be ignored and new warning messages will be issued.
- >Please refer to IP Configuration Reference for more information.

Netstat command to display current job-specific source IP address specifications



Report examples for new netstat SRCIP/-J command:

Not IPv6 enabled (SHORT format)

1	MVS TCP/II	P NETSTAT	CS V1R6	TCPIP Name:	TCPCS	20:30:49
I	Job Name	Type	Interface			
ı	USER1	IPv4	9.43.242.1			
ı	USER0*	IPv4	9.43.239.3			
١	JOB3	IPv4	9.67.240.2			

IPv6 enabled or request for LONG format

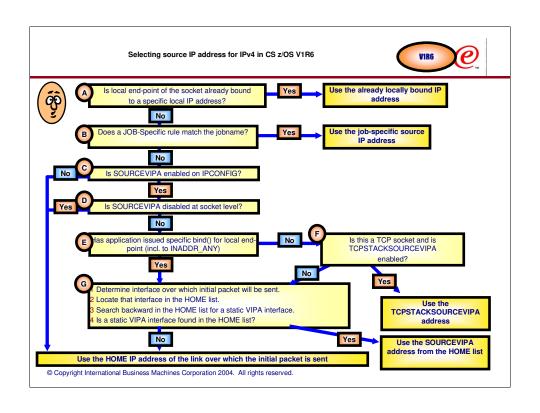
19

Job-specific source IP addressing - details



- Existing functions such as SOURCEVIPA and TCPSTACKSOURCEVIPA will be superceded by SRCIP statements if they are specified and match the jobname.
- ➤The specification of "*" for a jobname in the SRCIP statement will give equivalent functionality as the TCPSTACKSOURCEVIPA with the additional benefit of working with explicit binds.
- ➤ Valid values for TCPSTACKSOURCEVIPA are restricted to:

 fStatic VIPAs
 fActive Dynamic VIPAs





V1R6 Sysplex Autonomics

Copyright International Business Machines Corporation 2004. All rights reserved.



Sysplex Autonomic Overview



>TCP/IP sysplex technologies were originally introduced in OS/390 V2R8

- f Functional enhancements have been shipped in all OS/390 and z/OS releases since then
- Customers use the functions as the foundation for building highly available and scalable TCP/IP solution environments in their z/OS Sysplexes
- Customers have adapted the use of these functions rapidly and they are today in widespread use for business critical workloads in installations of all sizes ranging from small to very large

As these technologies themselves mature and use of them increases, it is time to take a critical look at the operational aspects and identify areas where proactive improvements can be made

- f A large experience base to harvest ideas from for operational improvements
 - •Reduce overall system impact of malfunctioning dependent components such as XCF signaling, or dynamic routing daemon (OMPROUTE) availability, etc.
 - •Improve operator interaction so operators can control the system without modifying the active configuration
- Apply autonomic principles to let the technology itself detect and react to a range of error conditions without operator intervention

Sysplex Distributor as a z/OS Sysplex resident load balancer, has the opportunity to include very detailed, real-time metrics in its decision logic

- f As the z/OS infrastructure improves in the areas of server-specific performance reporting, apply those improved metrics to the TCP/IP load balancing decision algorithms
- f Include connection processing time in the decision algorithms

TCP/IP sysplex functions Overview



OS/390 V2R8

f Initial dynamic VIPA support

OS/390 V2R10

- f Sysplex Distributor based on WLM feedback
- f Non-disruptive dynamic VIPA movement

z/OS V1R2

- f Sysplex Distributor to factor in QoS metrics
- f Sysplex Distributor support for fast connection reset
- f Sysplex Distributor as service manager for Cisco MNLB forwarding agents
- f Use of HiperSockets for same-CEC dynamic XCF connectivity

z/OS V1R4

- f Sysplex-wide SOURCEVIPA support
- f Sysplex-wide security associations (IPSec workload)
- f Sysplex Distributor support for passive mode FTP data connections

7/OS V1R

- f Dynamic Sysplex Distributor ports
- Increase VIPADISTRIBUTE ports from 4 to 64
- f Increase number of supported DVIPAs per stack from 256 to 1024
- f Sysplex Distributor round robin distribution
- ¹ Sysplex Distributor timer-based affinity support
- 1 Let VIPABACKUP stack activate DVIPA during sysplex-wide IPL
- f DynamicXCF support for IPv6 connectivity

TCP/IP sysplex recovery functions before z/OS V1R6

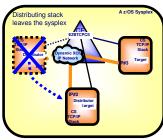


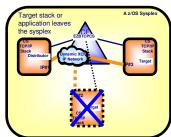
>TCP/IP sysplex recovery functions to protect against major hardware and software failures are triggered when a TCP/IP stack leaves the TCP/IP XCF group (terminates)

- 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
- f 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 a "unresponsive" condition, recovery functions are not triggered

- f 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 seize
- 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





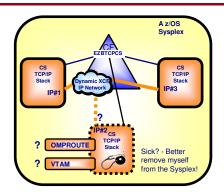
What can cause an unresponsive condition of TCP/IP?



- >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
 - $_{\it f}$ VTAM is malfunctioning, data link control services are not working properly, and IP packets cannot be received or sent
 - fTCP/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
 - f Abends/errors in the TCP/IP sysplex code components

TCP/IP sysplex autonomic problem determination and recovery in z/OS V1R6





The base 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.

>Add autonomic functions to monitor system health:

- ¹ Monitor storage usage (if critical for a *period of time*, initiate recovery action)
 - •CSM, TCP/IP private and ECSA
- f Monitor dependent networking functions (if unavailable for a period of time, initiate recovery action)
 - •OMPROUTE availability and health
 - VTAM availability
 - Dynamic XCF link availability
- Monitor for abends in critical sysplex-related functions
- f Monitor that specific internal functions are executed in a timely fashion (if unresponsive for a period of time, initiate recovery action)
 - •Using independent XCF status monitor component
- ➤ Optionally, delay joining a sysplex during initialization until OMPROUTE is active (already done for VTAM)
- ➤ Provide an MVS operator command to request a TCP/IP stack to leave the sysplex
- © Copyright International Business Machines Corporation 2004. All rights reserved.

TCP/IP Sysplex autonomic Configuration control



- > The built-in monitoring functions determine if this TCP/IP stack will remove itself from the sysplex and allow a healthy backup to take ownership of the sysplex duties (own DVIPAs, distribute workload, host a target application instance, etc.)
- Monitoring is always active, and if a condition is detected that is considered to be an error condition, an MVS console message (eventual action message) will always be issued.
- Configuration options in the TCP/IP profile determine if the TCP/IP stack will remove itself from the sysplex if an error condition is detected:

GLOBALCONFIG
SYSPLEXMONITOR
TIMERSECS seconds
RECOVERY | NORECOVERY
DELAYJOIN | NODELAYJOIN

- > Timersecs used to determine duration of the troubling condition before issuing messages or leaving the sysplex (if RECOVERY is specified) default value is 60 seconds
- >RECOVERY TCP/IP removes itself from the sysplex. Recommended but not the default value.
- ▶NORECOVERY TCP/IP does not remove itself from the sysplex. This is the default value.
- ▶ DELAYJOIN TCP/IP delays joining the sysplex during initialization until OMPROUTE is active.
- >NODELAYJOIN TCP/IP does not delay joining the sysplex.

© Copyright International Business Machines Corporation 2004. All rights reserved

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 TCP/IP to leave the sysplex (i.e., EZBTCPCS XCF group)
Vary TCPIP,,SYSPLEX,LEAVEGROUP

To have TCP/IP rejoin the sysplex group, a Vary Obey of the TCP/IP profile with sysplex configuration statements is needed.

Severe problems may require a TCP/IP stack restart

Actions when a problem is detected



When a problem condition is detected, the following actions occur:

- 1 An MVS console message is always issued
 - a The message is an eventual action message identifying the detected error condition
 - b In cases where the problem is cleared, the message will be DOMed (Deleted) from the MVS console again
- 2 If the RECOVERY option is specified in the SYSPLEXMONITOR configuration and this stack is not the only stack in the TCP/IP sysplex:
 - c The stack leaves the XCF group (EZBTCPCS) message EZZ9676E or EZZ9675E
 - d Depending on the actual problem that was detected, additional documentation is collected
- e All dynamic VIPA resources are removed from the stack's active configuration

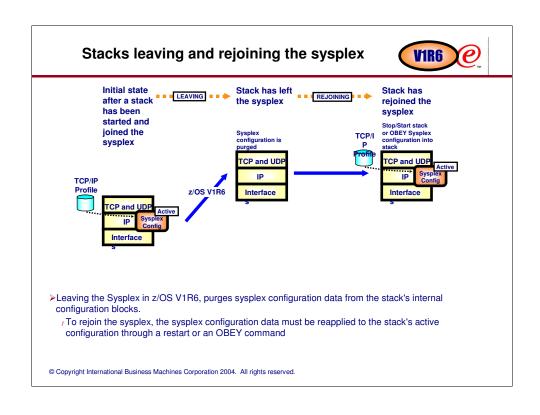
Message examples:

- EZZ9672E tcpstackname DETERMINED THAT OMPROUTE WAS NOT RESPONSIVE FOR AT LEAST timervalue SECONDS
- -EZZ9673E tcpstackname DETERMINED THAT DYNAMIC XCF CONNECTIVITY TO ALL PARTNERS WAS NOT AVAILABLE FOR AT LEAST timervalue SECONDS
- EZZ9679E tcpstackname DETERMINED THAT CSM WAS CRITICAL FOR AT LEAST timervalue SECONDS
- •EZZ9670E tcpstackname SYSPLEX PROCESSING ENCOUNTERED A NONRECOVERABLE ERROR abendreasoncode
- •EZZ9674E tcpstackname SYSPLEX PROCESSING WAS NOT RESPONSIVE FOR AT LEAST timervalue SECONDS

Leaving - rejoining the sysplex



- > When a TCP/IP stack leaves the sysplex (disconnects from the EZBTCPCS XCF group), it will in z/OS V1R6 remove all its internal sysplex related configuration data:
 - f The content of the VIPADYNAMIC block
- DynamicXCF definitions will not be removed and DynamicXCF IP connectivity may in fact remain active even after a stack has left the sysplex
- >To rejoin the sysplex, one of two methods can be used:
 - fStop the stack and start it again
 - •During normal initialization, all sysplex related configuration options in the TCP/IP profile will be processed and normal sysplex initialization will join the stack into the sysplex
 - f Vary OBEY a profile that includes a DynamicXCF statement and/or a VIPADYNAMIC block
 - •If the error condition was temporary and only affected sysplex related functions, while local functions continued to operate normally, a vary OBEY can be used after the sysplex-related error condition was cleared to have the stack rejoin the sysplex
 - •If a stack does not have a VIPADYNAMIC block (could be the case for a target-only stack), a vary OBEY of an IPConfig statement with the DynamicXCF definitions will make the stack rejoin the sysplex
 - •When the stack rejoins the sysplex, it will initiate takeback of any DVIPAs for which it is the primary owner (VIPADEFINEd)



Delay joining the sysplex until the TCP/IP environment is fully ready to do real work



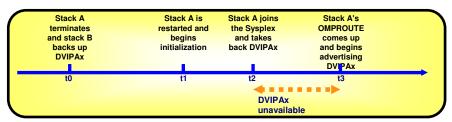
Delay joining the sysplex during stack initialization:

 $_{\it f}$ The case that is being addressed by this function is where a primary stack is restarted and attempts to take back the dynamic VIPA addresses for which it is the primary owner.

If it attempts to do so before its OMPROUTE is up and active, a window may occur where the dynamic VIPA address have been taken back from the backup stack, but hasn't yet been advertised by the restarted stack - resulting in a time window where that address isn't available

•That window would normally be very short (a few seconds), but could be longer depending on local operations procedures

_fThe base idea is that it is better to leave the address with the backup stack until the restarting stack is fully ready to take over its responsibilities



Operator command to request a TCP/IP stack to leave the sysplex



>Operator command to leave the sysplex

f Operator command must be issued on the system where the stack that is to leave the sysplex is running

f Allows an operator initiated recovery of an error condition

- •Monitoring issues messages, but NORECOVERY was configured
- •Recovery from other unforeseen error condition

 $_{\it f}$ For the stack to rejoin the Sysplex, a restart of the stack or an OBEYFILE command is required to redefine all the sysplex-related resources (DVIPAs, Distributor rules, etc.)

```
VARY TCPIP, [stackname], SYSPLEX, LEAVEGROUP
.....
EZZ00531 COMMAND SYSPLEX, LEAVEGROUP COMPLETED SUCCESSFULLY
```

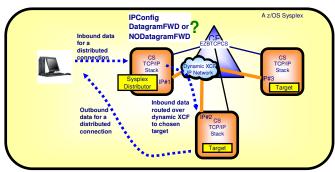
General IP forwarding no longer required for Sysplex Distributor



The distributing TCP/IP stack needs to forward both connection setup and inbound connection data over a dynamic XCF IP network to the chosen TCP/IP target stack in the sysplex.

 $_{\it f}$ Previous to z/OS V1R6 it was a requirement that the distributing stack had to have DATAGRAMFWD enabled

- •This option means that the TCP/IP stack is allowed to route IP packets in general from any interface to any interface (only way to limit this general routing capability was via firewall filters on z/OS)
- f In z/OS V1R6, use of Sysplex Distributor does not require DATAGRAMFWD to be enabled
 - •Sysplex Distributor can now be deployed without any risk of using a z/OS stack as a general intermediate routing node



Trademarks, Copyrights, and Disclaimers e-business (e)



CICS Cloudscape DB2 DB2 Universal Database

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 services of the product of the pro

Information is provided "AS IS" without warranty of any kind. THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY ETHER EXPRESS OF IMPLED. IBM EXPRESSLY DISCLAMIS ANY WARRANTY IS OF MERCHAMTRABLITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT. BIM shall state that the provided of the provided in the provide

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