

Sysplex Distributor and port definitions

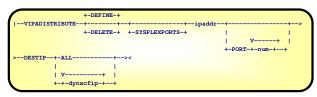


- >The initial implementation of Sysplex Distributor allowed a maximum of four ports per distributed Dynamic VIPA. APAR PQ65205(UQ70170) in z/OS V1R2 raised this limit to 64.
- >However, as large customers begin to tailor their applications for use with Sysplex Distributor, this limitation is becoming too restrictive.
- ➤CS z/OS V1R5 allows applications that bind specifically to an appropriately-configured distributed Dynamic VIPA to have more than four ports the number of ports is not limited by configuration, but rather by other considerations such as available storage.
- ▶To preserve current operation, current TCP/IP configuration works unchanged.
- •If a distributed Dynamic VIPA has a VIPADISTRIBUTE statement configured with ports specified, distribution occurs only for those ports (except as required to handle Passive-Mode (or Firewall-Friendly) FTP).
 •If the PORT parameter is omitted from the VIPADISTRIBUTE statement, and the application instances bind specifically to the distributed DVIPA on a target stack (so that TCP/IP can tell that distribution is requested), then an entry will be created in the DPT for that distributed DVIPA and port and target stack, and that stack and application instance will become a candidate for distribution of connections directed at that distributed DVIPA and port.
- >All TCP/IP stacks that participate in distribution in this manner must be at least V1R5:
 - •Sysplex Distributor routing stack (where VIPADEFINE/VIPADISTRIBUTE) are coded.
 - ·Backup routing stacks.
- ·- and all target stacks.

Dynamic port definitions for Sysplex Distributor - details



The PORT parameter of the VIPADISTRIBUTE statement is made optional:



- > The dynamic VIPA address will still be created as a target DVIPA on the stacks identified by the DESTIP parameter regardless of whether they have started any applications bound to that address yet.
- The Netstat VDPT/-O report now handles an effectively unlimited number of ports for a particular distributed DVIPA, but with IPADDR/-I and PORT/-P filter supports, this should not be a problem. An additional 'Flg' field on each VDPT entry, indicating 'D' for port was created dynamically.

Number of configured DVIPAs per stack raised from 256 to 1024



- ➤A configured Dynamic VIPA is one that is created in any of the following ways, and might or might not be active:
- Using VIPADEFINE
- •Using VIPABACKUP
- •Using an IOCTL SIOCSVIPA DEFINE when this stack had a covering VIPARANGE
- This includes the MODDVIPA utility
- •Using a BIND when this stack had a covering VIPARANGE
- ➤ Expanding the DVIPA limit from 256 to 1024 allows the user more flexibility in defining a network configuration. In association with this increase, many DVIPA associated control blocks have been moved from common to TCP/IP private storage. Thoughtful consideration should still be used when planning workload and the number of dynamic DVIPAs needed.
- >A stack is now limited to no more than 1024 configured or target DVIPAs at any one time.
- > If a pre-V1R5 stack is to back up DVIPAs defined on a V1R5 stack, the DVIPAs to be backed up by the pre-V1R5 stack must be among the first 256 DVIPA definitions on the V1R5 stack.
- >OSA-Express devices have a limit on the number of IP address (both IPV4 and IPV6 addresses) that can be registered to the device.
- •The limit is dependent on the microcode level of the OSA-Express device.
- •This limit applies across all TCPIP stacks that share the OSA-Express device.
- •When defining a large number of VIPAs, users should take care not to exceed this limit.
- If the limit is exceeded, IP addresses beyond the limit will not be registered with the OSA-Express devices, and incoming packets with those IP addresses will not be routed to the correct stack unless that stack is designated as the Primary Router.

OMPROUTE Restriction Notes



>The MTU size defined on OSPF_INTERFACE statements limits the size of advertisements that can be sent or received over that interface. OMPROUTE cannot build an advertisement whose size would exceed the largest MTU size of all its interfaces. Also, OMPROUTE cannot receive an advertisement that is larger than the largest MTU size defined of all its interfaces. In either of these cases, you will see the following message:

, EZZ7967I ADVERTISEMENT DISCARDED, OVERFLOWS BUFFER: LS TYPE x ID x.x.x.x ORG y.y.y.y

>When this happens on an originating host, that host will not be able send Router Link State Advertisements and therefore other hosts will not be able to calculate routes to any destinations (for example, VIPAs) owned by the originating host. OMPROUTE will terminate if it encounters this condition, because if it can't send its router LSA, it is useless as a router. When this happens on a receiving host, that host will not be able to compute routes to any destinations advertised in the discarded LSA. Also note that other OSPF implementations may have similar or stricter limitations, in which case they would be unable to receive or propagate large router LSAs received from OMPROUTE. These scenarios can severely affect network connectivity and routing capability. If large numbers of VIPA interfaces are going to be used, you are well-advised to examine OSPF MTU sizes throughout your network to ensure that large router LSAs can be propagated.

Normally this would not be a problem, as Link State Advertisements seldom exceed their allowed MTU sizes. However, if a large number of VIPA or dynamic VIPA interfaces are defined on a host, this may become a consideration. The size of the router LSA will include 52 bytes for headers, plus the number of bytes required to advertise the host's owned interfaces. The number of bytes required for each interface is:

/VIPA: 12 bytes plus 12 bytes for each VIPA subnet (see example below)

Point to Point: 24 bytes

Point to Multipoint: 12 bytes plus 12 bytes for each neighbor on the interface

/ All other types: 12 bytes

OMPROUTE Restriction Notes (continued)



```
For owned VIPA interfaces, OMPROUTE advertises both host and subnet routes. The size of router LSAs
      required can be minimized by careful subnet planning. For example, if the following definition exists in the
      OMPROUTE configuration file:
        OSPF_Interface
              IP_Address=3.3.3.*
              Name = VIPA1A
              Subnet_Mask=255.255.255.252
              Attaches_To_Area=1.1.1.1
              MTU=1024
              Cost0 = 1;
O
T
     If 101 VIPA interfaces, numbered 3.3.3.1 to 3.3.3.101, are activated then, in addition to the headers and any
      other owned interfaces, OMPROUTE would need 1512 bytes to advertise 126 links in its router LSA: 1 host
      route to each of the VIPAs, plus 33 subnet routes since each subnet contains only four addresses.
      ▶By contrast, if this definition exists in the OMPROUTE configuration file:
        OSPF_Interface
              IP_Address=3.3.3.*
              Name = VIPA1A
              Subnet_Mask=255.255.255.0
              Attaches_To_Area=1.1.1.1
              MTU=1024
              Cost0 = 1;
  © Copyright International Business Machines Corporation 2004. All rights reserved.
```

OMPROUTE Restriction Notes (continued)



➤If the same 100 VIPA interfaces are activated, OMPROUTE would advertise 101 links in its router LSA: 1 host route to each VIPA, plus 1 subnet route since all the VIPAs are in the same subnet. This would only require 1212 bytes to advertise the VIPAs. If the MTU size on the network is 1500, this can make the difference between being able to send or receive a router LSA and not.

>This limitation can further be circumvented by suppressing VIPA host routes (V1R2 and above only) by coding SUBNET=YES on the OSPF_INTERFACE statement for the VIPA interface(s). However, there are limits on when this can be done, as described in the IP Configuration Reference section on the OSPF_INTERFACE statement.

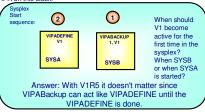
© Copyright International Business Machines Corporation 2004. All rights reserved.

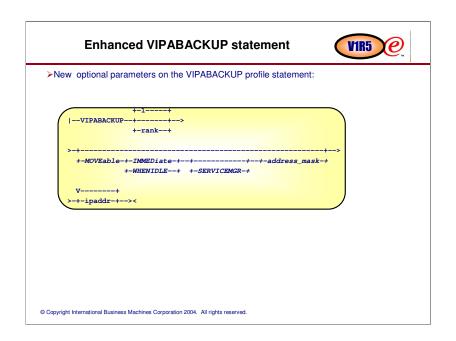
N O T

VIPABACKUP before VIPADEFINE



- >The original design for Dynamic VIPAs envisioned that the stack with the VIPADEFINE would be activated first, but this is not always the case.
- The VIPADEFINE profile statement contains definitions (MOVEABLE state and subnet mask) needed to activate the DVIPA. Currently, these are not present on the VIPABACKUP statement.
- >z/OS V1R5 adds MOVEABLE, SERVICEMGR, and the subnet mask as optional parameters on the VIPABACKUP statement
- _fThe presence of MOVEABLE and a subnet mask designates that the DVIPA may be activated on this stack if it is not active elsewhere in the Sysplex
- /MOVEABLE, a subnet mask, and optionally SERVICEMGR, provide the required information for activation of the DVIPA
- _f If the DVIPA is already active elsewhere in the sysplex, these parameters are ignored and the DVIPA is configured as a backup DVIPA on this stack.





VIPABACKUP - notes



➤MOVEable IMMEDiate

/Specifies that the dynamic VIPA or VIPAs whose address(es) will be specified as ipaddr may be activated on this TCP/IP when the DVIPA is not already active elsewhere in the Sysplex, and that when the TCP/IP where the DVIPA is defined by VIPADEFine is activated, the DVIPA will be activated on that TCP/IP immediately, while preserving TCP connections to this TCP/IP.

, This parameter is used only for activating the DVIPA when it is not already active in the sysplex. If the DVIPA is active when the VIPABACKUP statement is processed, this parameter is ignored.

MOVEable WHENIDLE

0

, Specifies that the dynamic VIPA or VIPAs whose address(es) will be specified as ipaddr may be activated on this TCP/IP when the DVIPA is not already active elsewhere in the Sysplex, and that when the TCP/IP where the DVIPA is defined by VIPADEFine is activated, the DVIPA will remain active on this TCP/IP until there are no more connections to the DVIPA on this TCP/IP.

, This parameter is used only for activating the DVIPA when it is not already active in the sysplex. If the DVIPA is active when the VIPABACKUP statement is processed, this parameter is ignored.

VIPABACKUP - notes



▶SERVICEMGR

, Indicates that Sysplex Distributor performs Multinode Load Balancing (MNLB) by functioning as a Service Manager (in place of Cisco's LocalDirector) for these distributed dynamic VIPAs. SERVICEMGR has no effect if a VIPADISTRIBUTE DEFINE statement does not exist for the VIPA or VIPAs. SERVICEMGR is optional, and may be specified on VIPABACKUP only when MOVEABLE is also specified.

 $_{\rm f}$ This parameter is used only for activating the DVIPA when it is not already active in the sysplex. If the DVIPA is active when the VIPABACKUP statement is processed, this parameter is ignored.

➤address_mask

0

/Specifies the subnet mask or prefix to be used when building the BSDROUTINGPARMS entry for this DVIPA when it is activated. This parameter may only be specified on a VIPABackup statement when MOVEable is also specified, and this parameter is required when MOVEable is specified on a VIPABackup statement. It is specified in standard dotted decimal notation.

, This parameter is used only for activating the DVIPA when it is not already active in the sysplex. If the DVIPA is active when the VIPABACKUP statement is processed, this parameter is ignored.

>Note: If MOVEable is not specified, this parameter will be treated as an IP address rather than as an address mask.



netstat VIPADCFG/-F report - notes Address Mask, Moveable, and SrvMgr display the values specified on the VIPABACKUP profile statement only if the DVIPA is currently active on this stack due to initial activation via VIPABACKUP. The values will be blank for classic VIPABACKUP DVIPAs or if the DVIPA has been activated elsewhere in the sysplex (even if this stack has subsequently done a takeover and the DVIPA is active here) Netstat VIPADCFG/-F display when IPv6 is enabled: MVS TCP/IP onetstat CS V1R5 TCPIP Name: TCPCS2 16:06:10 Dynamic VIPA Information: OTES IpAddr/PrefixLen: 203.1.1.94/24 Rank: 000100 Moveable: Immediate SrvMgr: Yes IpAddr/PrefixLen: 203.1.1.140 Rank: 000100 Moveable: SrvMgr: , Address mask is displayed as a prefix length , PrefixLen, Moveable, and SrvMgr display the values specified on the VIPABACKUP profile statement only if the DVIPA is currently active on this stack due to initial activation via VIPABACKUP. The values will be blank for classic VIPABACKUP DVIPAs or if the DVIPA has been activated elsewhere in the sysplex (even if this stack has subsequently done a takeover and the DVIPA is active here) © Copyright International Business Machines Corporation 2004. All rights reserved.

SNMP MIB variables - notes



➤ SNMP changes for VIPABACKUP DVIPAs

▶ibmMvsDVIPAStatusChange SNMP trap

•ibmMvsDVIPAMoveable and ibmMvsDVIPAServMgrEnabled will reflect actual values when the DVIPA status is active (initial DVIPA activation or takeover)

•currently these MIB objects always contain 'unknown' and 'false' for VIPABACKUP DVIPAs

>ibmMvsDVIPATable

N O T

•the following MIB objects will report current values for VIPABACKUP DVIPAs

-ibmMvsDVIPAMaskType

-ibmMvsDVIPAMaskAddr

-ibmMvsDVIPAMoveable

-ibmMvsDVIPAServMgrEnabled

•For VIPABACKUP DVIPAs in backup status:

–if the DVIPA has not been activated in the sysplex, the values of these MIB objects will be UNKNOWN, NULL or FALSE

-if the DVIPA is active in the sysplex, the values of these MIB objects are those that will be used in case of a takeover

Sysplex Distributor and timed affinity



Timer affinity

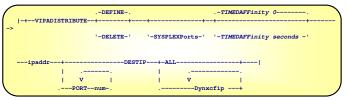
/ Specify whether or not a connection from a client (as identified by source IP address) to a particular server instance of several served by Sysplex Distributor shall establish an affinity for future connections from the same client (IP address) to the same Distributed DVIPA and port(s).

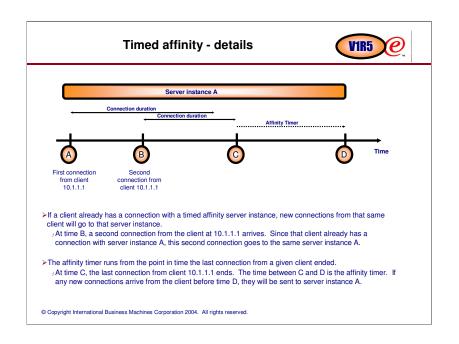
 $_{\it f}$ The affinity extends to a single instance of an application when shareport is used.

, TIMEDAFFINITY 0, the default, means that no affinity will be established when a new connection request is distributed to a particular server application instance by Sysplex Distributor. A non-zero (n seconds - range 1 to 9999) value means that when a new connection from a client is routed to a particular server instance, any subsequent connections will be routed to the same server instance until n seconds have elapsed after the last such connection was closed.

, If a server instance to which clients have affinity is taken out of service, all affinities with that server instance are purged allowing clients to re-connect and establish a new connection and affinity with an available server.

, The parameter can be changed using the VARY OBEY command.





Un-expected effects of affinity



Situations where many different connections from different true client instances all appear with the same source IP address:

- f Proxy applications (such as a Web HTTP server proxy) which initiate secondary connections on behalf of large numbers of different clients. The secondary connections into Sysplex Distributor would all look as though they came from the same client.
- f Firewall Network Address Translation (NAT) which keeps connection tables so it can map multiple client-side addresses into a single server-side source IP address (known as client-NATing)
- f Clients on z/OS or OS390 configured for SOURCEVIPA, or in general many instances of a client on the same network node which use one or a very few IP addresses.

WLM or round robin distribution

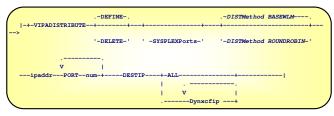


Round robin distribution

r Specify that distribution should be even among available servers at the time the connection request is received, via round robin distribution, or distribution using normal WLM LPAR displaceable capacity

- •BASEWLM (default) preserves current operation using WLM LPAR displaceable capacity If WLM not used the distribution is random (roughly even) and not round robin.
- •ROUNDROBIN causes distribution of future connection requests to be round robin Different stack chosen for each connection in order.

^fThe parameter can be changed using the VARY OBEY command.



Round Robin distribution method - details



- ➤ Note that this distribution method is not influenced by the number of server instances that are active on a target TCP/IP stack instance and listening on the same port (such as SHAREPORT specified on the PORT reservation statement). In other words, a target TCP/IP stack that has multiple active servers on the same port will not receive more connection requests than a target stack that has a single instance of that server active.
- >This option does not have any effect on incoming connection requests that have an active affinity established to a specific server instance (via the TIMEDAFFINITY parameter). When an affinity exists, it has priority over the DISTMethod setting.
- If DISTMethod is changed for an operational Distributed DVIPA, the change will only be applied to future connection requests. Existing connections will not be moved or affected in any way.
- ➤ If DISTMethod is not specified on a VIPADISTribute statement in a VARY OBEY, the state of DISTMethod will not be changed. That is, if DISTMethod was normal WLM, it will stay that way after the VARY OBEY, and if DISTMethod was previously specified as ROUNDROBIN before the VARY OBEY, it will still be ROUNDROBIN after the VARY OBEY.

Single system image (SSI) from an IP e-business perspective in the Sysplex z/OS V1R4 introduced new SysplexPorts Connect to DRVIPA1 capabilities that allow a single Sysplex-wide source VIPA address to be used for outbound Inbound TCP connections by all images in the Sysplex -SSI resulting in single system image capabilities for both inbound and outbound connections. Sysplex /z/OS V1R5 significantly improved the performance of SysplexPorts by allocating z/OS z/OS z/OS ephemeral ports in the CF in blocks of 64. Connect to DRVIPA1 Connect to Connect to Connect to Sysplex ? from ? from ? from SRCVIPA1 SRCVIPA2 **SRCVIPA3** z/OS z/OS We have single system image capability for inbound connections where a single distributed VIPA address can represent all images in the Sysplex - and remote users do not need to select a specific image when connecting to their server application. **Outbound** This function /But if we establish outbound connections from the requires a SSI images in the Sysplex, each image has its own source Coupling VIPA address - so there is no single system image from ? from Facility if the an outbound connection perspective - which has DRVIPA2 source VIPA implications in firewall filter setup, etc. address is Sysplex-wide. © Copyright International Business Machines Corporation 2004. All rights reserved.

When is a source VIPA address being used?





A source VIPA address is used when the following conditions are met:

•The SOURCEVIPA option has been enabled in the IPCONFIG statement

•An outbound connection or UDP association is being established from z/OS

•The application has not bound the local socket to a specific interface IP address before establishing an outbound connection or UDP association

The application has not disabled the use of SOURCEVIPA through a setsockopt call

SOURCEVIPA is not being used when outbound data is sent on a connection that was established inbound to z/OS (such as data sent as a response on a TN3270(E) connection that was established from a remote TN3270(E) client to the TN3270(E) server on z/OS).

An FTP outbound data connection is always established from a socket that was bound to the same server IP address as to where the control connection was directed.

Which source VIPA address is being used?



Two basic rule

•If TCPSTACKSOURCEVIPA is not enabled on the IPCONFIG statement, then the SOURCEVIPA address to use is selected based on the order of the HOME list

```
10.0.0.1 VIPALINK1
192.168.1.1 REALLINNA; Will use 10.0.0.1 as SOURCEVIPA
10.0.0.2 VIPALINK2
192.168.2.1 REALLINNB; Will use 10.0.0.2 as SOURCEVIPA
```

If the connection setup request is sent over REALLINKA, then 10.0.0.1 will be used as source VIPA. If the connection setup request is sent over REALLINKB, then 10.0.0.2 will be used as source VIPA.

- *If TCPSTACKSOURCEVIPA is enabled on the IPCONFIG statement, then the IP address specified as TCPSTACKSOURCEVIPA will be used as source VIPA address for all outbound TCP connections, while UDP and RAW associations will continue to use a source VIPA address based on the order of the HOME list
- >If the IP address that is specified on the TCPSTACKSOURCEVIPA option is used on multiple LPARs in the Sysplex, then the address must be defined as a VIPADISTRIBUTE IP address and the SYSPLEXPORT option should (highly recommended) be specified on the VIPADISTRIBUTE statement.

```
VIPADYNAMIC

VIPADEFINE 255.255.255.192 201.2.10.11 201.2.10.12

VIPADISTRIBUTE DEFINE SYSPLEXPORTS 201.2.10.11 PORT 9999

DESTIP 201.3.10.10 201.3.10.11

VIPABACKUP 100 201.2.10.13

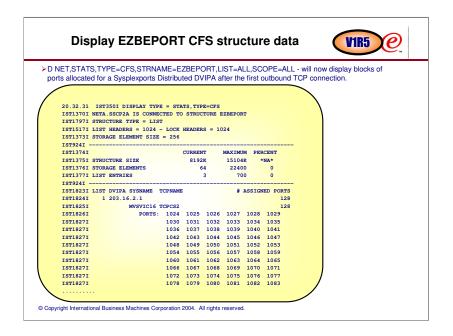
ENDIVIPADYNAMIC

IPCONFIG ... TCPSTACKSOURCEVIPA 201.2.10.11 ...
```

Sysplex port performance enhancements Reserve ports in 128-port blocks



- ➤ Sysplexports will now be allocated in blocks of 64 at a time per Sysplexports Distributed DVIPA from the coupling facility.
- >Ports from the block are then managed by the TCP stack.
- >The block of ports will be returned to the coupling facility after all ports within the block have been used.
- A new block of 64 ports will then be allocated on the next outbound TCP connection request with the Sysplexports Distributed DVIPA as the source IP address.
- >The block of 64 ports will not necessarily be in contiguous order but will range from low to high.



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 bipographical errors. IBM may make improvements and/or changes in the product(s) and/or program(s) described herein at any time tellulor could be an expected program of the product of the program of the product of the program of the product of the program of the progra

Information is provided "AS IS" without warrantly of any kind. THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR MAIN ED. IBM EXPRESS.** DISCLAMAS ANY WARRANTY SO THE MEXICAN PARTICULAR PURPOSE ON NO METER PROVIDED IN THIS DOCUMENT. BY MAIN EXPRESS OR MAIN EXPRESS OR A PARTICULAR PURPOSE ON NO METER PROVIDED IN THIS DOCUMENT. BY MAIN EXPRESS OR A PARTICULAR PURPOSE ON NO METER PROVIDED IN THIS DOCUMENT. BY MAIN EXPRESS OR A PARTICULAR PURPOSE ON NO METER PURPOSE ON N

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 BM benchmads in a controlled environment. All customer examples described are presented as illustrations of how those customers have used BM production and the results they may have solved. The actual throughout or performance that any user is increased that the production and the results they may have solved. The actual throughout or performance that any user is in reliable to the production and the workload processed. Therefore, no assurance on the given that are included user user flately experience they workload processed. Therefore, no assurance can be given that are included user user flately experience performance improvements equivalent to the ratics stated how the production of the

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

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