# Sysplex Networking Technologies and Considerations

Gus Kassimis - kassimis@us.ibm.com

IBM Enterprise Networking Solutions
Raleigh, NC, USA

Session: 16742
Monday, March 2, 2015: 4:30 PM-5:45 PM

#SHAREorg

SHARE is an independent volunteer-run information technology association that provides education, professional networking and industry influence.

Copyright (c) 2014 by SHARE Inc. Except where otherwise noted, this work is licensed under http://creativecommons.org/licenses/by-nc-sa/3.0/

# Trademarks, notices, and disclaimers

**The following terms are trademarks or registered trademarks of International Business Machines Corporation in the United States or other countries or both:**

- Advanced Peer-to-Peer Networking®
- AIX®
- alphaWorks®
- AnyNet®
- AS/400®
- BladeCenter®
- Candle®
- CICS®
- DataPower®
- DB2 Connect
- DB2®
- DRDA®
- e-business on demand®
- e-business (logo)
- e business(logo)®
- ESCON®
- FICON®
- GDDM®
- GDPS®
- Geographically Dispersed Parallel Sysplex
- HiperSockets
- HPR Channel Connectivity
- HyperSwap
- i5/OS (logo)
- i5/OS®
- IBM eServer
- IBM (logo)®
- IBM®
- IBM zEnterprise™ System
- IMS
- InfiniBand ®
- IP PrintWay
- IPDS
- iSeries
- LANDP®
- Language Environment®
- MQSeries®
- MVS
- NetView®
- OMEGAMON®
- Open Power
- OpenPower
- Operating System/2®
- Operating System/400®
- OS/2®
- OS/390®
- OS/400®
- Parallel Sysplex®
- POWER®
- POWER7®
- PowerVM
- PR/SM
- pSeries®
- RACF®
- Rational Suite®
- Rational®
- Redbooks
- Redbooks (logo)
- Sysplex Timer®
- System i5
- System p5
- System x®
- System z®
- System z9®
- System z10
- Tivoli (logo)®
- Tivoli®
- VTAM®
- WebSphere®
- xSeries®
- z9®
- z10 BC
- z10 EC
- zEnterprise
- zSeries®
- z/Architecture
- z/OS®
- z/VM®
- z/VSE

* All other products may be trademarks or registered trademarks of their respective companies.

**The following terms are trademarks or registered trademarks of International Business Machines Corporation in the United States or other countries or both:**

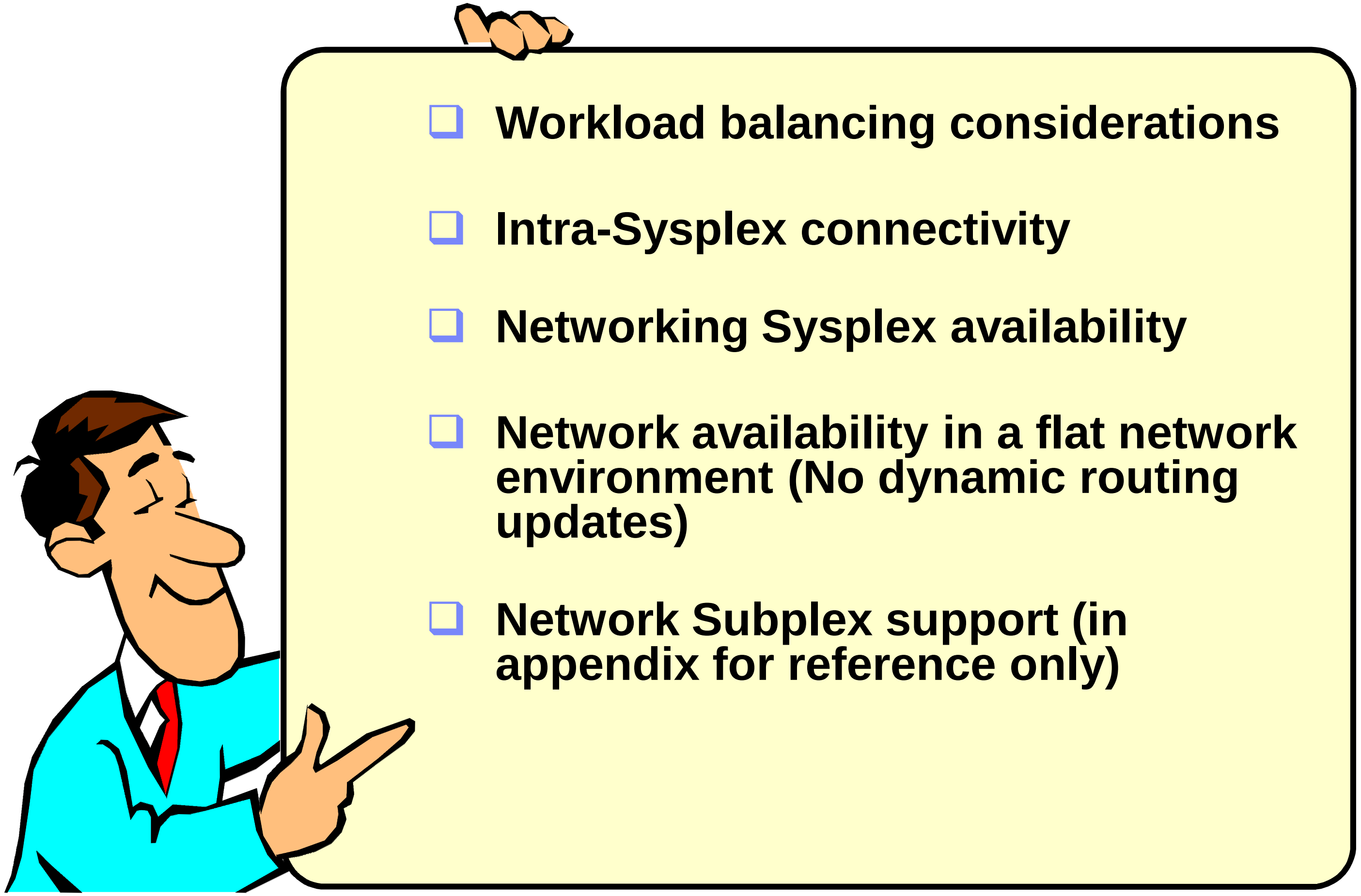
- Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
- Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license there from.
- 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 trademarks of Microsoft Corporation in the United States, other countries, or both.
- InfiniBand is a trademark and service mark of the InfiniBand Trade Association.
- Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
- UNIX is a registered trademark of The Open Group in the United States and other countries.
- Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
- ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
- IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

**Notes**:

- Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput 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 improvements equivalent to the performance ratios stated here.
- IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.
- All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.
- This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.
- All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.
- Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.
- Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

Refer to www.ibm.com/legal/us for further legal information.

# Agenda



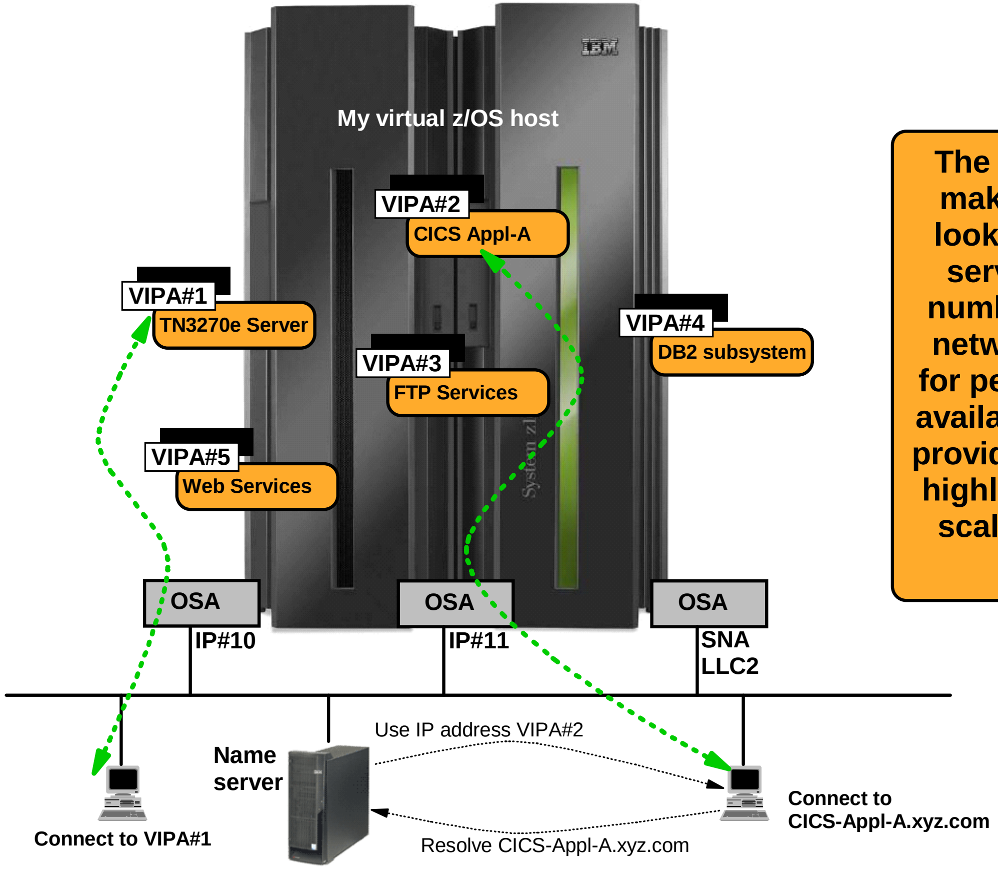
- □ Workload balancing considerations
- □ Intra-Sysplex connectivity
- □ Networking Sysplex availability
- □ Network availability in a flat network environment (No dynamic routing updates)
- □ Network Subplex support (in appendix for reference only)

***Disclaimer: All statements regarding IBM future direction or intent, including current product plans, are subject to change or withdrawal without notice and represent goals and objectives only. All information is provided for informational purposes only, on an “as is” basis, without warranty of any kind.***

© 2015 SHARE and IBM Corporation

# The network view of a Parallel Sysplex - a single large server with many network interfaces and many application services

- The promises of the Parallel Sysplex cluster environment are:
  - Application location independence
  - Ability to shift application workload between LPARs
  - Application single system image from the network
  - Application capacity on-demand
  - Component failure does not lead to application failure
- Gaining the benefits, depend on:
  - Carefully designed redundancy of all key hardware and software components in symmetric configurations
  - Supporting functions in z/OS and middleware
  - Cooperation by applications
  - Operations procedures



**The objective is to make the Sysplex look like one large server that has a number of physical network interfaces for performance and availability - and that provides a number of highly available and scalable services.**

**SNA and TCP/IP**

**✓Single-system image (SSI)**
**✓Scalable**
**✓Highly available**
**✓Secure**

© 2014 SHARE and IBM Corporation

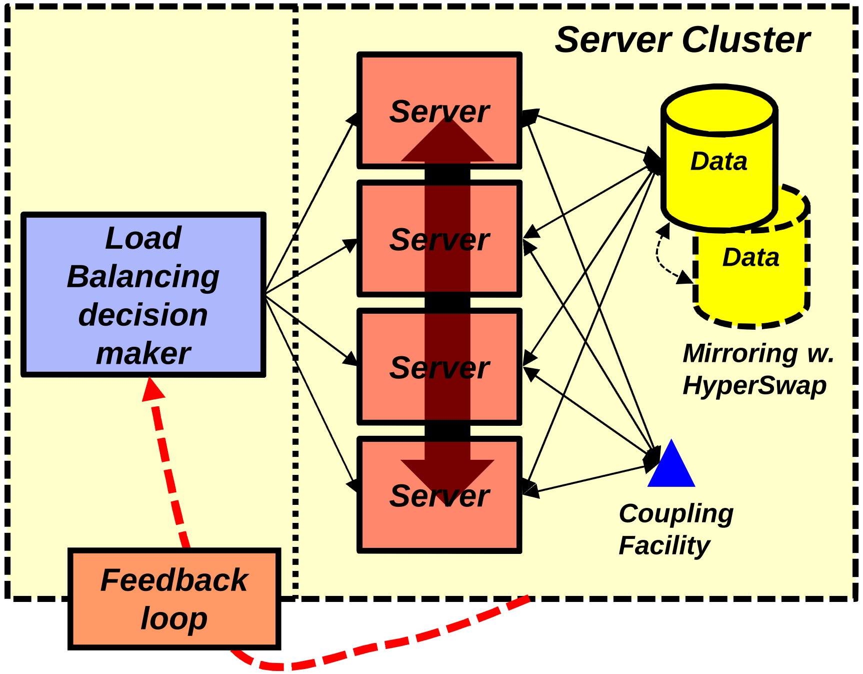
# Workload Balancing Considerations

© 2015 SHARE and IBM Corporation

# What are the main objectives of 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 a z/OS Sysplex that means the applications must be Sysplex-enabled and be able to share data across all LPARs in the Sysplex.*



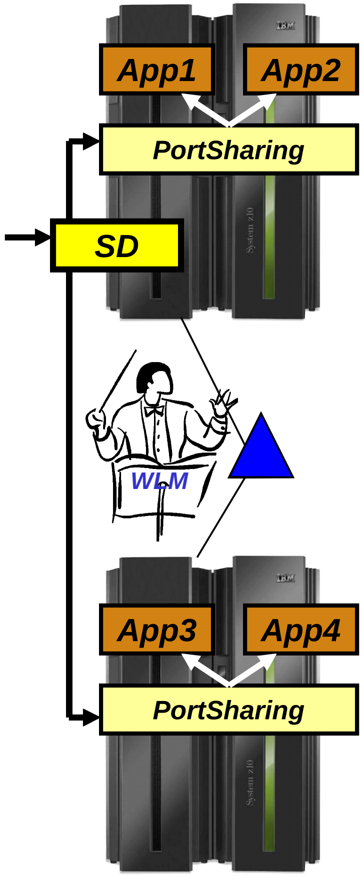
*In order for the load balancing decision maker to meet those objectives, it must be capable of obtaining feedback dynamically, such as server instance availability, capacity, performance, and overall health.*

© 2014 SHARE and IBM Corporation

# z/OS IP network workload balancing overview

- Two main technologies:
  - Sysplex Distributor
  - Port sharing

- Sysplex Distributor
  - Sysplex Distributor is a layer-4 load balancer
    - It makes a decision when it sees an inbound SYN segment for one of the Distributed Dynamic VIPA (DDVIPA) IP address/port combinations it load balances for
  - Sysplex Distributor uses MAC-level forwarding when connection routing takes place over XCF
  - Sysplex Distributor uses GRE when connection routing takes place over any network between the z/OS images
    - Based on definition of VIPAROUTE
  - All inbound packets for a distributed connection must be routed through the Sysplex Distributor LPAR
    - Only the Sysplex Distributor LPAR advertises routing ownership for a DDVIPA, so downstream routers will forward all inbound packets for a given DDVIPA to the distributing LPAR
  - All outbound packets from the server instances can take whatever route is most optimal from the server instance node back to the client



- Port sharing
  - PORTSHARING can be used within a z/OS node to distribute connections among multiple server address spaces within that z/OS node
    - SHAREPORT – TCP/IP Server Efficiency Factor (SEF) value used to perform a weighted round robin distribution to the server instances
    - SHAREPORTWLM – WLM input is used to select server for new connection

© 2014 SHARE and IBM Corporation

# Sysplex Distributor distribution method overview

- z/OS targets without WLM recommendations
  - ROUNDROBIN
    - Static distribution of incoming connections, does not account for target system capacity to absorb new workload
  - WEIGHTEDACTIVE
    - Incoming connections are distributed so the available server instances' percentage of active connections match specified weights

- z/OS targets with WLM recommendations
  - BASEWLM
    - Based on LPAR level CPU capacity/availability and workload importance levels
  - SERVERWLM
    - Similar to BASEWLM but takes into account WLM service class and how well individual application servers are performing (i.e. meeting specified WLM goals) and how much CPU capacity is available for the specific workload being load balanced
    - Enhanced to account for WLM provided server health
    - ***Generally, the recommended distribution method for Sysplex Distributor***

© 2014 SHARE and IBM Corporation

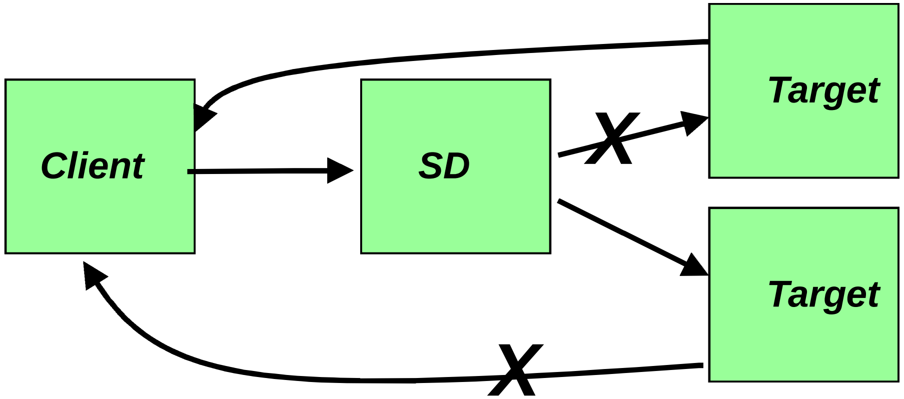
# *Sysplex Distributor distribution method overview …*

- New distribution methods:
  - Non z/OS targets (IBM WebSphere DataPower)
    - TARGETCONTROLLED
      - Incoming connections are distributed among available non-z/OS server instances based on CPU capacity and availability information from target node resident Sysplex Distributor agents.
  - HOTSTANDBY
    - Incoming connections are distributed to a primary server instance and only rerouted to a backup server instance (the “hot standby”) when the primary server instance is not ready, unreachable, or unhealthy.
    - Method added in ***z/OS V1R12***

© 2014 SHARE and IBM Corporation

# Sysplex Distributor built-in awareness of abnormal conditions

- TSR – Target Server Responsiveness
  - How healthy is the target system and application from an SD perspective? A percentage, 0-100%
  - Comprised of several individual health metrics:
    - TCSR – Target Connectivity Success Rate
      - Are connections being sent to the Target System making it there?
      - A Percentage: 100 is good, 0 is bad

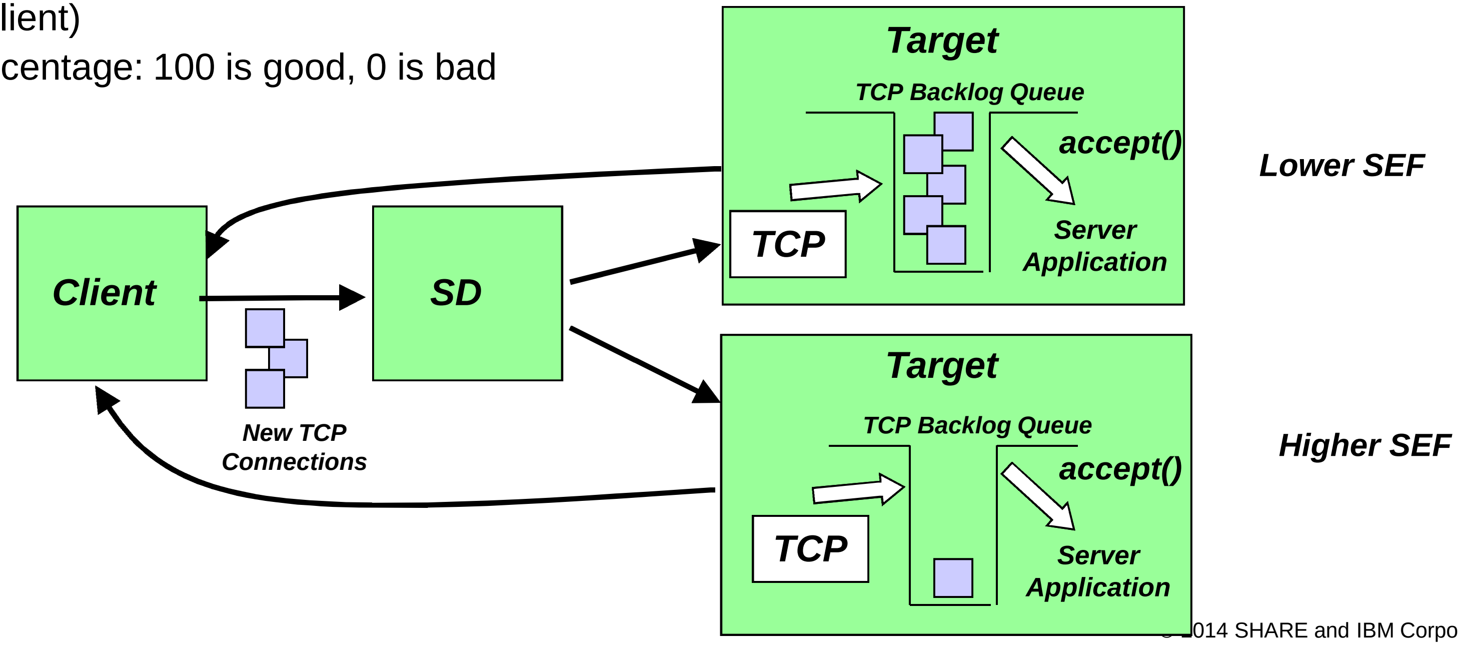


- CER – Connectivity Establishment Rate
  - Is connectivity between the target system and the client ok?
  - By monitoring TCP Connection Establishment state (requires 3 way handshake between client and server) we can detect whether a connectivity issue exists
  - A percentage: 100 is good, 0 is bad
  - Note: CER no longer part of TSR directly but is included in SEF and continues to be calculated and reported separately

© 2014 SHARE and IBM Corporation

# Sysplex Distributor built-in awareness of abnormal conditions

- TSR – Target Server Responsiveness (cont)
  - SEF – Server Efficiency Fraction
    - Is the target server application server keeping up with new connections in its backlog queue?
      - Is the new connection arrival rate higher than the application accept rate? (i.e. is backlog growing over time)
      - How many connections in the TCP backlog queue?  How close to maximum backlog queue depth? Did we have to drop any new connections because the backlog queue max was exceeded?
      - Is the server application hung?  (i.e. not accepting any connections)
      - Are the number of half-open connections on the backlog queue growing? (Similar to CER – One such scenario is when the target system does not have network connectivity to the client)
    - A Percentage: 100 is good, 0 is bad



© 2014 SHARE and IBM Corporation

# *Middleware/Application Issues and the “Storm Drain Problem”*

- TCP/IP and WLM are not aware of all problems experienced by load balancing targets (middleware/applications) – Examples:
  - 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 may appear to be completing the transactions quickly (using little CPU capacity) when they are actually being failed
- This is sometimes referred to as the *Storm Drain* Problem
  - 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"

© 2014 SHARE and IBM Corporation

# Improving WLM awareness of Application Health - Avoiding "Storm Drain" Issues

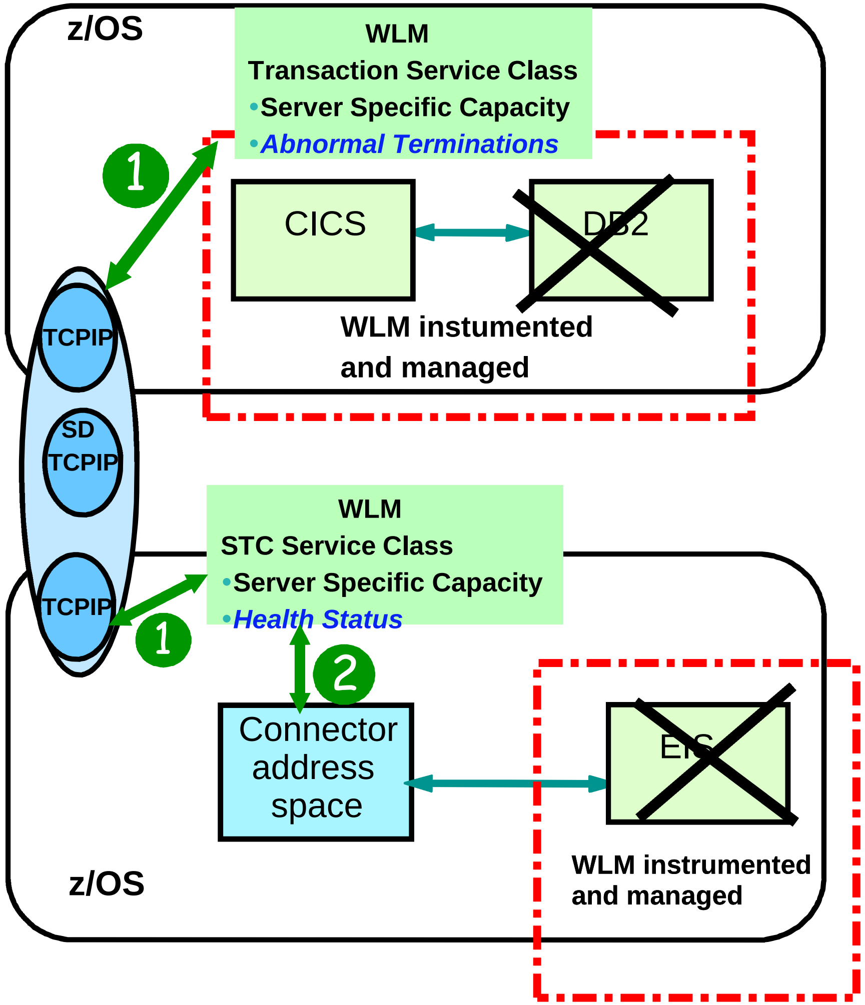
## Server Scenarios

### 1 IWM4SRSC WLM Service

- Used by Sysplex Distributor to obtain WLM recommendations
- Abnormal Termination information: Reported by 1st tier server when transactions can not complete because back end resource managers are not available
  - WLM uses this information to reduce the recommendation for ailing server

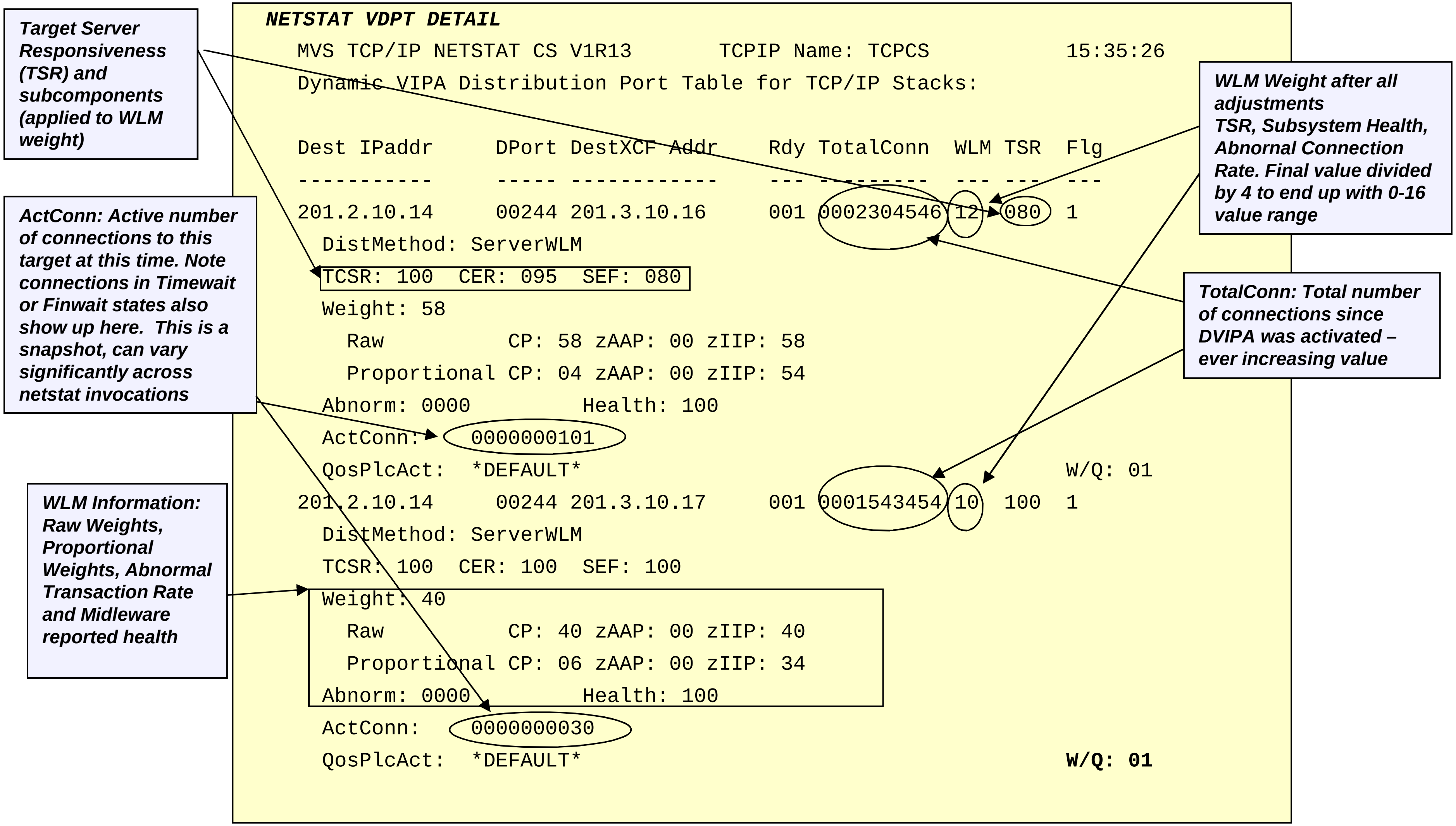
### 2 IWM4HLTH WLM Service

- Allows address spaces which are not instrumented with WLM to set a a health status which is also returned by IWM4SRSC
- The ServerWLM recommendations are reduced when the health is <100%
- Exploited by CICS Transaction Gateway, DB2 and LDAP



© 2014 SHARE and IBM Corporation

# Using Netstat VDPT Detail display to monitor Sysplex Distributor



*Target Server Responsiveness (TSR) and subcomponents (applied to WLM weight)*

*ActConn: Active number of connections to this target at this time. Note connections in Timewait or Finwait states also show up here. This is a snapshot, can vary significantly across netstat invocations*

*WLM Information: Raw Weights, Proportional Weights, Abnormal Transaction Rate and Midleware reported health*

```
NETSTAT VDPT DETAIL
   MVS TCP/IP NETSTAT CS V1R13       TCPIP Name: TCPCS          15:35:26
   Dynamic VIPA Distribution Port Table for TCP/IP Stacks:

   Dest IPaddr     DPort DestXCF Addr    Rdy TotalConn  WLM TSR  Flg
   -----------     ----- ------------    --- ---------- --- ---  ---
   201.2.10.14     00244 201.3.10.16     001 0002304546 12  080  1
     DistMethod: ServerWLM
     TCSR: 100  CER: 095  SEF: 080
     Weight: 58
       Raw          CP: 58 zAAP: 00 zIIP: 58
       Proportional CP: 04 zAAP: 00 zIIP: 54
     Abnorm: 0000          Health: 100
     ActConn:   0000000101
     QosPlcAct:  *DEFAULT*                                  W/Q: 01
   201.2.10.14     00244 201.3.10.17     001 0001543454 10  100  1
     DistMethod: ServerWLM
     TCSR: 100  CER: 100  SEF: 100
     Weight: 40
       Raw          CP: 40 zAAP: 00 zIIP: 40
       Proportional CP: 06 zAAP: 00 zIIP: 34
     Abnorm: 0000          Health: 100
     ActConn:   0000000030
     QosPlcAct:  *DEFAULT*                                  W/Q: 01
```

*WLM Weight after all adjustments TSR, Subsystem Health, Abnornal Connection Rate. Final value divided by 4 to end up with 0-16 value range*

*TotalConn: Total number of connections since DVIPA was activated – ever increasing value*

© 2014 SHARE and IBM Corporation

# *What impacts the final selection of a target server instance?*

| Technology | Target LPAR displaceable capacity as seen by WLM | Server instance performance as seen by WLM | Server instance self-perceived health (as reported to WLM) | Server instance TCP/IP perceived health (the TSR value) | QoS perceived network performance (the QoS fraction) |
|---|---|---|---|---|---|
| SD ROUNDROBIN | No | No | No | Yes<br>(if TSR=zero) | No |
| SD WEIGHTEDACTIVE | No | No | Yes | Yes | No |
| SD BASEWLM | Yes | No | No | Yes | Yes |
| SD SERVERWLM | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* |
| SD TARGETCONTROLLED | Yes<br>(SD agent) | No | No | No | No |
| SD HOTSTANDBY | No | No | Yes | Yes | No |
| PORT SHAREPORT | No | No | No | Yes<br>(Only SEF value) | No |
| PORT SHAREPORTWLM | No | Yes | Yes | Yes<br>(Only SEF value) | No |

© 2014 SHARE and IBM Corporation

# SERVERWLM method: what is displaceable LPAR capacity?

- LPAR capacity that is currently being used for less important workload than what we want to send to the LPAR

- An example:
  - New workload will run at Importance level 2
  - Which LPAR is best?
    - They both have 500 service units of displaceable workload
    - Before z/OS V1R11, they would be considered equally good targets
  - z/OS V1R11 can take importance level of displaceable workload into consideration
    - LPAR2 will be preferred since the importance level of the workload we're displacing is lower than the workload we would displace on LPAR1

***New workload at IL=2 (can displace IL=3 to IL=7 workload)*** →

***IL 0: High***
***IL 7: Low***

**LPAR1**

| IL | SUs |
|---|---|
| 0 | 0 |
| 1 | 0 |
| 2 | 0 |
| 3 | 500 |
| 4 | 0 |
| 5 | 0 |
| 6 | 0 |
| 7 | 0 |

**LPAR2**

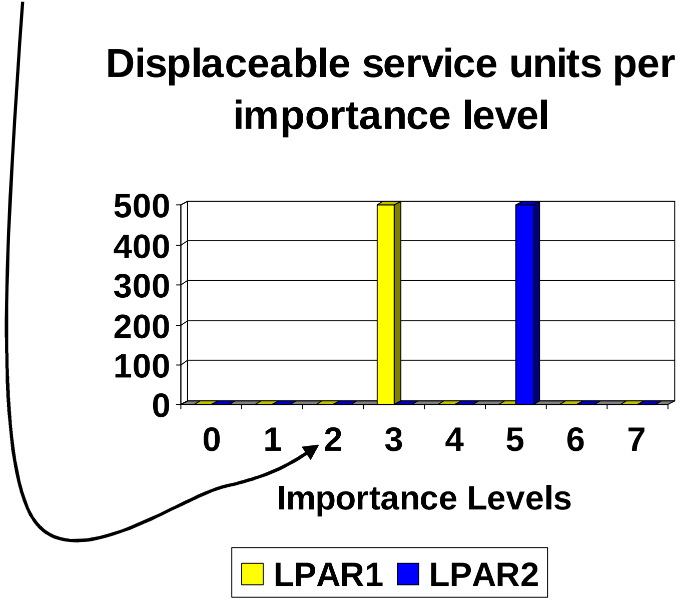
| IL | SUs |
|---|---|
| 0 | 0 |
| 1 | 0 |
| 2 | 0 |
| 3 | 0 |
| 4 | 0 |
| 5 | 500 |
| 6 | 0 |
| 7 | 0 |

© 2014 SHARE and IBM Corporation

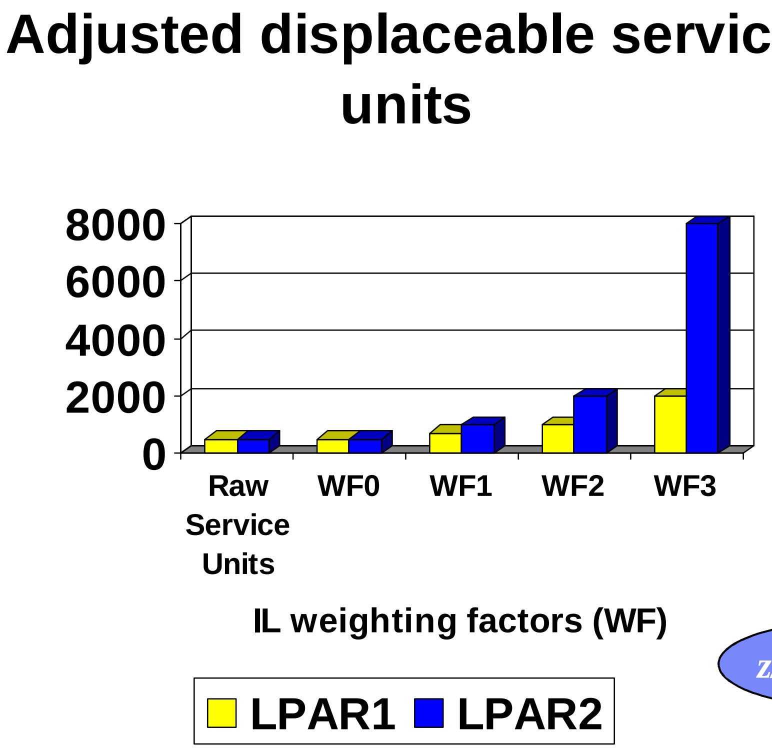
# *SERVERWLM method: How much should importance levels influence the workload distribution?*

- Importance level weighting factor of zero (IL0) means no change as compared to pre-R11 behavior
- Importance level weighting factors of one through 3 (IL1 through IL3), gradually shifts new workloads towards LPARs with the lowest importance level work to displace
  - In this example, LPAR2

**New workload to run at Importance Level 2**



**Displaceable service units per importance level**



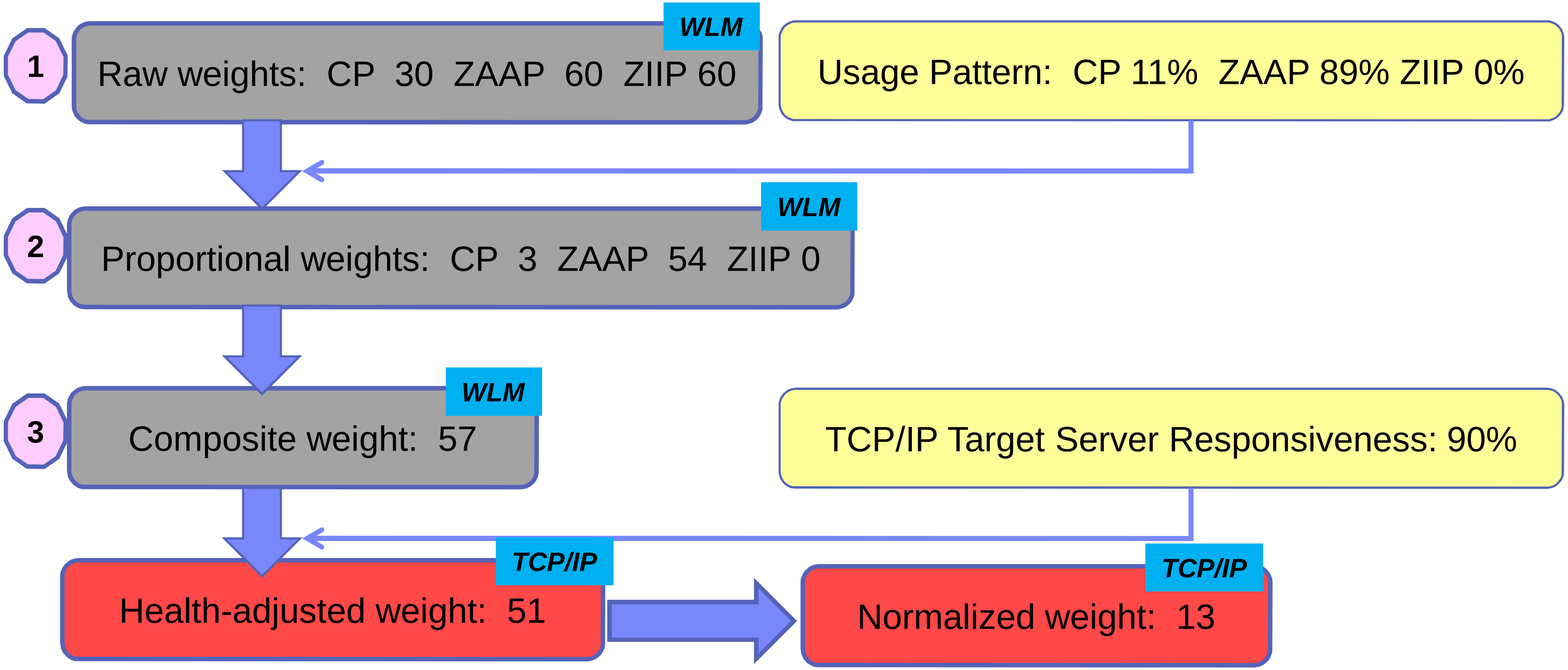
**Adjusted displaceable service units**

*z/OS V1R11*

© 2014 SHARE and IBM Corporation

# SERVERWLM method: specialty processors - overview

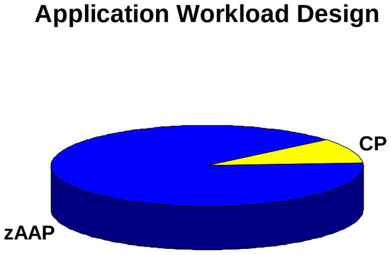
- **When using WLM server-specific weights.**
  - WLM returns
    - The raw CP, zAAP, and zIIP system weights.
    - Proportional weights – raw weight modified by actual server usage pattern as observed by WLM
    - Composite weight

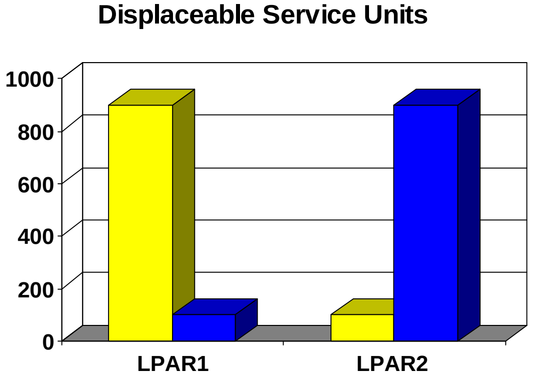

1
WLM
Raw weights: CP 30 ZAAP 60 ZIIP 60
Usage Pattern: CP 11% ZAAP 89% ZIIP 0%
2
WLM
Proportional weights: CP 3 ZAAP 54 ZIIP 0
3
WLM
Composite weight: 57
TCP/IP Target Server Responsiveness: 90%
TCP/IP
Health-adjusted weight: 51
TCP/IP
Normalized weight: 13

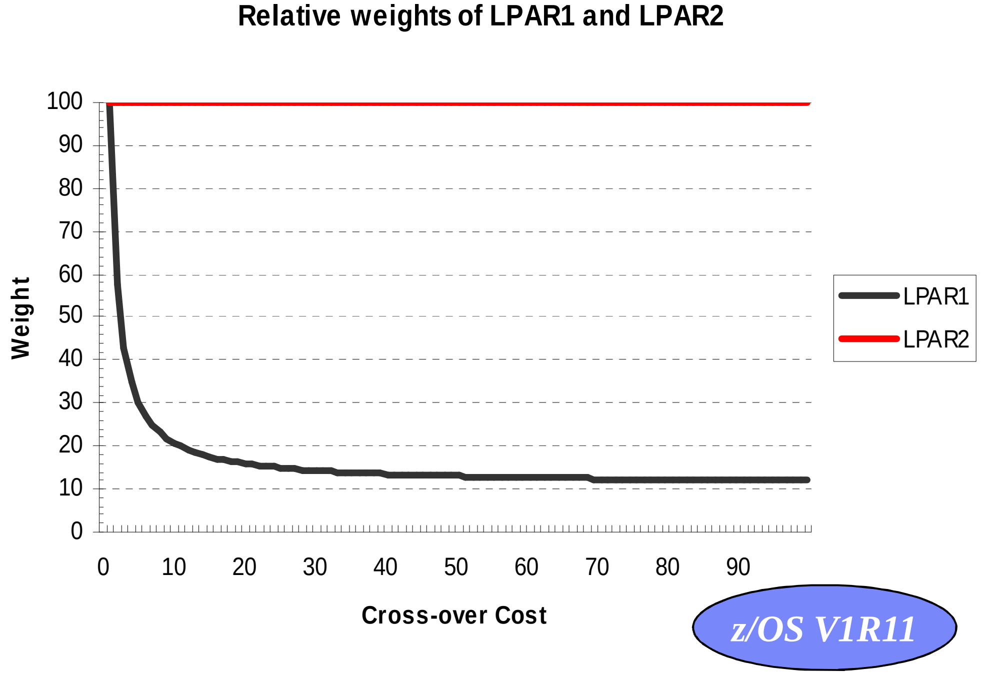
© 2014 SHARE and IBM Corporation

# SERVERWLM method: zIIP and zAAP cross-over to CP

- Application designed to use 10% CP and 90% zAAP
- LPAR1 and LPAR2 are targets
- LPAR1:
  - Has 900 CP SUs and 100 zAAP SUs that can be displaced
- LPAR2:
  - Has 100 CP SUs and 900 zAAP SUs that can be displaced
- Without a cross-over cost, the two targets are equally good to receive new workload
  - The way it always worked prior to z/OS V1R11
- As a cross-over cost is applied, LPAR1 is less attractive than LPAR2
- Cross-over cost can be set to a value between 1 and 100
  - 1: as before z/OS V1R11
  - 100: maximum penalty for cross-over







z/OS V1R11

© 2014 SHARE and IBM Corporation

# SERVERWLM method:  Configuring and displaying the new SERVERWLM options

- The new configuration parameters are
  - Only valid when server-specific recommendations are being used
  - Only used by WLM when all systems in the sysplex are V1R11 or later
- These parameters can affect performance
  - Importance Level values range from 0 (no impact) to 3 (aggressive weighting).
    - Guideline – use Moderate (IL 1) value initially.
  - Crossover cost values range from 1 (no impact) to 100 (crossover cost very expensive).
    - Guideline – Use a low cost initially.

```
VIPADISTRIBUTE
      DISTMETHOD SERVERWLM  PROCXCOST ZIIP 5 ZAAP 20 ILWEIGHTING 1
      201.2.10.11  PORT 8000
      DESTIP  ALL
```

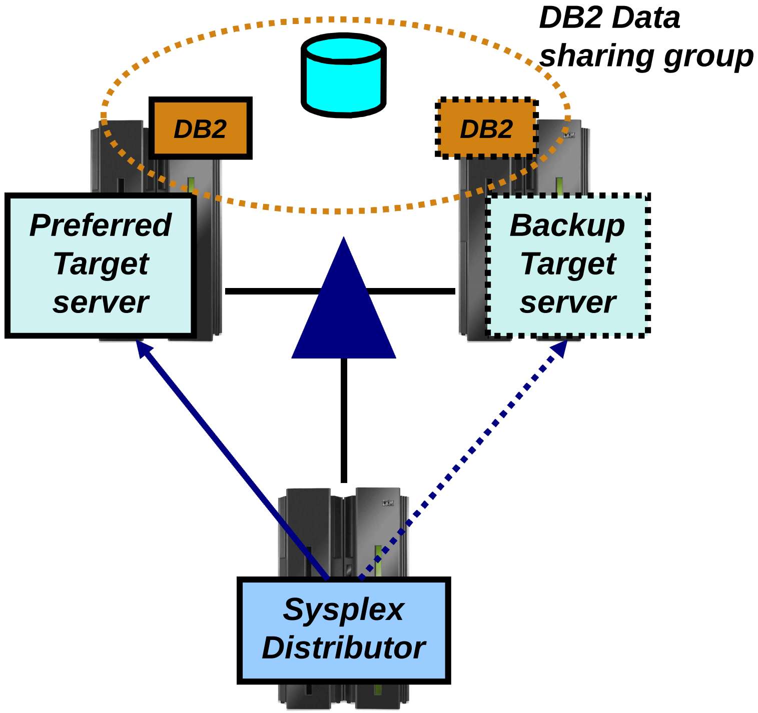
```
NETSTAT VIPADCFG DETAIL
VIPA Distribute:
  Dest:         201.2.10.11..8000
    DestXCF:    ALL
      SysPt:    No  TimAff: No  Flg: ServerWLM
      OptLoc:   No
      ProcXCost:
        zAAP: 020  zIIP: 005
      ILWeighting: 1
```

© 2014 SHARE and IBM Corporation

# Sysplex Distributor hot standby support

z/OS V1R12

- Data Sharing provides for highly scalable and highly available configuration
  - Additional application instances and LPARs can be cloned and added to increase capacity and improve performance
  - But what if the workload can comfortably fit within a single LPAR?
    - Data Sharing becomes primarily an availability feature
    - A Hot Standby configuration would allow all workload to be routed to a single LPAR
      - Minimizing data sharing overhead!
      - While retaining high availability!
- This configuration is currently possible via Policy Agent
  - Using QoS policy:
    - ibm-policyGroupForLoadDistribution:TRUE
- But several users had requested a simpler mechanism for doing this via the TCP/IP profile



© 2014 SHARE and IBM Corporation

# Sysplex Distributor hot standby support…

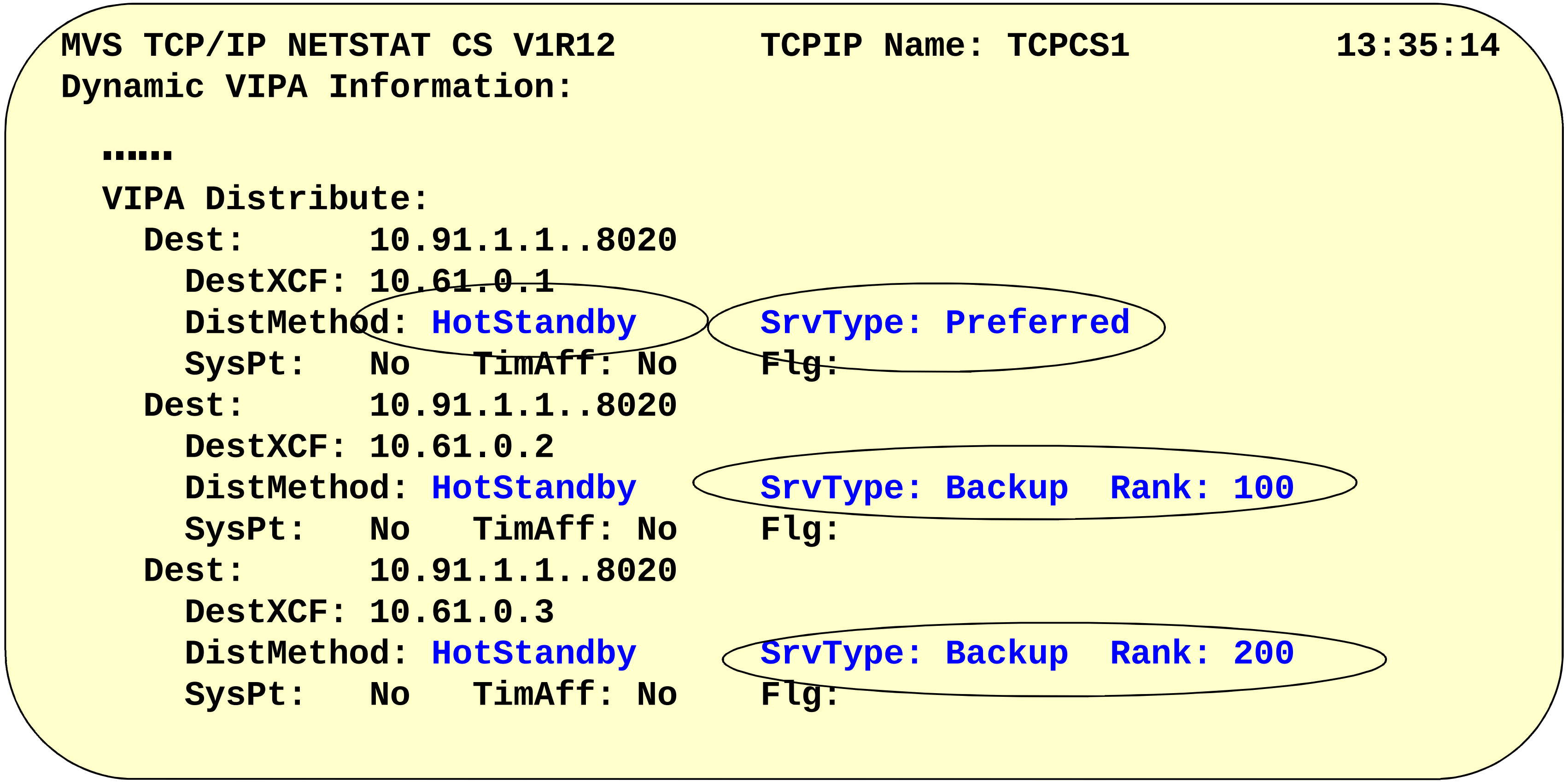
- Have a single target server to receive all new connection requests
  - While other target servers are active but not receiving any new connection requests
  - Automatically route traffic to a backup target server when the active target server is not available
- Enable using a new HOTSTANDBY distribution method
  - One preferred target
    - AUTOSWITCHBACK option - switch to the preferred target if it becomes available
      - No auto switch back if reason for original switch was health problems
        - Use a V TCPIP Quiesce and Resume sequence
  - And one or more backup targets ranked in order of preference
  - A target is not available when:
    - Not ready OR
    - Route to target is inactive  OR
    - If HEALTHSWITCH option configured – target is not healthy when
      - TSR = 0% OR
      - Abnormal terminations = 1000 OR
      - Server reported Health = 0%

```
VIPADEFINE DVIPA1
VIPADISTRIBUTE DISTMETHOD HOTSTANDBY
  AUTOSWITCHBACK HEALTHSWITCH
  DVIPA1 PORT nnnn
  DESTIP  XCF1 PREFERRED
  DESTIP  XCF2 BACKUP 50
  DESTIP  XCF3 BACKUP 100
```

© 2014 SHARE and IBM Corporation

# Sysplex Distributor hot standby support…

## Netstat VIPADCFG/-F VIPADistribute



```
MVS TCP/IP NETSTAT CS V1R12       TCPIP Name: TCPCS1          13:35:14
Dynamic VIPA Information:

  ......
  VIPA Distribute:
    Dest:       10.91.1.1..8020
      DestXCF: 10.61.0.1
      DistMethod: HotStandby      SrvType: Preferred
      SysPt:    No   TimAff: No    Flg:
    Dest:       10.91.1.1..8020
      DestXCF: 10.61.0.2
      DistMethod: HotStandby      SrvType: Backup  Rank: 100
      SysPt:    No   TimAff: No    Flg:
    Dest:       10.91.1.1..8020
      DestXCF: 10.61.0.3
      DistMethod: HotStandby      SrvType: Backup  Rank: 200
      SysPt:    No   TimAff: No    Flg:
```

© 2014 SHARE and IBM Corporation

# Sysplex Distributor hot standby support…

- Determining current active Target: Netstat VDPT/-O
- Server Type
  - Preferred or Backup (based on configuration)
- Flags changed to include server state
  - Active – this server is receiving new connections, Backup – this server is in standby mode

```
MVS TCP/IP NETSTAT CS V1R12       TCPIP Name: TCPCS1          14:18:18
Dynamic VIPA Destination Port Table for TCP/IP stacks:
Dest:          10.91.1.1..8020
  DestXCF:     10.61.0.1
  TotalConn:  0000000000  Rdy: 000  WLM: 10  TSR: 100
  DistMethod: HotStandby            SrvType: Preferred
  Flg: Backup
Dest:          10.91.1.1..8020
  DestXCF:     10.61.0.2
  TotalConn:  0000000000  Rdy: 001  WLM: 10  TSR: 100
  DistMethod: HotStandby            SrvType: Backup
  Flg: Backup
Dest:          10.91.1.1..8020
  DestXCF:     10.61.0.3
  TotalConn:  0000000000  Rdy: 001  WLM: 10  TSR: 100
  DistMethod: HotStandby            SrvType: Backup
  Flg: Active
```

© 2014 SHARE and IBM Corporation

# *Sysplex Distributor Polling Intervals*

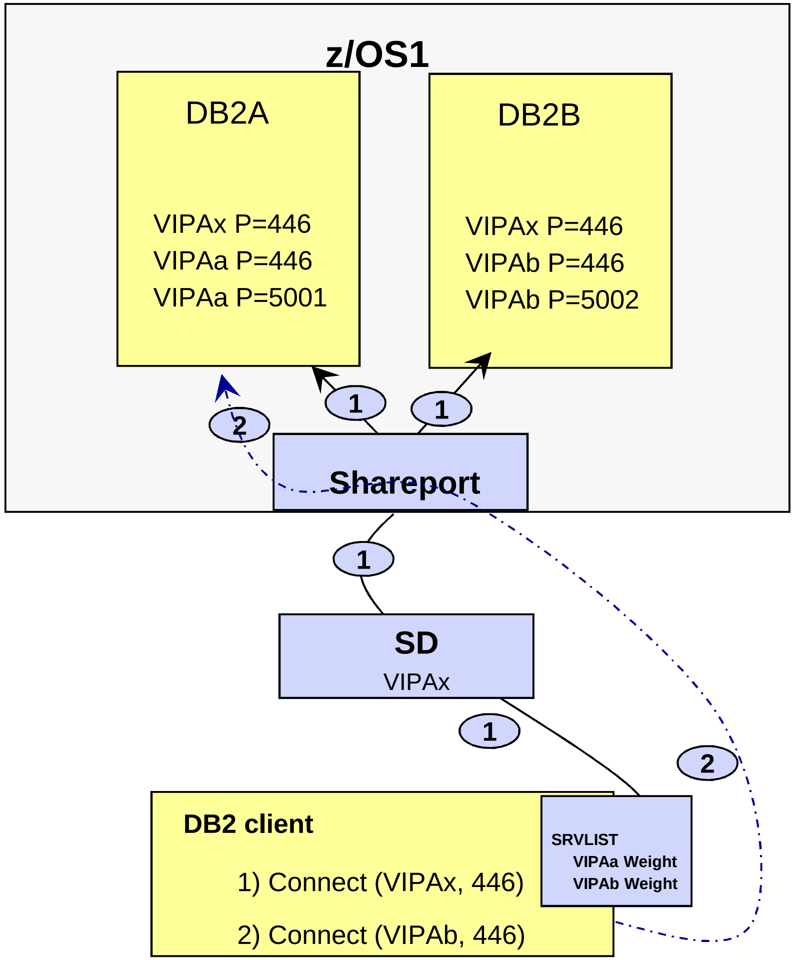
- By default Sysplex Distributor queries WLM every 60 seconds
  - During this same interval Sysplex Distributor also calculates and health metrics such as SEF, TSR, etc.
  - In environments where changes in workload conditions can occur very quickly a smaller polling interval is often more desirable
    - Allows Sysplex Distributor to have more current WLM recommendations and health metrics
    - The polling interval can be set via the ***SYSPLEXWLMPOLL*** keyword on ***GLOBALCONFIG*** statement
      - Specified in seconds (1-180)
      - A value of 10 seconds is recommended for obtaining the most recent WLM recommendations
      - ***Must be configured on all*** TCP/IP stacks in the Sysplex environment (Distributor and Targets)
        - Each stack implements its own timer

© 2014 SHARE and IBM Corporation

# Dynamic VIPA affinity support – Problem Scenario

- **Multiple DB2 members deployed in the same z/OS image**
- **Currently 2 methods for creating Dynamic VIPAs to represent each member (Member-specific DVIPAs)**

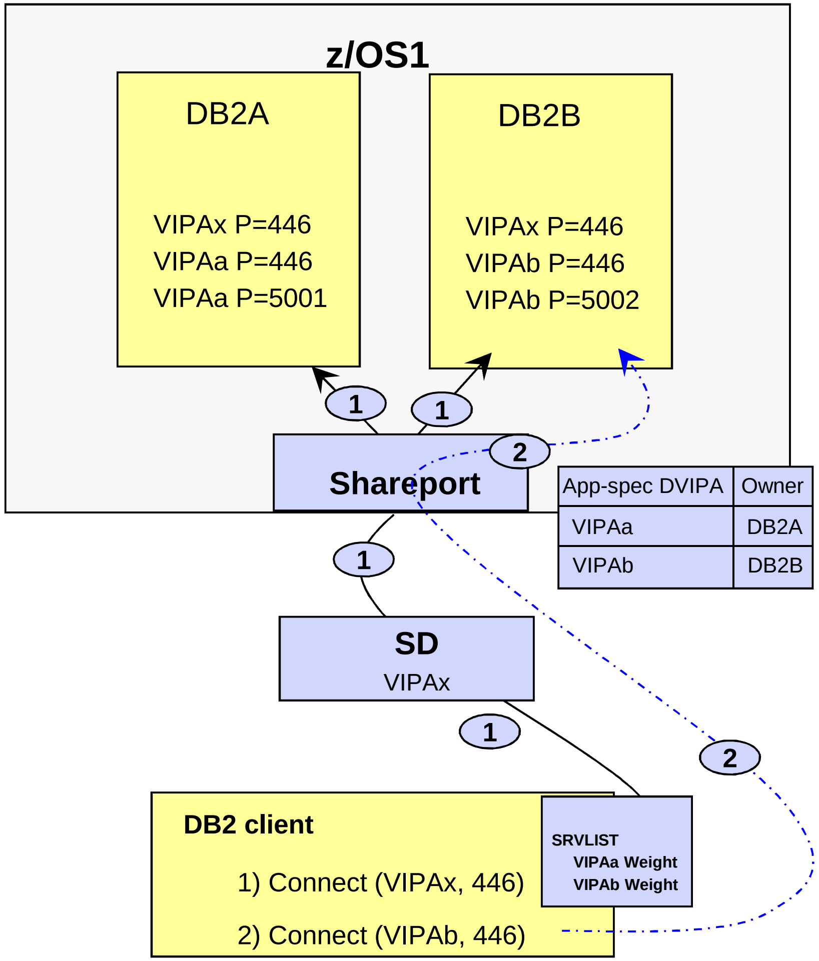
1. BIND-specific approach: Using the TCP/IP profile PORT BIND specific support each member binds its listening socket to a specific member DVIPA
   - ***All is well,*** incoming connections get routed to the appropriate member
2. INADDR_ANY approach: Dynamic VIPAs configured via DB2 BSDS (Bootstrap dataset)
   - DB2 programmatically creates the DVIPAs during initialization and ***binds listening sockets to*** **INADDR_ANY**
     - o Allows each DB2 member to be reached using any IP address (including the IPv4 and IPv6 DVIPAs, DB2 Location Alias IP addresses)
     - o However if multiple DB2 members in the same z/OS system use the BSDS method, ***incorrect routing of incoming connections is possible***
     - o Resulting in suboptimal load balancing



© 2015 SHARE and IBM Corporation

# Affinity for application-instance DVIPAs - Solution

- **Provide a capability to create an application instance DVIPA with affinity**
  - DVIPA affinity determined by creating address space (i.e. DDF instance)
  - Incoming connections to an “affinity” DVIPA are handled in a special manner by SHAREPORT processing
    - When multiple listening sockets for the target port are available find the listening socket owned by the address space that created the DVIPA
    - If an address space with affinity is not found with a listening socket on the target port then route the connection to any listening socket that can accept it (i.e. bound to INADDR_ANY
      - Allows the DVIPA to be used by non-DB2 applications, such as incoming FTP connections to the member-specific DVIPA address
      - Only works while the DVIPA is still active
- **Support to create DVIPA with affinity is provided on**
  - Socket APIs (SIOCSVIPA and SIOCSVIPA6 IOCTLs)
  - MODDVIPA utility program
- **Allows a sysplexWLB connection using a BSDS DVIPA (INADDR_ANY approach) to land on the intended DB2 member even when multiple DB2 members coexist on the same z/OS image.**
  - Will require new DB2 exploitation support
    - APAR PI08208 (available on DB2 V10 and DB2 V11)



| App-spec DVIPA | Owner |
|---|---|
| VIPAa | DB2A |
| VIPAb | DB2B |

© 2015 SHARE and IBM Corporation

# Intra-Sysplex connectivity

© 2015 SHARE and IBM Corporation

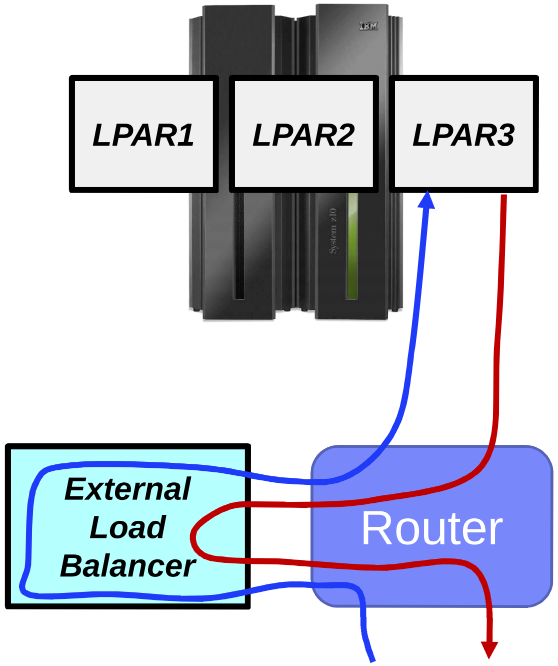
# Load Balancing - Inbound and outbound routing paths

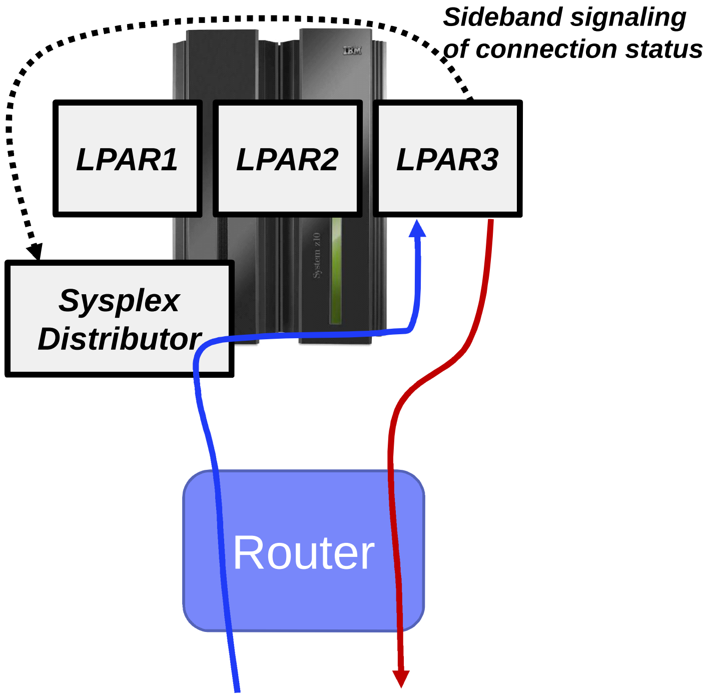
***Most external load balancers***

***The external load balancer generally uses server IP address NATing for the inbound flows, but it can also use Generic Routing Encapsulation (GRE). For outbound it either uses client IP address (and port) NATing or it relies on the associated router to enforce policy-based routing that directs the outbound packets back via the load balancer.***

***Sysplex Distributor***

***Sysplex Distributor does not use NAT; it uses MAC-level forwarding for the inbound flows, which requires the target servers are on a directly connected network (XCF network), or the use of GRE (VIPAROUTE). Sysplex Distributor does not need to be in the outbound path, so no control of outbound flows are needed.***

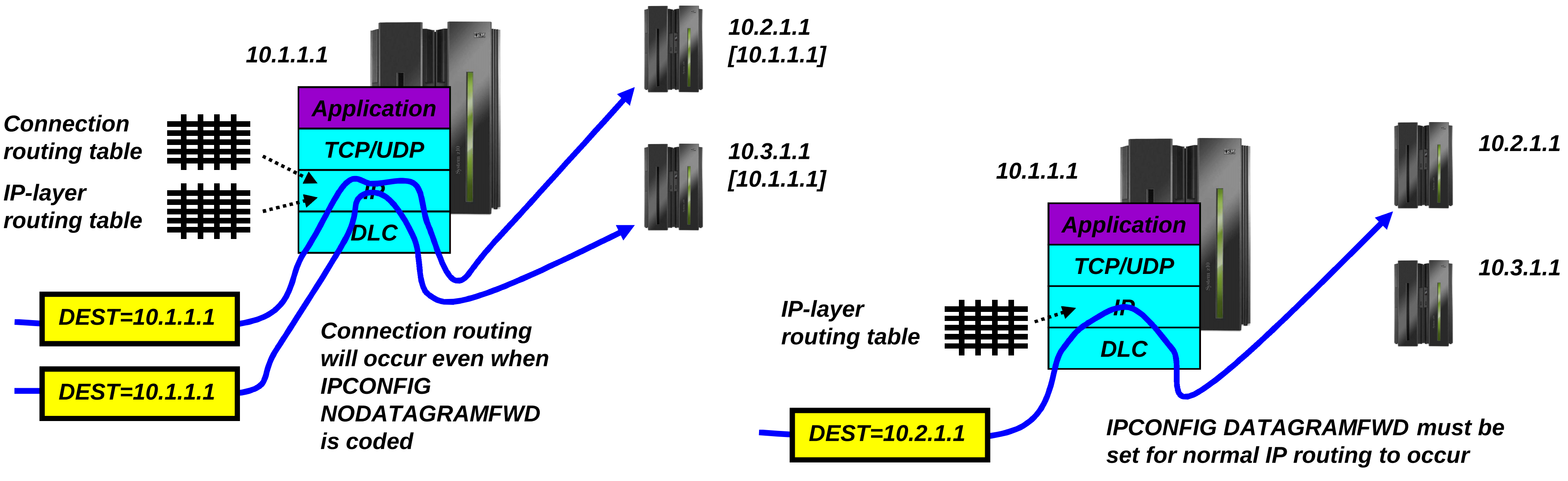




© 2015 SHARE and IBM Corporation

# Two types of intra-Sysplex/subplex routing

- **Connection routing**
  - IP routing decision based upon connection routing table (CRT), destination IP address and specific connection (4-tuple)
    - Packets to the same IP address, but belonging to two different connections, may go to two different targets
  - Used by Sysplex Distributor
  - Used by movable Dynamic VIPA support
  - Not subject to the setting of IPCONFIG DATAGRAMFWD/NODATAGRAMFWD
- **Normal IP routing**
  - IP routing decision based upon IP-layer routing table and destination IP address
    - All packets to the same IP address are treated the same
  - Forwarding to z/OS TCP/IP stacks through another z/OS TCP/IP stack
  - Subject to the setting of the IPCONFIG DATAGRAMFWD/NODATAGRAMFWD option



*Connection routing will occur even when IPCONFIG NODATAGRAMFWD is coded*

*IPCONFIG DATAGRAMFWD must be set for normal IP routing to occur*

© 2015 SHARE and IBM Corporation

# The role of XCF, ISTIQDIO HiperSockets, and external LAN interfaces in a z/OS Sysplex/subplex

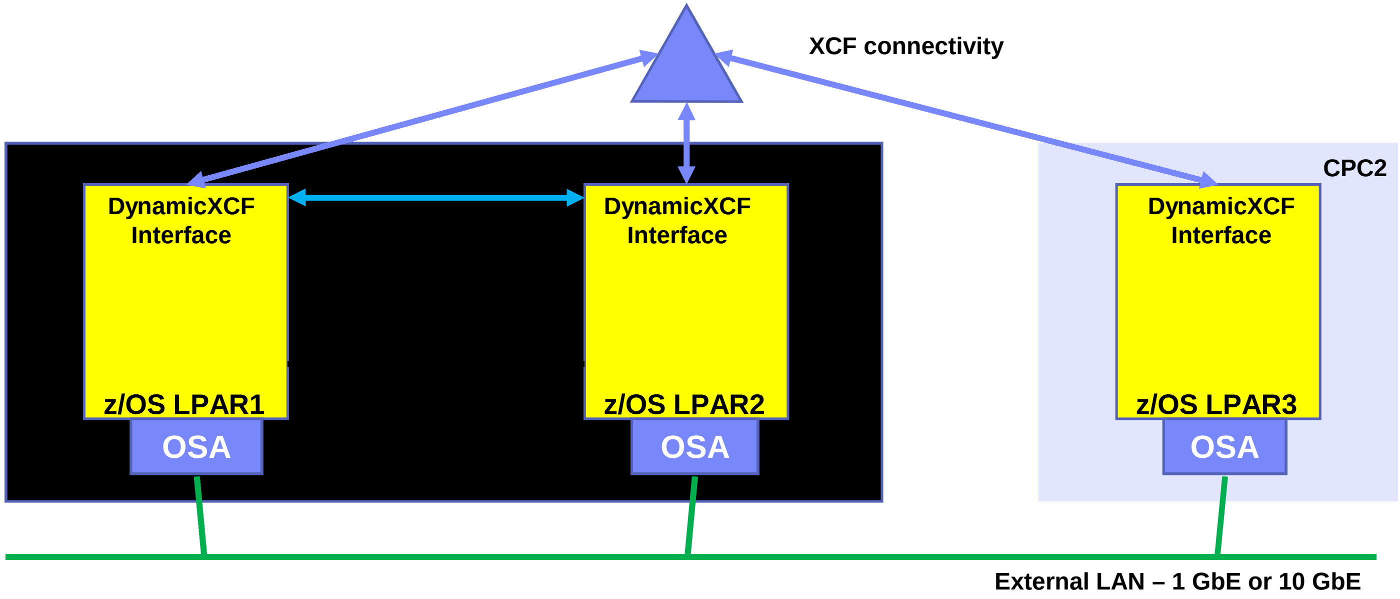
- **XCF**
  - All XCF control messaging between stacks in a Sysplex (DVIPA availability, etc.) – always go via XCF messages
  - DynamicXCF SD connection routing (but only if no VIPAROUTE defined)
  - If so configured through static or dynamic routing, normal IP routing between LPARs – normally not recommended

- **ISTIQDIO HiperSockets**
  - If ISTIQDIO is defined (in VTAM) DynamicXCF SD connection routing between LPARs on same CPC goes this way instead of XCF
  - Considered part of the DynamicXCF network interface – no separate DEVICE/LINK or INTERFACE definitions

- **External LAN or a manually defined HiperSockets LAN**
  - If VIPAROUTE defined, then  used for SD connection routing between LPARs
    - VIPAROUTE is generally recommended
  - Normal IP routing

**Only define DynamicXCF interfaces as OSPF interfaces, if you want to be able to use XCF as a last-resort connectivity between z/OS stacks.**

**If you have “enough” redundancy built into your OSA adapters, data center switches, and switch connectivity, you may not need to ever use XCF for normal IP routing.**



© 2015 SHARE and IBM Corporation

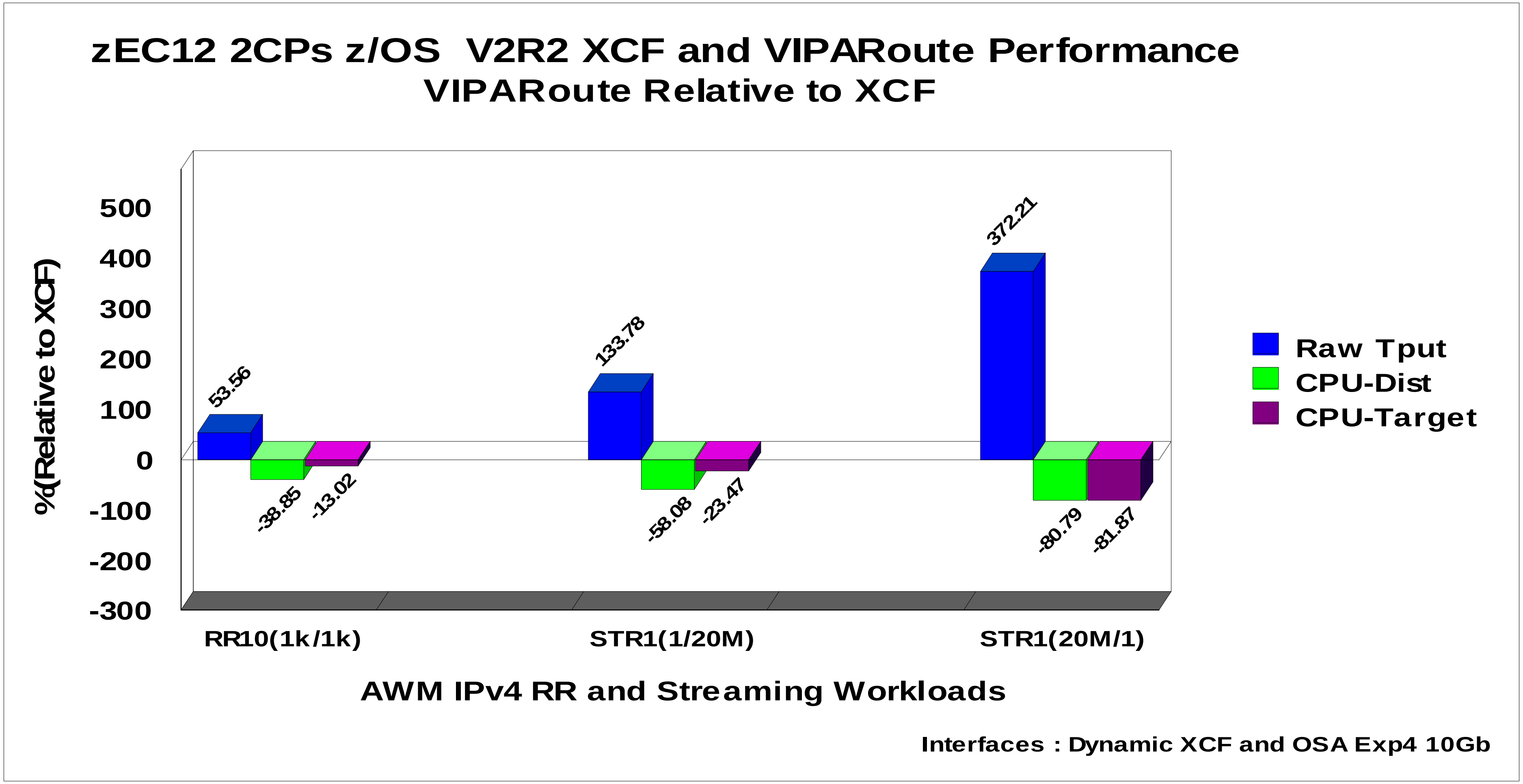
# So what should I use for what type of routing?

- **VIPAROUTE** is often the best choice for connection routing
  - Exploits network redundancy
  - Often as fast or faster than XCF
  - Does not use Coupling Facility CPU cycles, which often is a limited resource

| | **Exchange control messages between stacks in a Sysplex or Subplex** | **Sysplex Distributor connection routing (forwarding inbound packets for distributed connections)** | **General IP routing between stacks in a Sysplex or Subplex** |
|---|---|---|---|
| **XCF messaging** | Always | Yes - If no VIPAROUTE specified (or for traffic associated with SWSA and MLS) | Can be used (not recommended) |
| **IUTIQDIO (Dedicated HiperSockets LAN)** | Never | Yes - If defined in VTAM start options and no VIPAROUTE defined. Used for connection routing to LPARs on same CPC only. | Can be used (not recommended since XCF will be used for LPARs on other CPCs) |
| **All other connectivity options between stacks in a Sysplex or Subplex (OSA, HiperSockets, Channel links, etc.** | Never | Yes - If VIPAROUTE is defined | Always |

© 2015 SHARE and IBM Corporation

# VIPAROUTE vs XCF – Performance Comparison



***Significant throughput benefits for both request/response (RR) and streaming data (STR) patterns***

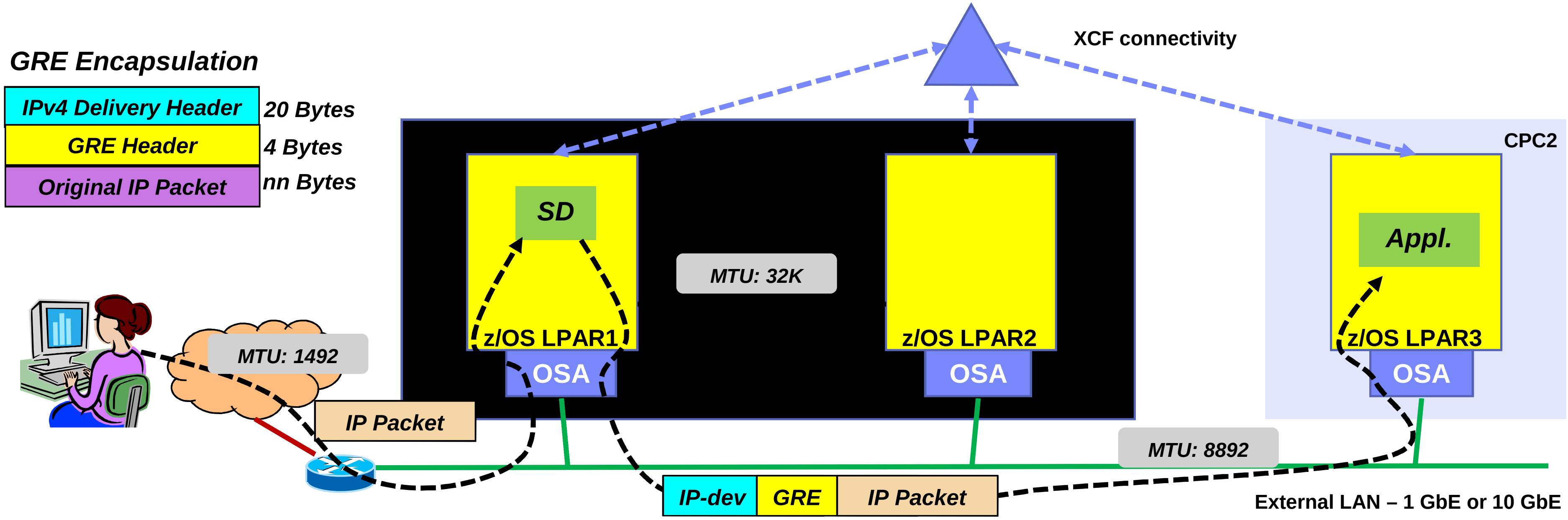
***Significant reduction in networking CPU overhead on both Sysplex Distributor and Target LPARs***

***Note: The performance measurements were collected in IBM internal tests using a dedicated system environment. The results obtained in other configurations or operating system environments may vary significantly depending upon environments used. Therefore, no assurance can be given, and there is no guarantee that an individual user will achieve performance or throughput improvements equivalent to the results stated here.***

© 2015 SHARE and IBM Corporation

# VIPAROUTE and MTU size considerations

- When VIPAROUTE is used, the distributing stack adds a GRE header to the original IP packet before forwarding to the target stack
- Two ways to avoid fragmentation between distributing and target stacks:
  - Have clients use path MTU discovery
    - z/OS will factor in the GRE header size (24 bytes) when responding with next-hop MTU size
    - Not always possible to control distributed nodes' settings from the data center
  - Use jumbo-frames on the data center network
    - The access network will typically be limited to Ethernet MTU size (1492 bytes), while the data center network will be able to use jumbo frame MTU size (8892 bytes)
    - Adding the GRE header will not cause fragmentation in this scenario



© 2015 SHARE and IBM Corporation

# VIPAROUTE fragmentation avoidance

- VIPAROUTE has been used extensively by many users to offload sysplex distributor forwarded traffic from XCF links
  - When used in combination with QDIO Accelerator for SD can result in dramatically reduced overhead for SD forwarding
- Fragmentation is still a concern for several customers
  - Resulting from the extra 24 bytes that are needed for the GRE header
  - Path MTU Discovery helps but doesn’t solve the issue in some environments (where ICMP messages cannot flow across FWs)
  - Fragmentation can cause significant performance degradation

© 2015 SHARE and IBM Corporation

***New in V2R2***

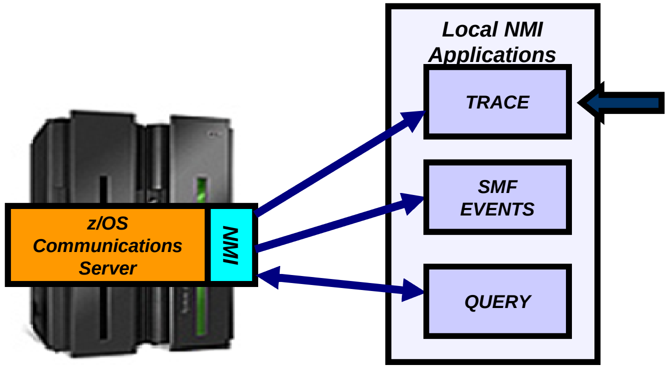
# VIPAROUTE fragmentation avoidance (cont)

- V2R2 introduces a new autonomic option that will automatically reduce the MSS (Maximum Segment Size) of a distributed connection by the length of the GRE header.
  - This will allow the client TCP stack to build packets that allow for the 24 bytes of the GRE header to be added without any fragmentation being required.
  - ADJUSTDVIPAMSS on GLOBALCONFIG
    - Defaults to *AUTO* - Enables adjusted MSS
      - On target TCP/IP stacks when VIPAROUTE is being used
      - On Sysplex Distributor stack if it is also a target and VIPAROUTE is defined
    - Option *ALL* – Enables adjusted MSS for all connections using a DVIPA (distributed or not)
    - Option *NONE* - If you are already exploiting VIPAROUTE and know that there's no fragmentation possible in your environment you can disable this function
    - *Note:* This is a stack specific option.  If the default is not taken then it must be configured on all systems (and TCP/IP stacks) in the sysplex

© 2015 SHARE and IBM Corporation

# z/OS V1R12 - Enhance Packet Trace for Sysplex Distributor VIPAROUTE traffic

- Apply Packet Trace filters to Sysplex Distributor VIPAROUTE traffic
  - Sysplex Distributor encapsulates VIPAROUTE traffic with GRE header, for IPv4 traffic, or an IPv6 header, for IPv6 traffic
    - Existing filter support only operates on the outer packet header, not the encapsulated packet
  - Packet Trace can now filter on the destination DVIPA address and/or the ports located inside the encapsulated packet

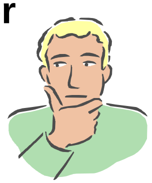


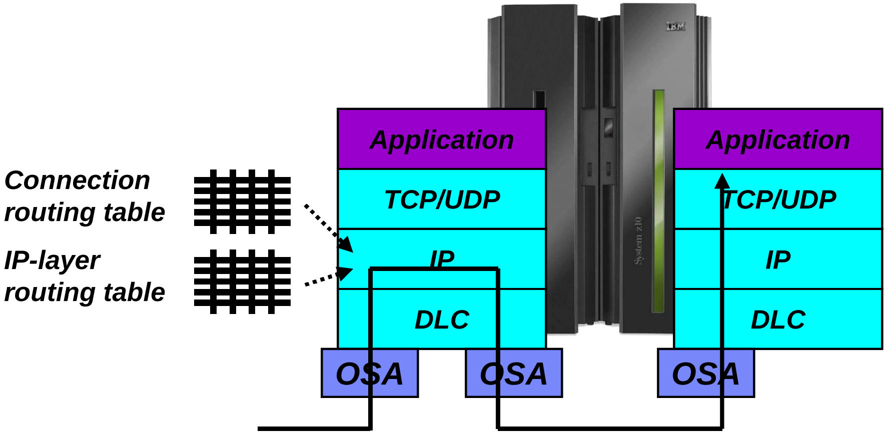
- In addition, the next hop address is now included in the packet trace

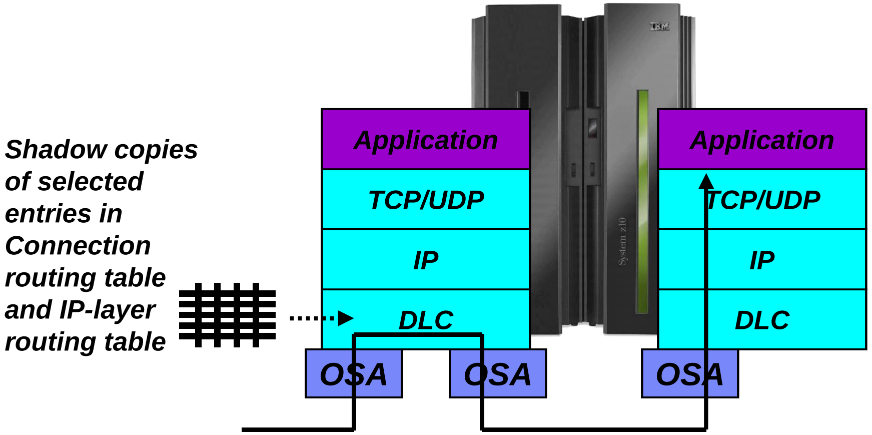
© 2015 SHARE and IBM Corporation

# z/OS V1R11 QDIO and iQDIO routing accelerator

- **Provides fast path IP forwarding for these DLC combinations:**
  - Inbound OSA-E QDIO → Outbound OSA-E QDIO or HiperSockets
  - Inbound HiperSockets → Outbound OSA-E QDIO or HiperSockets
- **Adds Sysplex Distributor (SD) acceleration**
  - Inbound packets over HiperSockets or OSA-E QDIO
  - When SD gets to the target stack using either:
    - Dynamic XCF connectivity over HiperSockets
    - VIPAROUTE over OSA-E QDIO
- **Improves performance and reduces processor usage for such workloads**
- **Restrictions:**
  - QDIO routing accelerator is IPv4 only
  - Mutually exclusive with IPSECURITY
  - Requires IP Forwarding to be enabled (for non-SD acceleration)
  - No acceleration for:
    - Traffic which requires fragmentation in order to be forwarded
    - VIPAROUTE over HiperSockets
    - Incoming fragments for an SD connection
    - Interfaces using optimized latency mode (OLM)






Shadow copies of selected entries in Connection routing table and IP-layer routing table
Application
TCP/UDP
IP
DLC
Application
TCP/UDP
IP
DLC
OSA
OSA
OSA

© 2015 SHARE and IBM Corporation

V2R1

# QDIO acceleration coexistence with IP filtering …

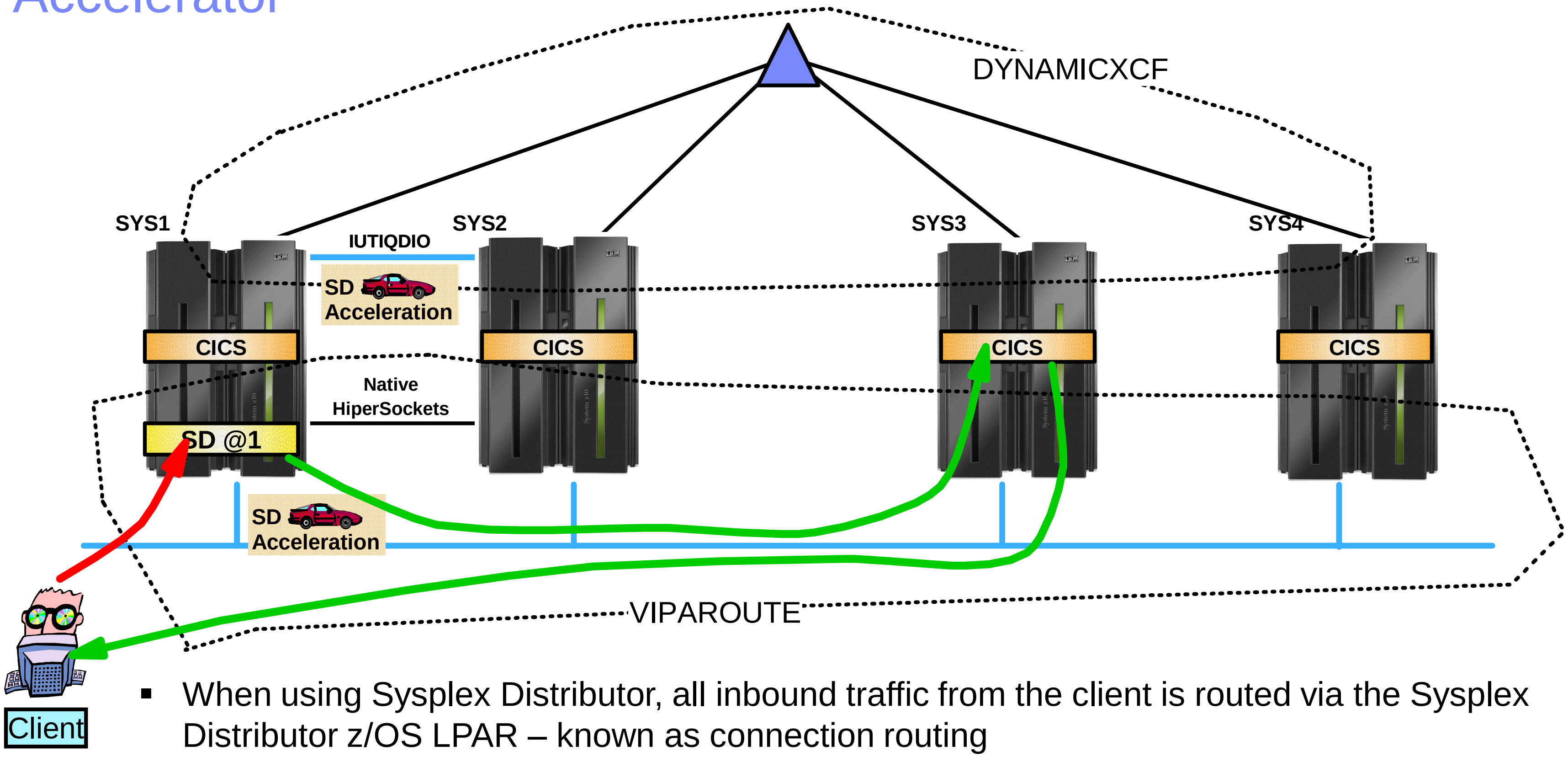
- Tthere are valid cases where it makes sense to specify QDIOACCELERATOR with IPSECURITY - cases where IP filtering is not needed for the QDIO Accelerator traffic:
  - The routed traffic is destined for a target that’s doing its own endpoint filtering
  - IPSECURITY is only specified to enable IPSec on the local node

- V2R1 will allow QDIOACCELERATOR to be specified with IPSECURITY in the TCPIP profile under certain conditions:

| **IP filter rules & defensive filter rules permit all routed traffic?** | **IP filter rules & defensive filter rules require routed traffic to be logged?** | **QDIO acceleration permitted?** |
|---|---|---|
| N | N | **N** |
| N | Y | **N** |
| Y | Y | **N** |
| Y | N | **Y** |
| **Sysplex Distributor traffic always forwarded** | | |

© 2015 SHARE and IBM Corporation

# Sysplex Distributor connection routing benefits from QDIO/iQDIO Accelerator

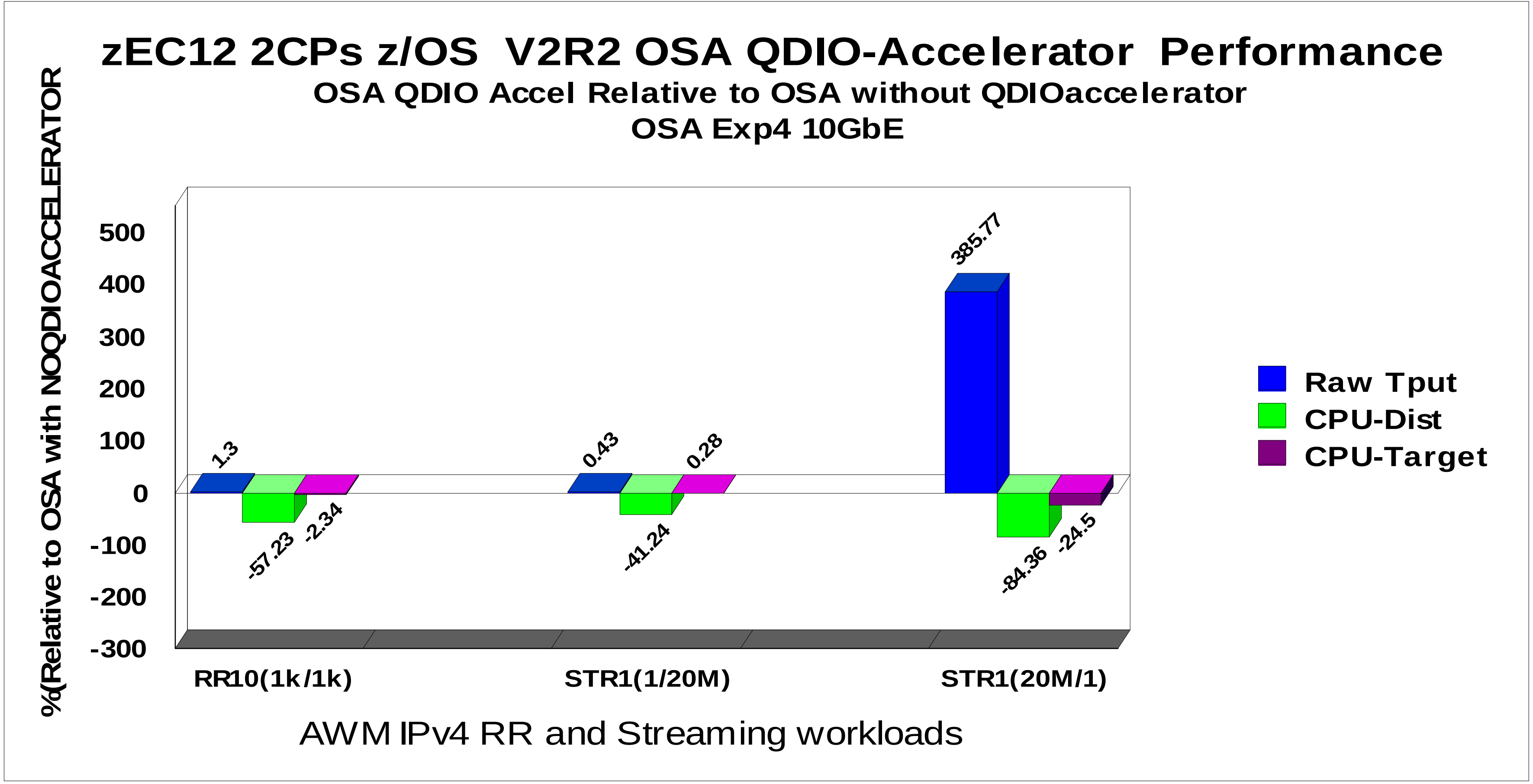


- When using Sysplex Distributor, all inbound traffic from the client is routed via the Sysplex Distributor z/OS LPAR – known as connection routing
  - – Outbound traffic goes directly back to the client
- When inbound packets to Sysplex Distributor is over QDIO or iQDIO (HiperSockets), Sysplex Distributor will perform accelerated connection routing when outbound is a DYNAMICXCF iQDIO interface - or when the outbound interface is a QDIO network interface
  - – Helping reduce CPU overhead and latency in the Sysplex Distributor LPAR (SYS1)

© 2015 SHARE and IBM Corporation

# Sysplex Distributor Accelerator Performance

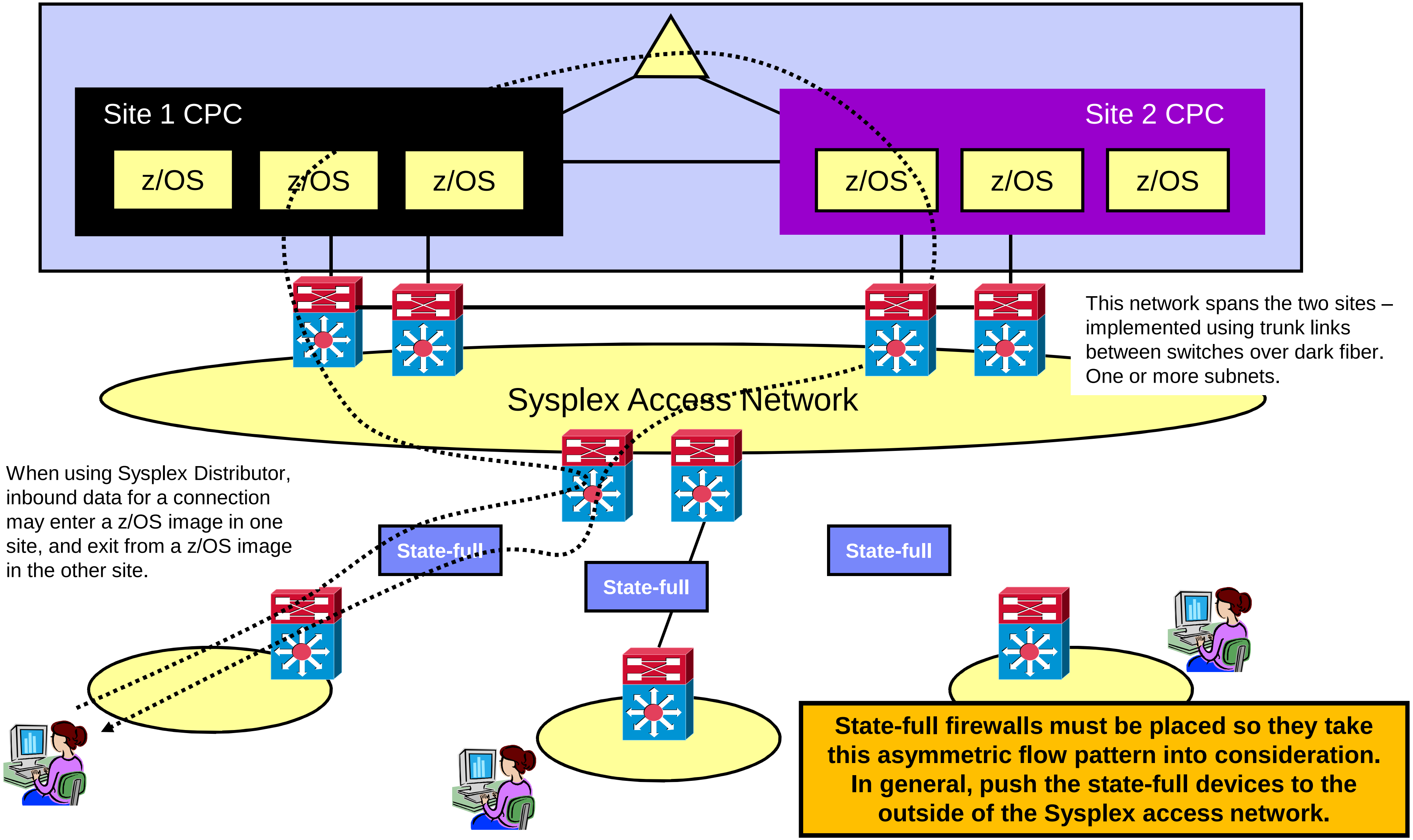
- ✓*Intended to benefit all existing Sysplex Distributor users*
- ✓*Request/Response data pattern (1K request, 1K response, 10 concurrent sessions) – RR10*
- ✓*Streaming data pattern (1/20M – 1 byte in, 20MB response, 20M/1 – 20MB in, 1 byte response)*
- ✓*Percentages relative to no acceleration (both using VIPAROUTE and 10GbE OSA Express 4)*



Note: The performance measurements discussed in this presentation are preliminary z/OS V2R2 Communications Server numbers and were collected using a dedicated system environment. The results obtained in other configurations or operating system environments may vary.

© 2015 SHARE and IBM Corporation

# State-full firewalls and multi-site Sysplex – shared Sysplex access network



This network spans the two sites – implemented using trunk links between switches over dark fiber. One or more subnets.

When using Sysplex Distributor, inbound data for a connection may enter a z/OS image in one site, and exit from a z/OS image in the other site.

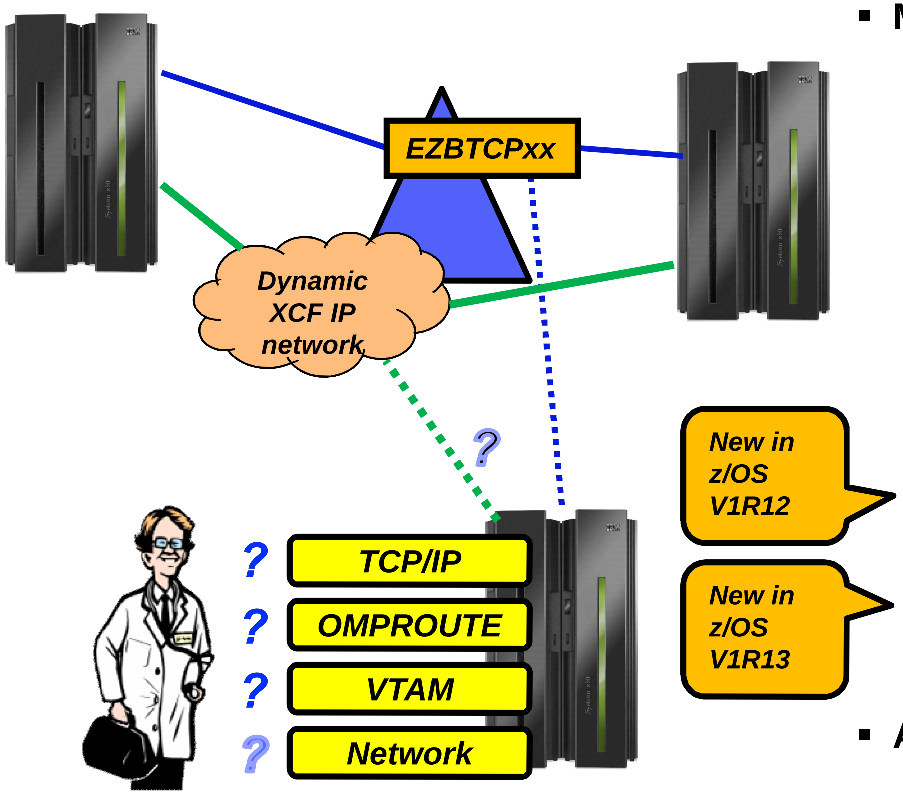
**State-full firewalls must be placed so they take this asymmetric flow pattern into consideration. In general, push the state-full devices to the outside of the Sysplex access network.**

© 2015 SHARE and IBM Corporation

# Networking Sysplex availability

© 2015 SHARE and IBM Corporation

# Sysplex autonomics extended with CSM storage constrained monitoring



EZBTCPxx

Dynamic XCF IP network

?

? TCP/IP

? OMPROUTE

? VTAM

? Network

New in z/OS V1R12

New in z/OS V1R13

- **Monitoring:**
  - Monitor CS health indicators
    - Storage usage critical condition (>90%) - CSM, TCPIP Private & ECSA
      - For more than TIMERSECS seconds
  - Monitor dependent networking functions
    - OMPROUTE availability
    - VTAM availability
    - XCF links available
  - Monitor for abends in Sysplex-related stack components
    - Selected internal components that are vital to Sysplex processing
      - Does not include "all" components
  - Selected network interface availability and routing
  - ***Monitor for repetitive internal abends in non-Sysplex related stack components***
    - ***5 times in less than 1 minute***
  - ***Detect when CSM FIXED or CSM ECSA has been constrained (>80% utilization) for multiple monitoring intervals***
    - ***For 3 times the TIMERSECS value***

- **Actions:**
  - Remove the stack from the IP Sysplex (manual or automatic)
    - 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 (manual or automatic)

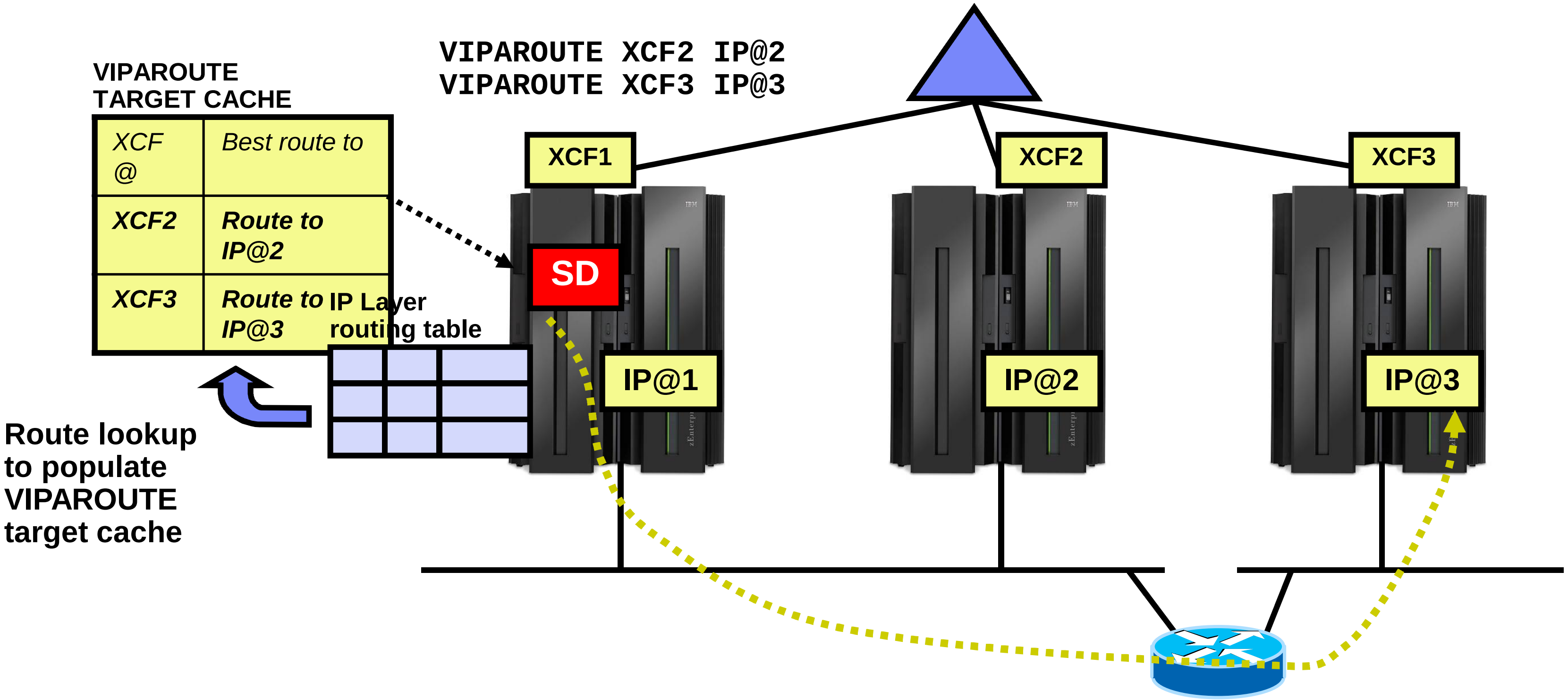
***Sick? Better remove myself from the IP Sysplex!***

***Feeling better? Maybe it's time to rejoin the IP Sysplex***

© 2015 SHARE and IBM Corporation

# VIPAROUTE target cache update during initialization – z/OS V1R13

- When using VIPAROUTE, a VIPAROUTE target cache is used to minimize the time it takes to route a Sysplex Distributor packet
- The target cache is updated every 60 seconds, which in some cases have caused delays during a primary stack's take-back of a distributed DVIPA
- z/OS V1R13 shortens the interval for VIPAROUTE route lookups in situations where the stack joins a Sysplex, or OMPROUTE is restarted
  - Will now start with 5 seconds, and gradually increase to 60 seconds



VIPAROUTE TARGET CACHE

| *XCF @* | *Best route to* |
|---|---|
| ***XCF2*** | ***Route to IP@2*** |
| ***XCF3*** | ***Route to IP@3*** |

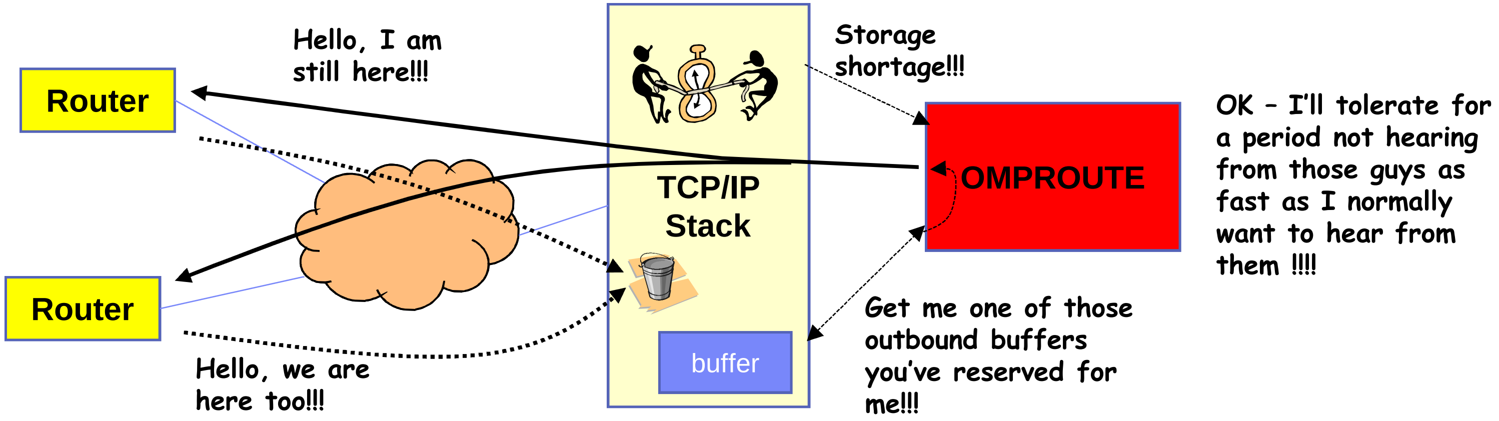
```
VIPAROUTE XCF2 IP@2
VIPAROUTE XCF3 IP@3
```

**Route lookup to populate VIPAROUTE target cache**

© 2015 SHARE and IBM Corporation

# z/OS V1R11 storage shortages and OMPROUTE availability

- OMPROUTE and the TCP/IP stack work together to make OMPROUTE more tolerant of storage shortage conditions:
  - TCP/IP stack informs OMPROUTE of stack storage shortage conditions
  - During a storage shortage, OMPROUTE temporarily suspends requirement for periodic routing updates from neighbor routers
  - TCP/IP stack ensures that dispatchable units for OMPROUTE can always obtain the control blocks  that they require
  - TCP/IP stack satisfies storage requests for OMPROUTE as long as storage remains available
- Temporarily keeps OMPROUTE from timing out routes due to lack of routing updates from neighbor routers during a storage shortage
- Decreases likelihood of OMPROUTE exiting or failing to send routing updates to neighbor routers


Hello, I am still here!!!
Router
Router
TCP/IP Stack
buffer
OMPROUTE
Storage shortage!!!
OK – I'll tolerate for a period not hearing from those guys as fast as I normally want to hear from them !!!!
Get me one of those outbound buffers you've reserved for me!!!
Hello, we are here too!!!

© 2015 SHARE and IBM Corporation

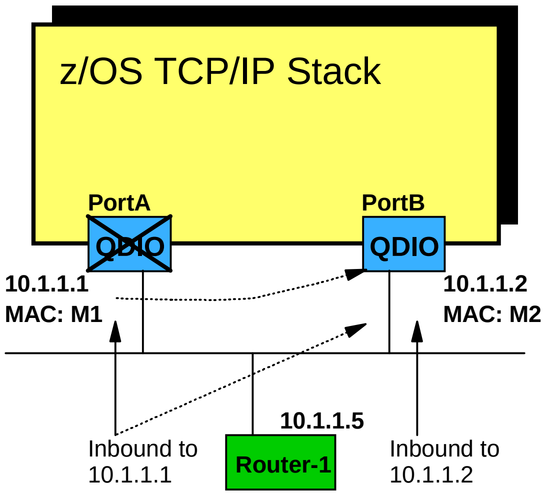
# Network availability in a flat network environment (No dynamic routing updates)

© 2015 SHARE and IBM Corporation

# Interface resilience without dynamic routing

**Requirement for this feature to function properly:**

- At least two adapters attached to the same network (broadcast media) - referred to as a LAN group.
- Adapters must use either LCS or QDIO
- The two adapters should be two physical adapters for real availability benefits



10.x.y.0/24

**Router's initial ARP Cache**

| IP address | Mac address |
|---|---|
| 10.1.1.1 | M1 |
| 10.1.1.2 | M2 |

**Router's ARP Cache after movement of 10.1.1.1 to PortB**

| IP address | Mac address |
|---|---|
| 10.1.1.1 | M2 |
| 10.1.1.2 | M2 |

## Example: PortA fails or is shut down

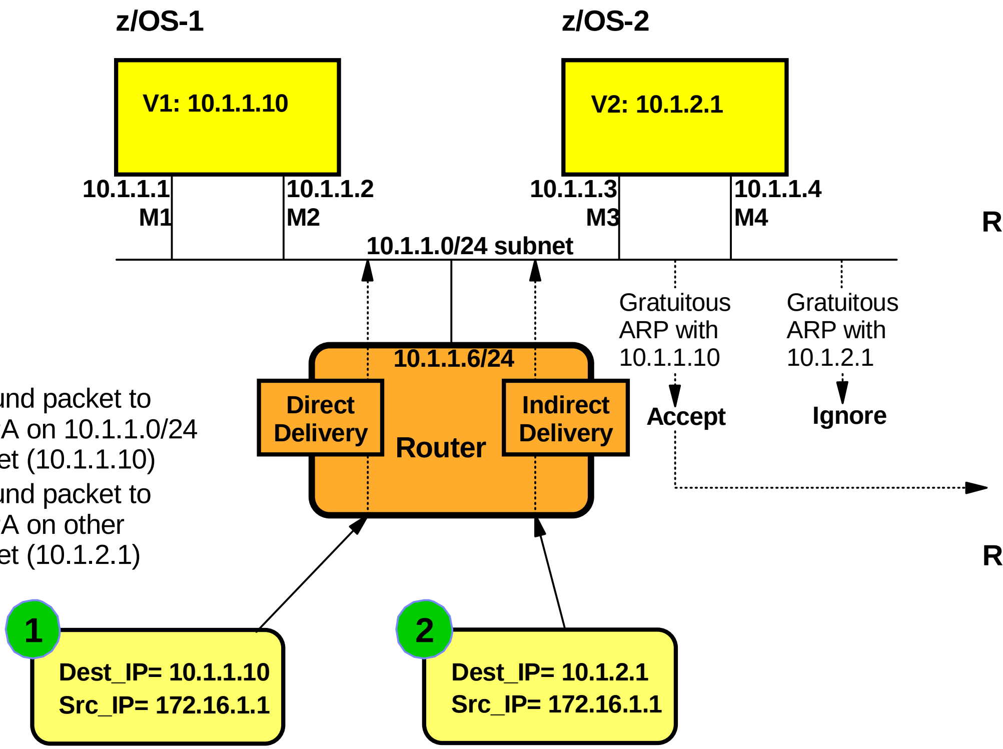
1. The z/OS TCP/IP stack moves address 10.1.1.1 to the other QDIO adapter (PortB), which is on the same network (same network prefix) as PortA was.
2. The z/OS TCP/IP stack issues a gratuitous ARP for IP address 10.1.1.1 with the MAC address of PortB (M2) over the PortB adapter
3. Downstream TCP/IP nodes on the same subnet with that IP address in their ARP cache, will update their ARP caches to point to M2 for IP address 10.1.1.1 and will thereafter send inbound packets for both 10.1.1.1 and 10.1.1.2 to MAC address M2

© 2015 SHARE and IBM Corporation

# Some (restricted) support of dynamic VIPA without dynamic routing

**Note**: Some router vendors may have added capabilities in this area that make their product work slightly different from what is described here!

1. Inbound packet to DVIPA on 10.1.1.0/24 subnet (10.1.1.10)
2. Inbound packet to DVIPA on other subnet (10.1.2.1)



**Router's ARP Cache**

| IP address | Mac address |
|---|---|
| 10.1.1.1 | M1 |
| 10.1.1.2 | M2 |
| 10.1.1.3 | M3 |
| 10.1.1.4 | M4 |
| 10.1.1.10 | M1 |

**Router's IP Routing Table**

| Destination | Via |
|---|---|
| 10.1.1.0/24 | Direct delivery |
| 10.1.2.0/24 | 10.1.1.3 |
| Default | n.n.n.n |

z/OS VIPA addresses in a flat network configuration without dynamic routing must be allocated out of the same subnet as the directly attached network that all members of the Sysplex are attached to - in this example, the 10.1.1.0/24 subnet.

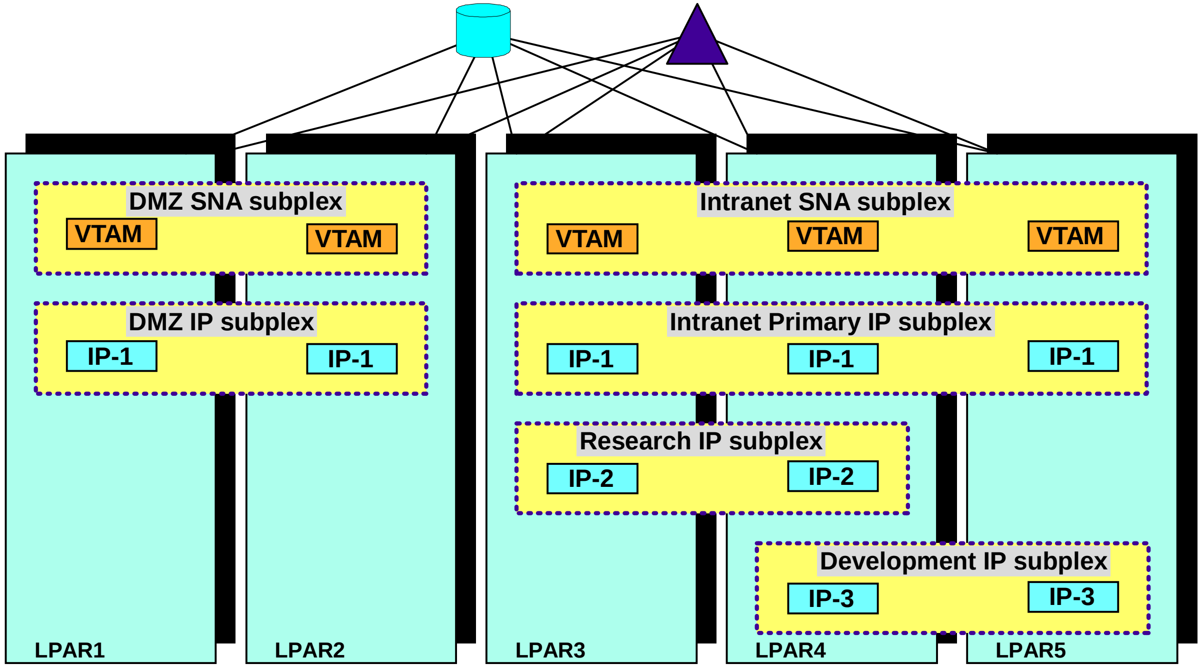
All LPARS must be attached to one and the same IP subnet via OSA ports. Network interfaces belonging to other IP subnets cannot be used for re-routing around failed OSA ports. Availability of the network to which the OSA ports are attached becomes of outmost importance and must generally be based on what is known as Layer-2 availability functions in the switches.

© 2015 SHARE and IBM Corporation

# Network Subplex support

© 2015 SHARE and IBM Corporation

# Networking sub-plexing within a z/OS Sysplex



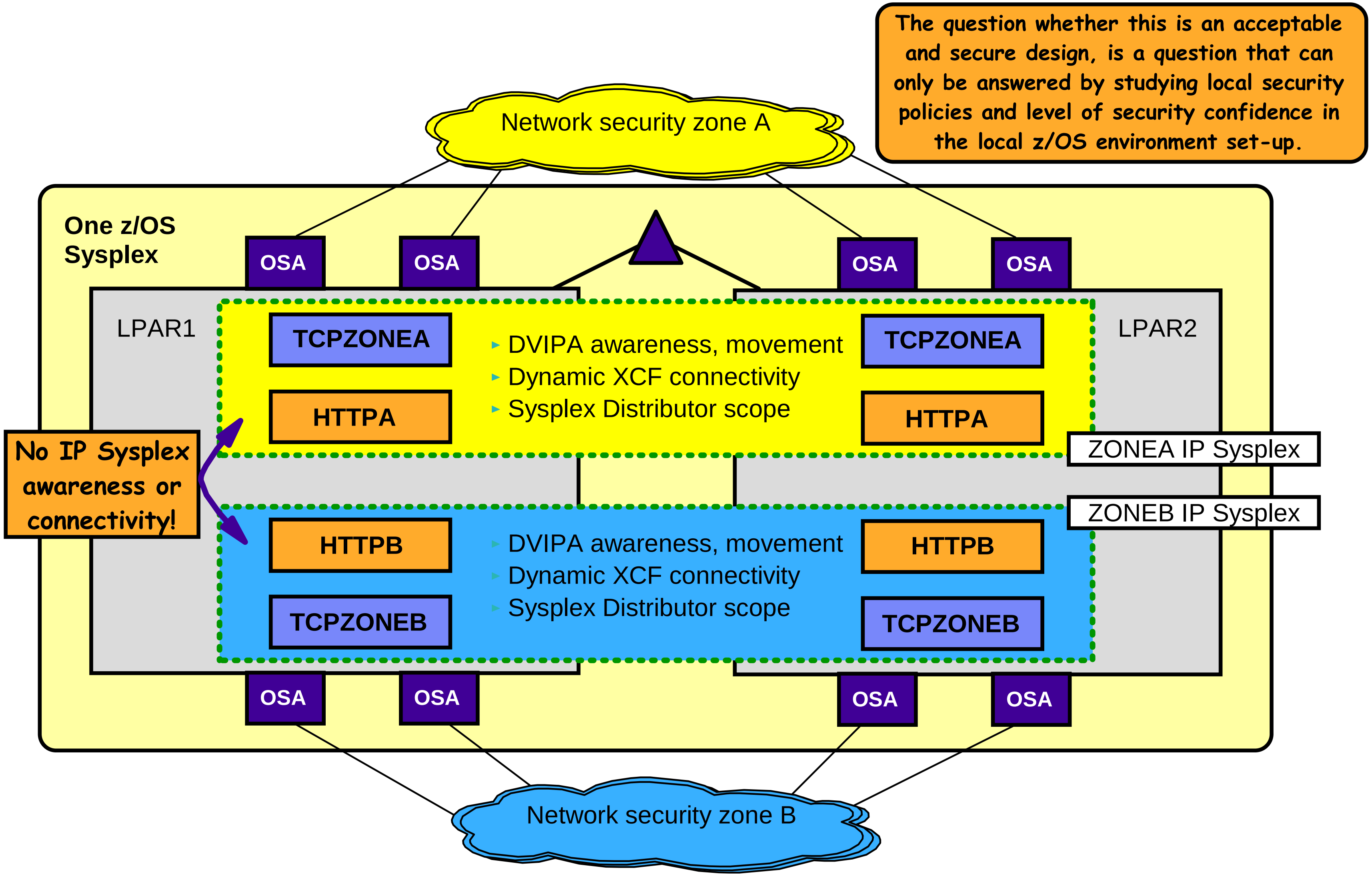
**Networking Subplexing - Sysplex partitioning from a network perspective**

- One SNA subplex per LPAR
- An IP subplex cannot span multiple SNA subplexes
- Different IP stacks in an LPAR may belong to different IP subplexes
- Standard RACF controls for stack access and application access to z/OS resources need to be in place.

- **Networking subplex scope:**
  - VTAM Generic Resources (GR) and Multi-Node Persistent Session (MNPS) resources
  - Automatic connectivity - IP connectivity and VTAM connectivity over XCF (including dynamic IUTSAMEH and dynamic HiperSockets based on Dynamic XCF for IP)
    - HiperSockets VLANID support also added as part of this support
  - IP stack IP address (including dynamic VIPA) awareness and visibility
  - Dynamic VIPA movement candidates
  - Sysplex Distributor target candidates

© 2015 SHARE and IBM Corporation

# An example of sub-plexing within a z/OS Sysplex



The question whether this is an acceptable and secure design, is a question that can only be answered by studying local security policies and level of security confidence in the local z/OS environment set-up.

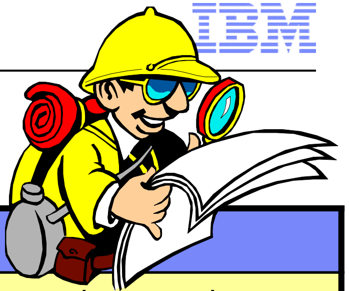
No IP Sysplex awareness or connectivity!

- DVIPA awareness, movement
- Dynamic XCF connectivity
- Sysplex Distributor scope

- DVIPA awareness, movement
- Dynamic XCF connectivity
- Sysplex Distributor scope

© 2015 SHARE and IBM Corporation

# For more information



| URL | Content |
|---|---|
| http://www.twitter.com/IBM_Commserver | IBM z/OS Communications Server Twitter Feed |
| http://www.facebook.com/IBMCommserver | IBM z/OS Communications Server Facebook Page |
| https://www.ibm.com/developerworks/mydeveloperworks/blogs/IBMCommserver/?lang=en | IBM z/OS Communications Server Blog |
| http://www.ibm.com/systems/z/ | IBM System z in general |
| http://www.ibm.com/systems/z/hardware/networking/ | IBM Mainframe System z networking |
| http://www.ibm.com/software/network/commserver/ | IBM Software Communications Server products |
| http://www.ibm.com/software/network/commserver/zos/ | IBM z/OS Communications Server |
| http://www.redbooks.ibm.com | ITSO Redbooks |
| http://www.ibm.com/software/network/commserver/zos/support/ | IBM z/OS Communications Server technical Support – including TechNotes from service |
| http://www.ibm.com/support/techdocs/atsmastr.nsf/Web/TechDocs | Technical support documentation from Washington Systems Center (techdocs, flashes, presentations, white papers, etc.) |
| http://www.rfc-editor.org/rfcsearch.html | Request For Comments (RFC) |
| http://www.ibm.com/systems/z/os/zos/bkserv/ | IBM z/OS Internet library – PDF files of all z/OS manuals including Communications Server |
| http://www.ibm.com/developerworks/rfe/?PROD_ID=498 | RFE Community for z/OS Communications Server |
| https://www.ibm.com/developerworks/rfe/execute?use_case=tutorials | RFE Community Tutorials |

***For pleasant reading ….***

© 2015 SHARE and IBM Corporation

# Please fill out your session evaluation

- Sysplex Networking Technologies and Considerations
- Session # 16472
- QR Code:

IBM Comm Server

Find us on Facebook at
http://www.facebook.com/IBMCommserver

Follow us on Twitter at
http://www.twitter.com/IBM_Commserver

Read the z/OS Communications Server blog at
http://tinyurl.com/zoscsblog

Visit the z/OS CS YouTube channel at
http://www.youtube.com/user/zOSCommServer

© 2015 SHARE and IBM Corporation

## Sysplex Networking Technologies and Considerations

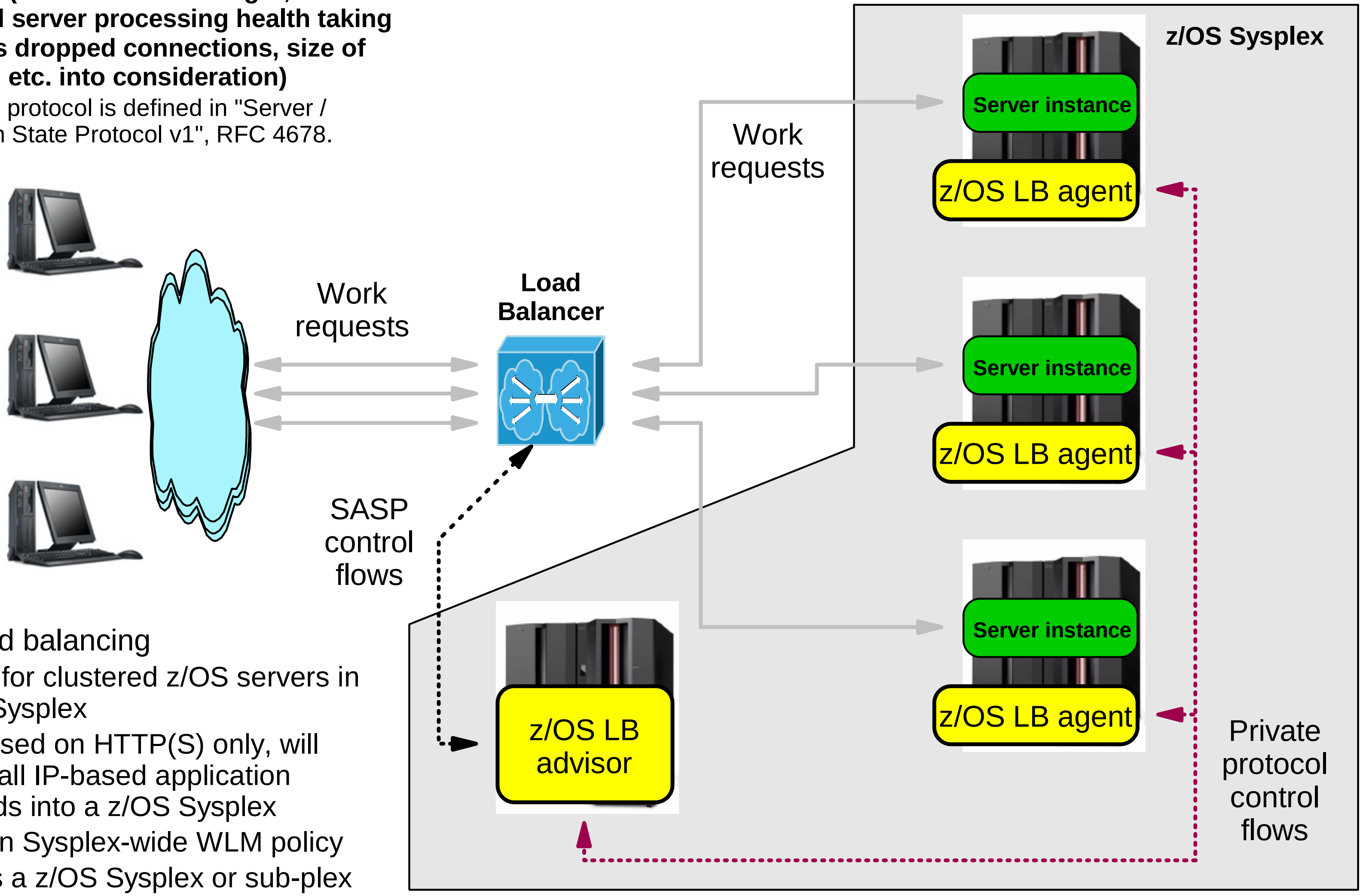
# Appendix

© 2015 SHARE and IBM Corporation

# z/OS support of external load balancers

**The SASP control flows provide relative weights per server instance (based on WLM weight, server availability, and server processing health taking such metrics as dropped connections, size of backlog queue, etc. into consideration)**

- The SASP protocol is defined in "Server / Application State Protocol v1", RFC 4678.



z/OS workload balancing

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

© 2015 SHARE and IBM Corporation

# Basing automated operations on TCP/IP start-up messages

```
GlobalConfig SysplexMonitor  ; Enable Sysplex autonomics
             TimerSecs 60    ; Check interval (default)
             AutoRejoin      ; Rejoin automatically
             DelayJoin       ; Delay joining till OMPROUTE is up
             DynRoute        ; Interface mon w. dyn routes
             MonInterface    ; Interface monitoring
             Recovery        ; Remove myself automatically
;
DEVICE OSAQDIO4  MPCIPA
LINK   QDIO4     IPAQENET    OSAQDIO4 MONSYSPLEX
;
INTERFACE QDIO6
  DEFINE IPAQENET6
  PORTNAME OSAQDIO4
  MONSYSPLEX
```

This set of SysplexMonitor definitions will automatically leave and join the Sysplex based on the availability and health of selected Sysplex TCP/IP resources.

When starting TCP/IP, the stack will not join the Sysplex until OMPROUTE is up and running and has learned dynamic routes over at least one monitored network interface (those coded with the MONSYSPLEX keyword)

```
*EZD1166E TCPCS DELAYING SYSPLEX PROFILE PROCESSING - OMPROUTE IS NOT
 ACTIVE
```

```
*EZD1211E TCPCS DELAYING SYSPLEX PROFILE PROCESSING - ALL MONITORED
 INTERFACES WERE NOT ACTIVE
```

```
*EZD1212E TCPCS DELAYING SYSPLEX PROFILE PROCESSING - NO DYNAMIC ROUTES
 OVER MONITORED INTERFACES WERE FOUND
```

```
 EZD1176I TCPCS HAS SUCCESSFULLY JOINED THE TCP/IP SYSPLEX GROUP EZBTCPCS
 EZD1214I INITIAL DYNAMIC VIPA PROCESSING HAS COMPLETED FOR TCPCS
```

Some applications do resolver calls when they start up. If they are started after TCPIP is up, but OMPROUTE has not learned the needed routes, resolver calls that need to use the DNS may fail. So, there is a need to start these applications after a route has been learned.
If you are not using AUTOLOG with DELAYSTART DVIPA to start server address spaces, let your automation software kick off on the EZD1214I message.

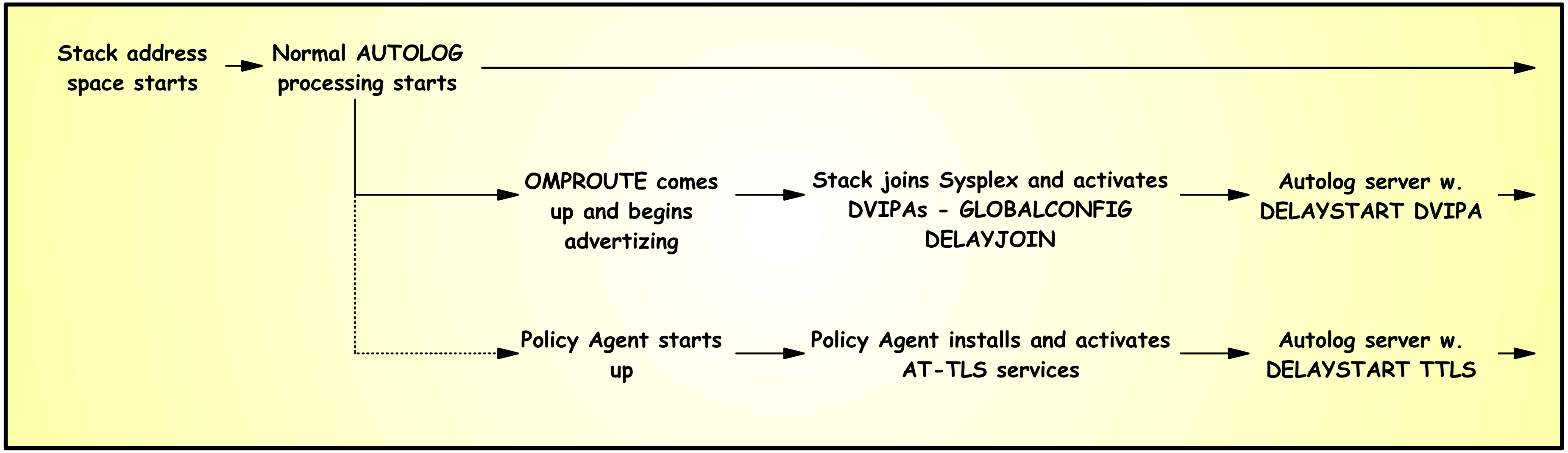
© 2015 SHARE and IBM Corporation

# z/OS V1R10 implemented improved AUTOLOG sequencing of the TCP/IP start-up process

- **The pre-V1R10 AUTOLOG DELAYSTART option delays application start until DVIPAs are configured and active**
- **z/OS V1R10 adds another option to AUTOLOG DELAYSTART that can be used to delay start of an application until AT-TLS services are available**

```
           +-DVIPA-+
DELAYSTART-+-------+----
           +-TTLS--+
```

**If DELAYSTART is specified without a sub-option, it defaults to DVIPA**



Stack address space starts → Normal AUTOLOG processing starts →

OMPROUTE comes up and begins advertizing → Stack joins Sysplex and activates DVIPAs - GLOBALCONFIG DELAYJOIN → Autolog server w. DELAYSTART DVIPA →

Policy Agent starts up → Policy Agent installs and activates AT-TLS services → Autolog server w. DELAYSTART TTLS →

© 2015 SHARE and IBM Corporation