z/OS 3.1 IBM Education Assistant

Solution Name: Persistent Quiesce/Resume support for Sysplex Distributor targets

Solution Element(s): z/OS Communications Server

July 2023

Agenda

- Trademarks
- Objectives
- Overview
- Usage & Invocation
- Interactions & Dependencies
- Upgrade & Coexistence Considerations
- Installation & ConfigurationSummary
- Appendix

Trademarks

- See url http://www.ibm.com/legal/copytrade.shtml for a list of trademarks.
- Additional Trademarks:
  - None

Objectives

Provide a high-level overview of the following Communications Server function in z/OS 3.1:

- Persistent QUIESCE/RESUME support for Sysplex Distributor targets

Overview

- Who
  - Users who need to quiesce sysplex distributed workload during system startup or for system maintenance in a persistent manner.
- What
  - Sysplex distribution for specific distributed DVIPA and port can be quiesced across the entire sysplex.
- Wow
  - The inability to persistently quiesce sysplex distribution of new connections for specific workloads (distributed DVIPA and port) across the entire sysplex is lifted. Additionally, the quiesced state will persist across events such as:
    - The application server being stopped and restarted
    - The server stack or system LPAR being stopped and restarted
    - The distributed DVIPA moving – Takeover and/or takeback

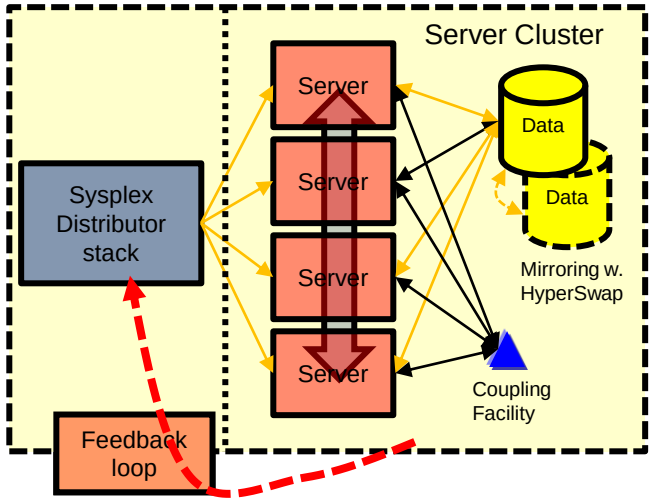
Overview

Background information – Sysplex Distributor Objective

What are the main objectives and advantages of Sysplex Distributor network workload balancing?

- **Performance**
  - Workload management across a cluster of server instances
  - One server instance on one hardware node may not be sufficient to handle all the workload
- **Availability**
  - As long as one server instance is up-and-running, the “service” is available
  - Individual server instances and associated hardware components may fail without impacting overall availability
- **Capacity management / horizontal growth**
  - Transparently add/remove server instances and/or hardware nodes to/from the pool of servers in the cluster
- **Single System Image**
  - Give users one target hostname to direct requests to
  - Number of and location of server instances is transparent to the user

All server instances must be able to provide the same basic service



In order for the load balancing decision maker to meet those objectives, it must be capable of obtaining feedback dynamically, such as server instance availability, capacity, performance, and overall health.

Overview Background information – Sysplex Distributor

- Sysplex Distributor on z/OS
  - Provides the ability to load balance connections to same-servers on multiple z/OS target stacks
    - Incorporates various health and load balancing metrics in target selection
    - Target stacks inform Distributor stack of server availability
      - When server(s) is / is not available for connections
    - Distribution is at the Distributed Dynamic Virtual IP Addr (DRVIPA) and Port granularity
      - Configured on Sysplex Distributor stack – Distributing stack
        - The distributing stack can also be a target stack
      - One or more stacks can be configured as backup stacks to the DRVIPA
        - Can inherit VIPADISTRIBUTE config from distributing stack
      - One or more stacks can be configured as target stacks for the workload
        - A target stack can also be a backup stack

Overview Background information – VIPADISTRIBUTE config

- Sample TCP configuration on Sysplex Distributor stack:

```
IPCONFIG SYSPLEXROUTING DYNAMICXCF 193.9.200.1 255.255.255.0 1
…
VIPADYNAMIC
VIPADEFINE 255.255.255.0 10.91.1.1

VIPADISTRIBUTE DISTMETHOD SERVERWLM OPTLOCAL
          10.91.1.1 PORT 9999 8888 DESTIP ALL
ENDVIPADYNAMIC
```

- Sample TCP configuration on Sysplex Distributor Backup stack:

```
IPCONFIG SYSPLEXROUTING DYNAMICXCF 193.9.200.2 255.255.255.0 1
…
VIPADYNAMIC
VIPABACKUP 255.255.255.0 10.91.1.1
ENDVIPADYNAMIC
```

- Sample TCP configuration on Sysplex Distributor Target stack:

```
IPCONFIG SYSPLEXROUTING DYNAMICXCF 193.9.200.3 255.255.255.0 1
```

Overview Background information – Stopping Sysplex Distribution using config

- VIPADISTRIBUTE DELETE ipaddr PORT port_num DESTIP ALL
  - On distributing stack:
    - All sysplex distributing infrastructure deleted for ipaddr and port_num
    - Stops distributing new connections to target stacks
      - However, new connections can still be serviced by distributing stack (DVIPA still active)
  - On target stacks:
    - Target DVIPA will be deleted after existing connections terminate
  - On backup stacks:
    - Distributing stack's VIPADISTRIBUTE information deleted
      - Backup stack could have its own VIPADISTRIBUTE that would be used if takeover occurs
        - Therefore, sysplex distribution could resume subsequent to a takeover

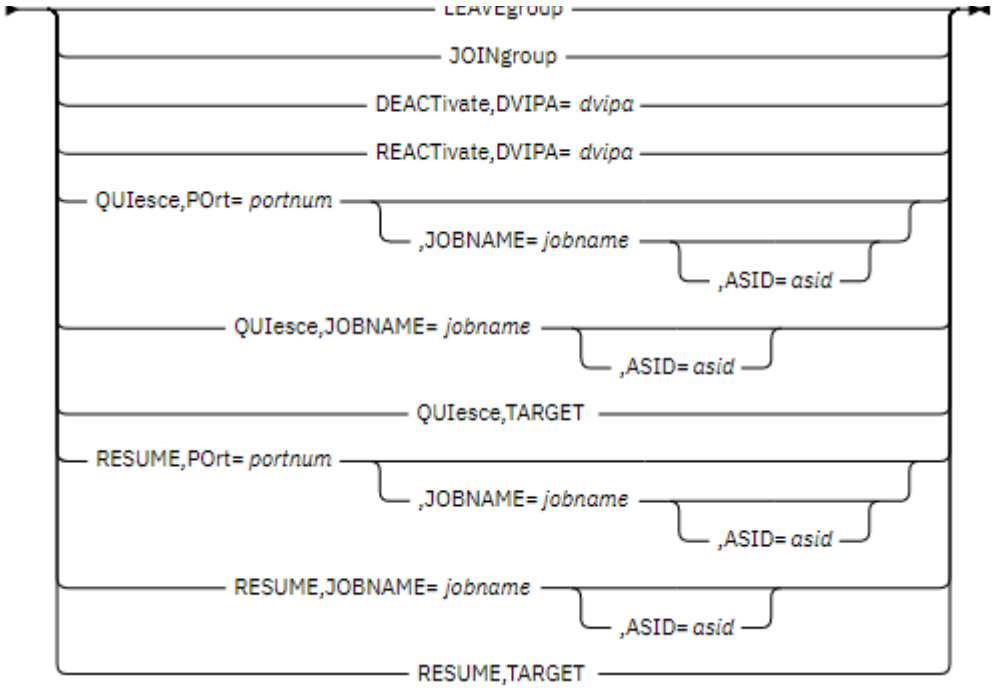
Overview Background information – Stopping Sysplex Distribution for specific target

VARY TCPIP,,SYSPLEX,QUIESCE,PORT

- On target stack:
  - Command to stop sysplex distribution of new connections to specified port
    - Informs distributing stack that this application on this target stack should not receive new connections – existing connections are unaffected
    - A SYSPLEX,RESUME command can be issued to reactivate the application for sysplex distribution
    - However, the quiesce status is **not** preserved across the application being restarted
      - Stopping and restarting the application or restart the target stack

Overview Background information – SYSPLEX command options

VARY TCPIP,,SYSPLEX command options



Overview Problem statement: Preserve QUIESCE state across the sysplex

- There exists no way to preserve a quiesced sysplex distribution status for a particular application workload over:
  - All instances of an application across the sysplex
  - Application stop and restart
  - Target stack stop and restart
- Customer requires a way to quiesce one or more sysplex distributed workloads until they take an overt action to resume such workloads
  - The quiesce state must be maintained across the 3 scenarios above

Overview Solution: Provide Sysplex-wide distribution state controls - Config

- New VIPADISTRIBUTE optional keyword PAUSE added
  - Added to Base Options
  - Default is PAUSE not specified (like SYSPLEXPORTS keyword)
  - Must be specified on initial VIPADIST statement for a DVIPA/PORT pair
  - If PAUSE configured, inbound connections to specified DVIPA and PORT(s) will be rejected – reset
  - Allows for distributing stack to not distribute specific workload(s) at stack start up
  - Must use new Vary SYSPLEX DISTRESume command to resume
  - Normal sysplex distribution setup processing still occurs
    - Target stack processing unchanged, infrastructure created etc..

Overview Solution: Provide Sysplex-wide distribution state controls – Command

- New Vary TCPIP,,SYSplex command options: DISTPAuse and DISTRESume
  - Valid on sysplex distributing stack only
  - Provides ability to resume or pause distribution across the sysplex for a DVIPA and optionally a PORT
    - A specific, configured PORT, or all ports configured for distribution with that DVIPA
  - Once a command is issued (DISTPAUSE or DISTRESUME) the distributing state will persist across:
    - Distributing and Target stack(s) being restarted
      - Requires at least one VIPABACKUP stack
    - Applications being stopped and restarted
    - Distributed DVIPA takeovers and takebacks
  - Valid on any VIPADISTRIBUTE DVIPA
    - Not just those configured with the new PAUSE keyword

Overview Solution: Impact to new connections

- When distribution for a DVIPA and PORT is paused:
  - New connection requests to that DVIPA and PORT are rejected on the sysplex distributing stack and a reset is initiated
  - Optlocal is disabled on target stacks if configured on the VIPADISTRIBUTE statement
    - On target stacks, locally initiated connections to the DVIPA and PORT will not be allowed to connect to local server
  - TimedAffinity is not honored
    - New connections matching an existing TimedAffinity established connection will be rejected
  - Existing connections are unaffected

Overview Solution: DVIPA takeovers and takebacks

- A change in the distribution status is propagated to all backups for the DRVIPA
  - Communicated via XCF messaging with a backup DPT
    - Backup DPT contains information pertaining to distribution status for affected DVIPA and PORT pairs
  - This allows a backup stack during takeover to assume the last distribution status before the primary went down
  - This allows a primary stack to regain the status after retaking ownership of the DVIPA
    - Current owning backup stack will send updated backup DPT to primary
    - Primary will then assume the current status of distribution for this DVIPA and PORT(s)

Overview Solution: Considerations

- Down level stack
  - A down-level stack will not honor this new function as a distributing stack or as a backup stack
    - A down-level stack may participate as a target-only stack
- No Backup stacks present
  - To preserve the PAUSE state across distributing stacks stopping and restarting, there must be at least one backup stack present
- Multiple VIPADISTRIBUTE statements for same DVIPA and Port must either all contain the PAUSE keyword, or all must exclude it

Usage & Invocation – New configuration option

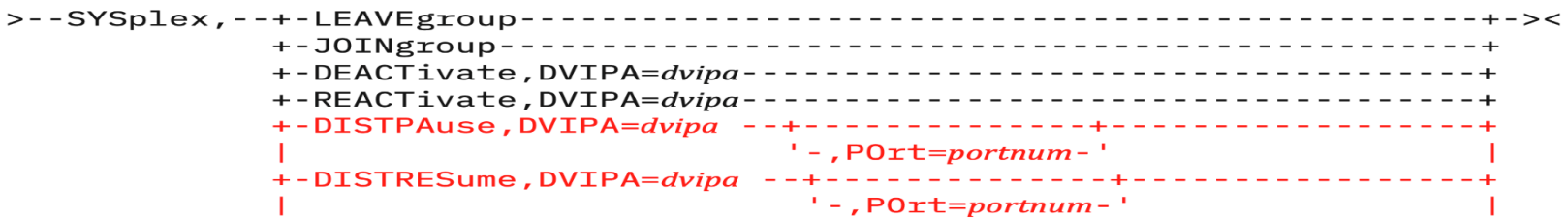
- New keyword PAUSE added to the VIPADISTRIBUTE statement

```
Base Options (These can be specified in any order)

   .-DEFINE-.
|--+--------+-------------------------------------------------->
   '-DELEte-'

   .-DISTMethod BASEWLM -| BASEWLM distribution method options |------------.
>--+-------------------------------------------------------------------------+-->
   '-DISTMethod--+-ROUNDROBIN-----------------------------------------------+-'
                 +-SERVERWLM -| SERVERWLM distribution method options |---+
                 +-WEIGHTEDActive-----------------------------------------+
                 '-HOTSTANDBY -| HOTSTANDBY distribution method options |-'

   .-NOOPTLOCAL----------.
>--+---------------------+--+--------------+--+-------+--------->
   |           .-1-----. |  '-SYSPLEXPorts-'  '-PAUSE-'
   '-OPTLOCAL--+-------+-'
               '-value-'
```

```
IPCONFIG SYSPLEXROUTING DYNAMICXCF 193.9.200.1 255.255.255.0 1
…
VIPADYNAMIC
VIPADEFINE 255.255.255.0 10.91.1.1

VIPADISTRIBUTE DISTMETHOD SERVERWLM OPTLOCAL PAUSE
        10.91.1.1 PORT 9999 8888 DESTIP ALL
ENDVIPADYNAMIC
```

Usage & Invocation – New Vary TCPIP,,SYSPLEX options

- New DISTPAUSE and DISTRESUME options added to the Vary TCPIP,,SYSPLEX command




```
VARY TCPIP,TCPDIST,SYSplex,DISTPAUSE,DVIPA=10.91.1.1,PORT=9999
```

Pauses sysplex distribution for new connections to DIVPA 10.91.1.1 port 9999

```
VARY TCPIP,TCPDIST,SYSplex,DISTPAUSE,DVIPA=10.91.1.1
```

Pauses sysplex distribution for new connections to DIVPA 10.91.1.1 and all ports

```
VARY TCPIP,TCPDIST,SYSplex,DISTRESUME,DVIPA=10.91.1.1,PORT=9999
```

Resumes sysplex distribution for connections to DIVPA 10.91.1.1 port 9999

Usage & Invocation – Diagnostics and monitoring (VIPADCFG)

- Netstat VIPADCFG – Displays Dynamic VIPA configuration
  - Includes VIPADISTRIBUTE statements
    - Use to confirm PAUSE keyword was configured

```
Long format:
VIPA DISTRIBUTE:
  DEST:      10.91.1.1..9999
    TARGET: ALL
    DISTMETHOD: BASEWLM
    SYSPT:   NO   TIMAFF: NO    FLG:  OPTLOCAL, PAUSE

Short format:
 VIPA DISTRIBUTE:
   IP ADDRESS        PORT   TARGET           SYSPT  TIMAFF  FLG
   ----------        ----   ------           -----  ------  ---
   10.91.1.1         9999   ALL              NO     NO      OP
```

Usage & Invocation – Diagnostics and monitoring (VDPT)

- Netstat VDPT – Displays Sysplex Distributor information
  - Issued on sysplex distributing stack
  - Updated to indicate if distribution for specific DVIPA and PORT has been paused
    - Use to determine why new connections might be rejected and reset
    - Will display PAUSED or P if distribution currently paused

```
Long format:
DEST:         10.91.1.1..9999
  TARGET:     10.61.0.1
  TOTALCONN: 0000000000  RDY: 001  WLM: 08  TSR: 100
  DISTMETHOD: BASEWLM
  FLG: PAUSED

Short format:
 DYNAMIC VIPA DESTINATION PORT TABLE FOR TCP/IP STACKS:
 DEST IPADDR     DPORT TARGET          RDY TOTALCONN  WLM TSR FLG
 -----------     ----- ------          --- ---------  --- --- ---
 10.91.1.1       09999 10.61.0.1       001 0000000000 08  100 P
```

Usage & Invocation – Diagnostics and monitoring (SMF/NMI)

- SMF Type 119 Subtype 4 (TCP/IP profile event record) - new flag to indicate if the PAUSE keyword is configured on the VIPADISTRIBUTE statement
- Can be obtained via the TCP/IP Callable NMI with the GetProfile request

Offset	Name	Length	Format	Description
4(X'4')	NMTP_DDVSFlags	2	Binary	X'0001', NMTP_DDVSPAUSE: If set, PAUSE keyword was specified on the VIPADISTRIBUTE statement. Target stacks can be prevented from receiving DVIPA sysplex distributed connections at TCP/IP stack startup by configuring the PAUSE keyword on the VIPADISTRIBUTE statement for a specific DVIPA and optional port(s).

Interactions & Dependencies

- Software Dependencies: None
- Hardware Dependencies: None
- Exploiters: None

Upgrade & Coexistence Considerations

- To exploit this solution, all systems in the Plex must be at the new z/OS level: No
  - The primary stack (Sysplex Distributor stack) and backup stacks must be at the new z/OS level to fully exploit this new function and maintain quiesce persistence
  - Target only systems can be down-level and work with this new function
- List any toleration/coexistence APARs/PTFs: None

Installation & Configuration

- Installation: None
- Configuration: None beyond what was covered earlier

Summary

- Ability to temporarily suspend (PAUSE) sysplex distribution for specific DVIPA and PORT pairs across the sysplex is now supported
- Suspend can be initiated at stack startup by adding new PAUSE keyword to VIPADISTRIBUTE statement
- Suspend and resume can be initiated with new DISTPAUSE and DISTRESUME options on the Vary TCPIP,,SYSPLEX command at any time
- The state of the pause and resume will persist across DVIPA takeovers and takebacks
- The state of the pause and resume will persist across stack and application termination and restart.

Appendix

z/OS Communications Server Publications

- z/OS Communications Server: IP Configuration Guide
- z/OS Communications Server: IP System Administrator's Commands
- z/OS Communications Server: New Function Summary
- z/OS Communications Server: IP Programmer's Guide and Reference
- z/OS Communications Server: IP Diagnosis Guide
- z/OS Communications Server: IP Configuration Reference