Sysplex Networking Technologies and Considerations

Gus Kassimis - kassimis@us.ibm.com

IBM Enterprise Networking Solutions
Raleigh, NC, USA

Session: 14682
Monday, March 10, 2014: 4:30 PM-5:30 PM

SHARE
in Anaheim

Sysplex Networking Technologies and Considerations

Session number:	14682
Date and time:	Tuesday, March 11, 2014: 4:30 PM-5:30 PM
Program:	Enterprise Data Center
Project:	Communications Server
Track:	Tracks: Network Systems and z/OS Systems Programming
Classification:	Technical
Speaker:	Gus Kassimis, IBM
Abstract:	In this session, we will review z/OS Communication Server networking technologies that help maximize the value of the parallel sysplex environment. Recent enhancements in the area of High Availability and workload balancing will be discussed, including considerations for enabling these features. Topics will also include best practices recommendations for intra-sysplex connectivity and external connectivity into the sysplex environment, including techniques for achieving network interface resilience in a flat network and the impacts of using state-full firewalls in the Sysplex access network.

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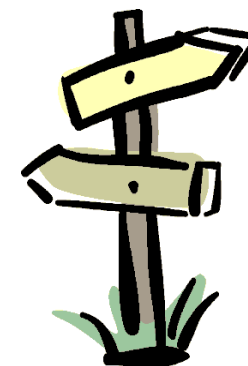
Disclaimer: All statements regarding IBM future direction or intent, including current product plans, are subject to change or withdrawal without notice and represent goals and objectives only. All information is provided for informational purposes only, on an "as is" basis, without warranty of any kind.

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

My virtual z/OS host

VIPA#2
CICS Appl-A

VIPA#1
TN3270e Server

VIPA#4
DB2 subsystem

VIPA#3
FTP Services

VIPA#5
Web Services

OSA
IP#10

OSA
IP#11

OSA
SNA LLC2
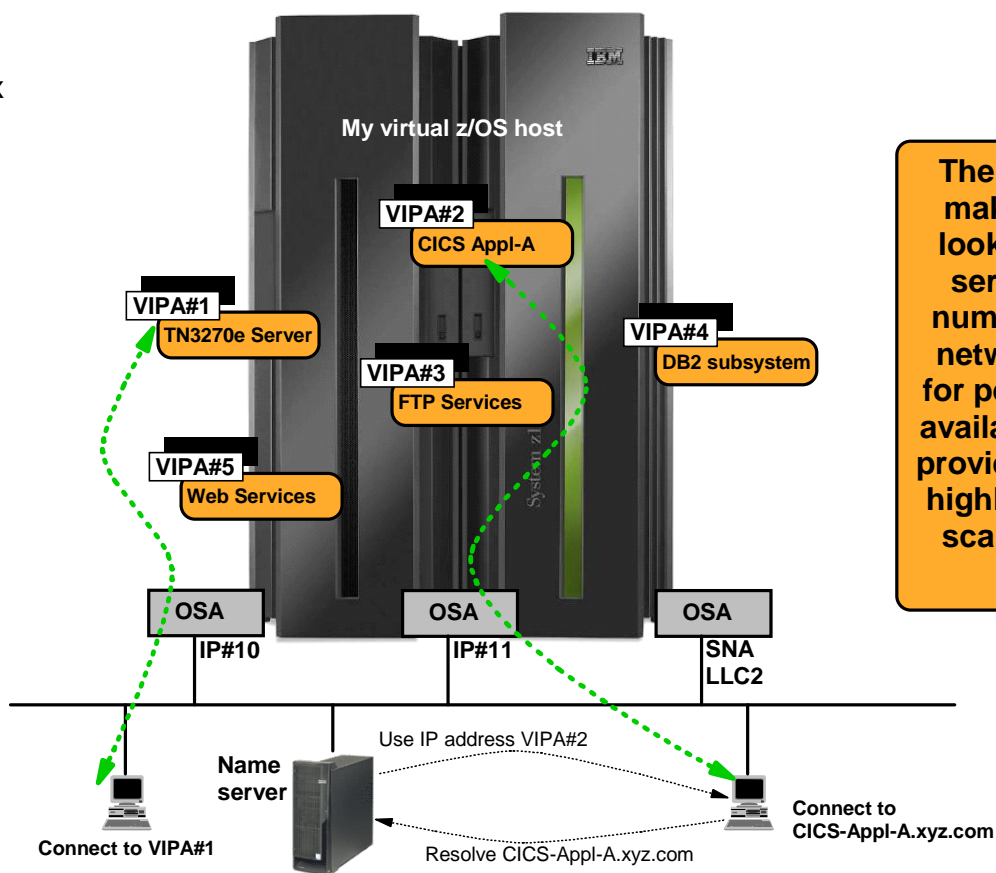
The objective is to make the Sysplex look like one large server that has a number of physical network interfaces for performance and availability - and that provides a number of highly available and scalable services.

Use IP address VIPA#2

Name server

Connect to VIPA#1

Resolve CICS-Appl-A.xyz.com

Connect to CICS-Appl-A.xyz.com

SNA and TCP/IP

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

Sysplex Networking Technologies and Considerations

Workload Balancing Considerations

IBM®

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.

Server Cluster

Load Balancing decision maker

Server

Server

Server

Server

Data

Data

Mirroring w. HyperSwap

Coupling Facility

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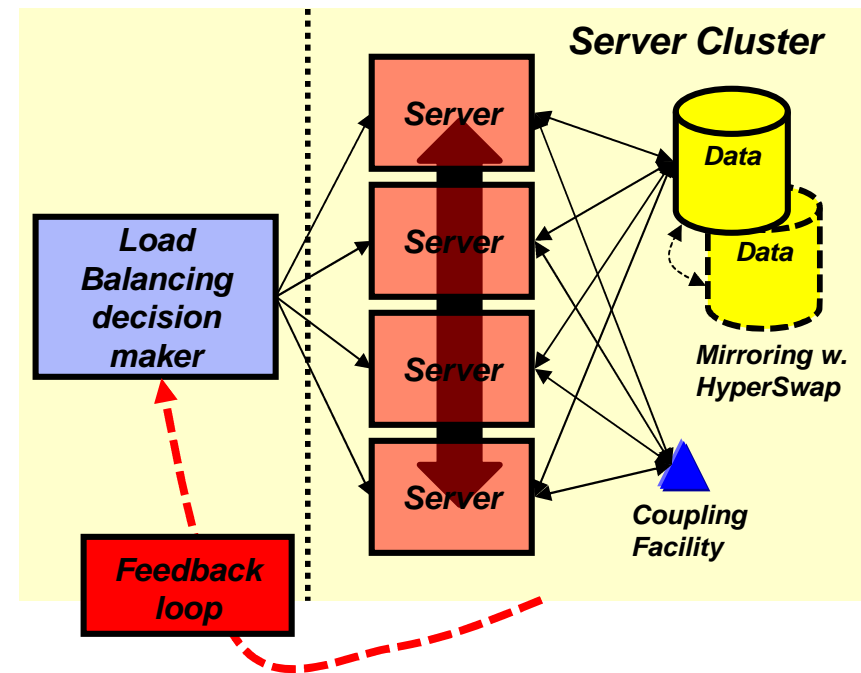
In order for the load balancing decision maker to meet those objectives, it must be capable of obtaining feedback dynamically, such as server instance availability, capacity, performance, and overall health.

© 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
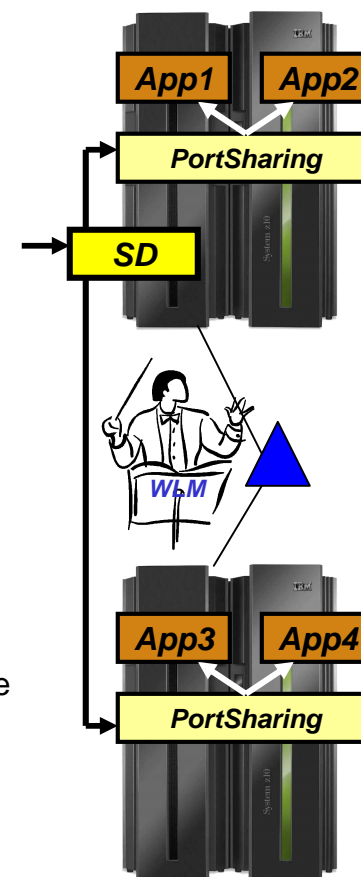
Sysplex Distributor distribution method overview

- 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
    - Method added in z/OS V1R9

- 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 as well in z/OS V1R8
    - *Generally, the recommended distribution method for Sysplex Distributor*

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.
      - Method added in *z/OS V1R11*

  - 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*

The full Sysplex Distributor recommendation mechanism
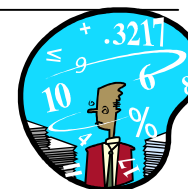
- **TSR**: Target Server Responsiveness fraction, which is a compound health-metric per target server (range from 0 (bad) to 100 (good)):
  - **TCSR**: Target Connectivity Success Rate. Connectivity between the distributing stack and the target stack - are the new connection requests reaching the target? (0 is bad, 100 is good)
  - **SEF**: Server accept Efficiency Fraction. Target Server accept efficiency - is the server accepting new work? (0 is bad, 100 is good)
  - **QoS**: QoS fractions. Taking retransmits and packet loss into consideration. (0 is good, 100 is bad)
  - Number of half-open connections also impact the TSR value
- **CER**: Connection Establishment Rate. Network connectivity between Server and client - are new connections being established? (0 is bad, 100 is good)
  - CER no longer impacts the TSR value, but is still calculated and included in netstat displays
  - Connection establishment problem detection is now part of SEF
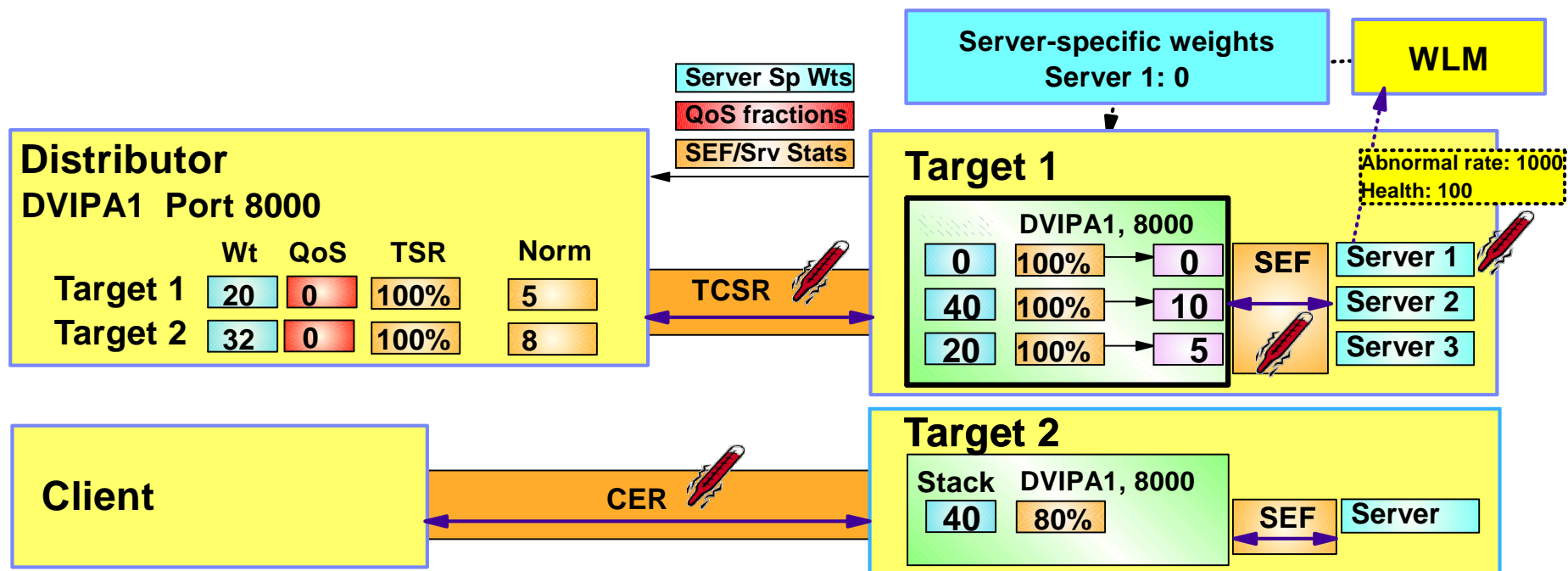
What impacts the final selection of a target server instance?

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 (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 (SD agent)	No	No	No	No
SD HOTSTANDBY	No	No	Yes	Yes	No
PORT SHAREPORT	No	No	No	Yes (Only SEF value)	No
PORT SHAREPORTWLM	No	Yes	Yes	Yes (Only SEF value)	No

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

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

Importance Levels

☐ LPAR1 ■ LPAR2
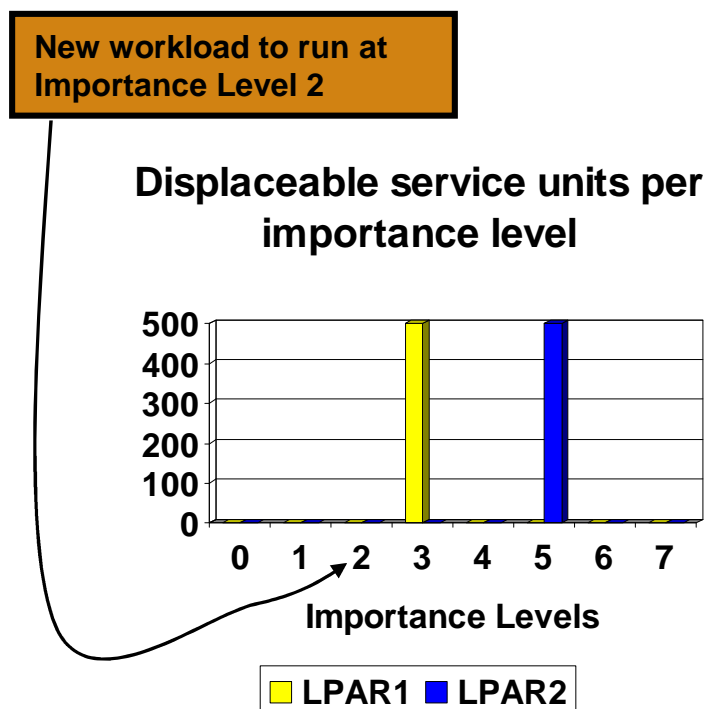
Adjusted displaceable service units

IL weighting factors (WF)

☐ LPAR1 ■ LPAR2
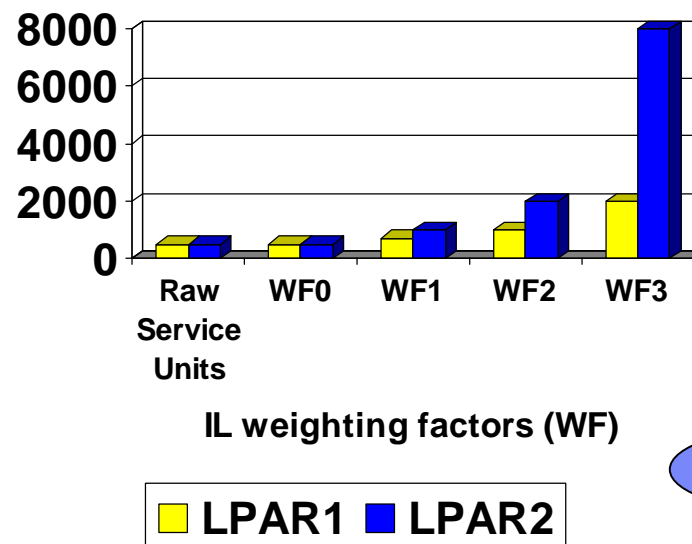
z/OS V1R11

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
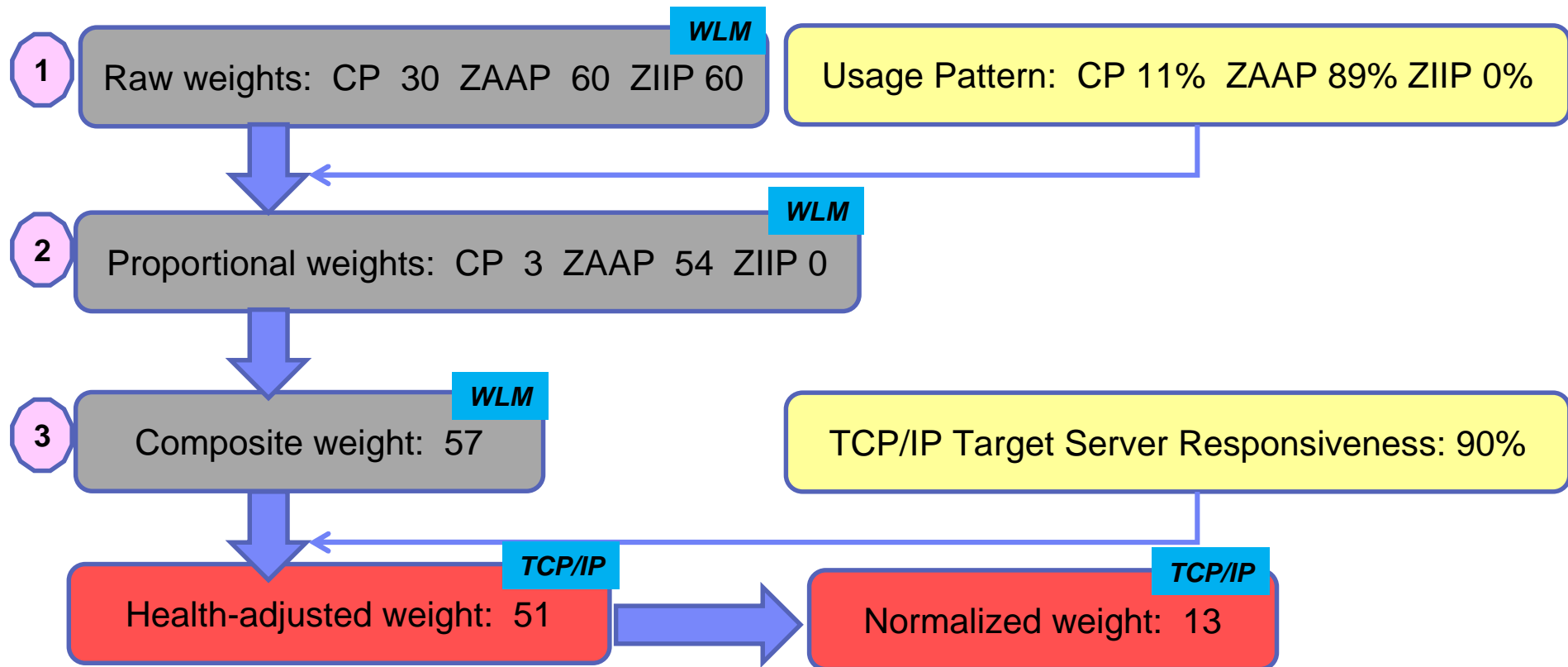
SERVERWLM method: zIIP and zAAP cross-over to CP

- 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

Application Workload Design

Displaceable Service Units

Relative weights of LPAR1 and LPAR2

Weight

Cross-over Cost

LPAR1
LPAR2
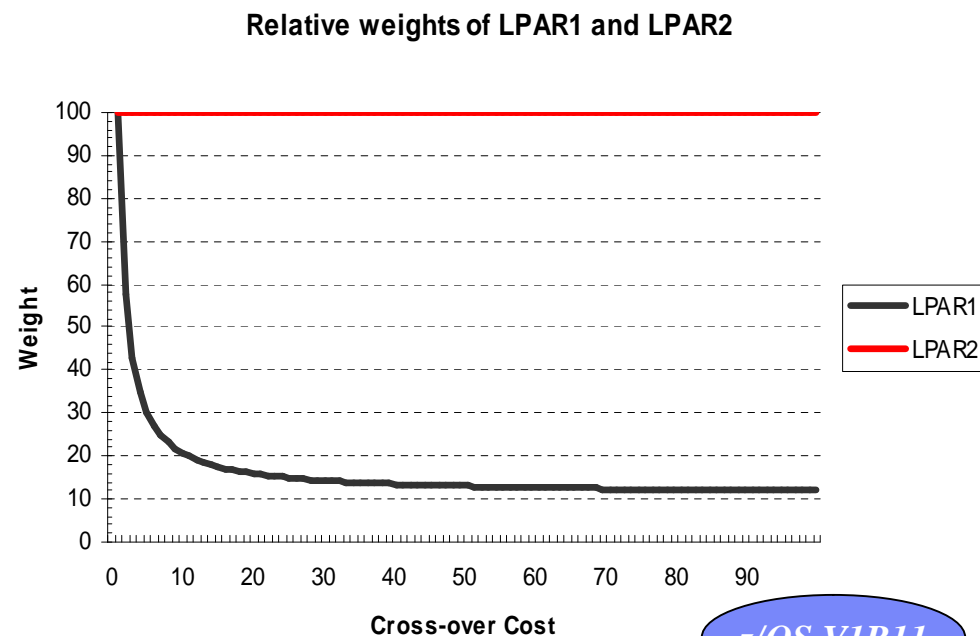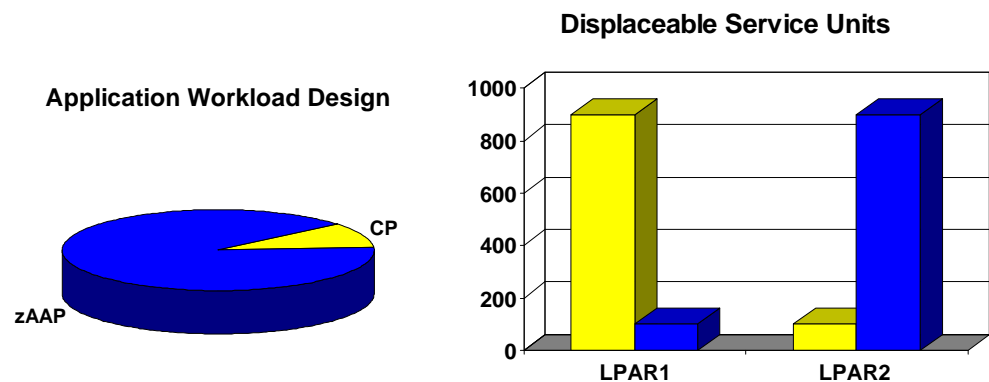
z/OS V1R11

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
```

Sysplex Distributor hot standby support

- 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

DB2 Data sharing group

DB2

DB2

Preferred Target server

Backup Target server

Sysplex Distributor
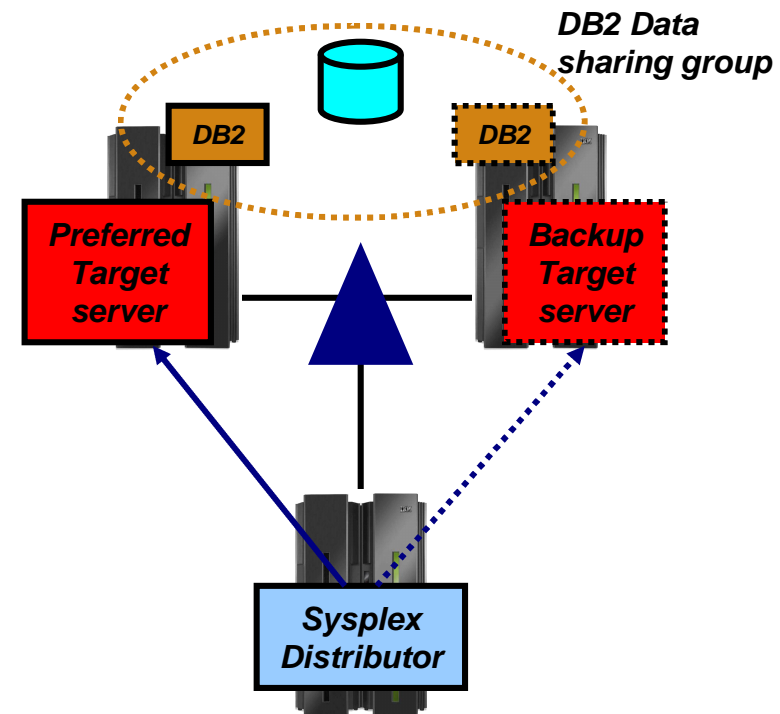
Sysplex Distributor hot standby support…

- 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
```

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:
```

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
```

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

z/OS1

DB2A

VIPAx P=446
VIPAa P=446
VIPAa P=5001

DB2B

VIPAx P=446
VIPAb P=446
VIPAb P=5002

2   1   1

Shareport

1

SD
VIPAx

1

2

DB2 client

1) Connect (VIPAx, 446)

2) Connect (VIPAb, 446)

SRVLIST
VIPAa Weight
VIPAb Weight
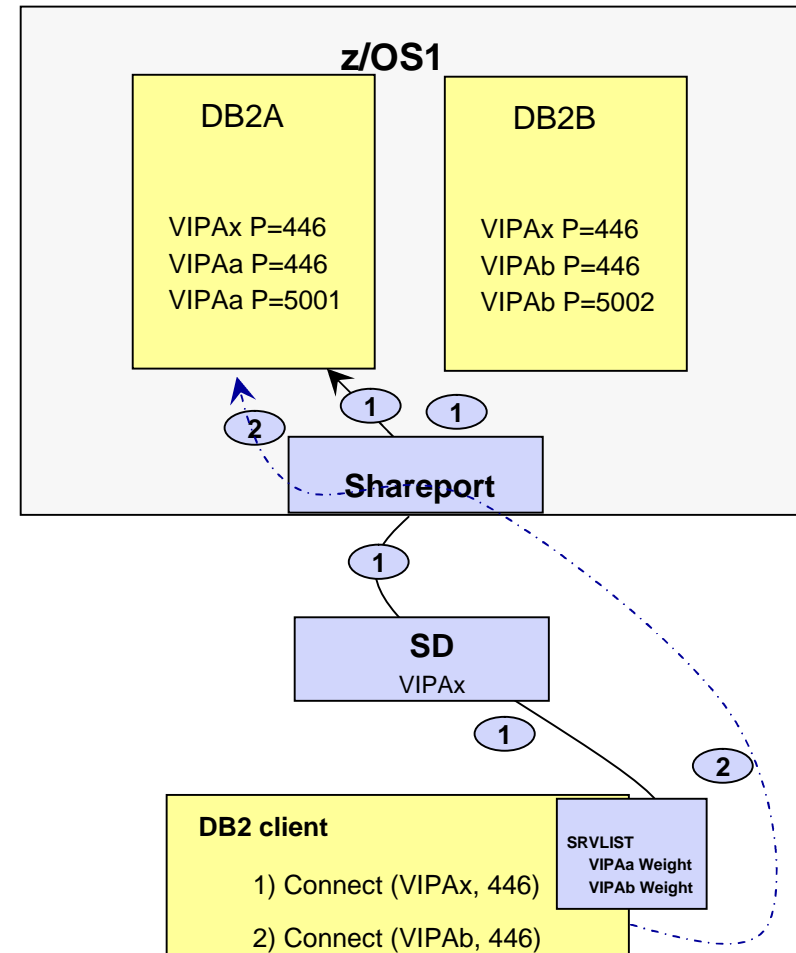
Affinity for application-instance DVIPAs - Solution

- **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)

z/OS1

DB2A

VIPAx P=446
VIPAa P=446
VIPAa P=5001

DB2B

VIPAx P=446
VIPAb P=446
VIPAb P=5002

Shareport

App-spec DVIPA	Owner
VIPAa	DB2A
VIPAb	DB2B

SD
VIPAx

DB2 client

1) Connect (VIPAx, 446)

2) Connect (VIPAb, 446)

SRVLIST
VIPAa Weight
VIPAb Weight
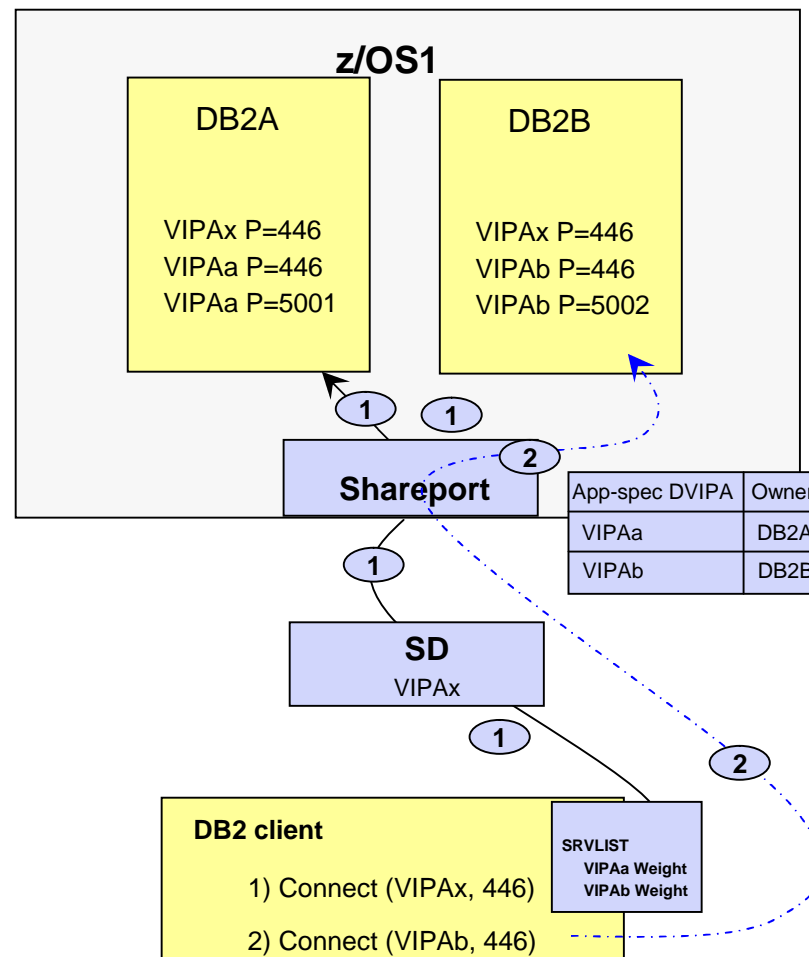
Sysplex Networking Technologies and Considerations

Intra-Sysplex connectivity

IBM

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.
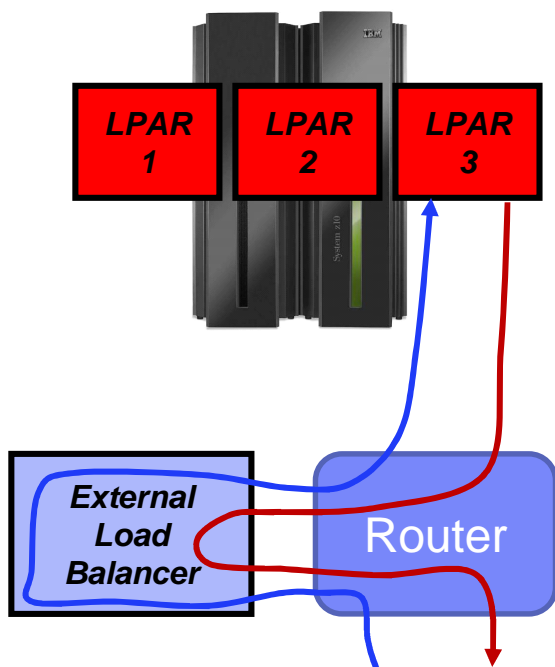
Sysplex Distributor

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.

Sideband signaling of connection status

LPAR 1   LPAR 2   LPAR 3

External Load Balancer

Router

Sysplex Distributor

Router
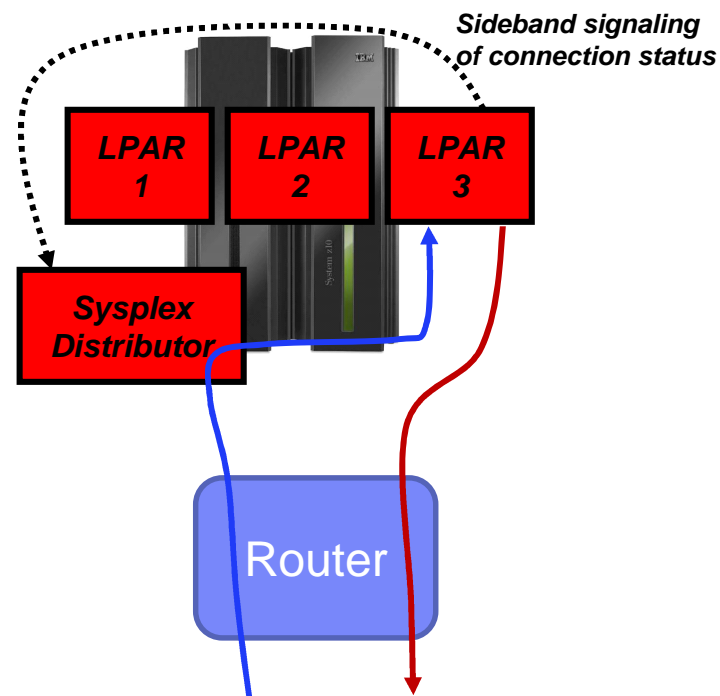
Two types of intra-Sysplex/subplex routing

- **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

10.1.1.1

Application

TCP/UDP

Connection routing table

IP-layer routing table

IP

DLC

10.2.1.1
[10.1.1.1]

10.3.1.1
[10.1.1.1]

DEST=10.1.1.1

DEST=10.1.1.1

Connection routing will occur even when IPCONFIG NODATAGRAMFWD is coded

10.1.1.1

Application

TCP/UDP

IP-layer routing table

IP

DLC

10.2.1.1

10.3.1.1

DEST=10.2.1.1

IPCONFIG DATAGRAMFWD must be set for normal IP routing to occur
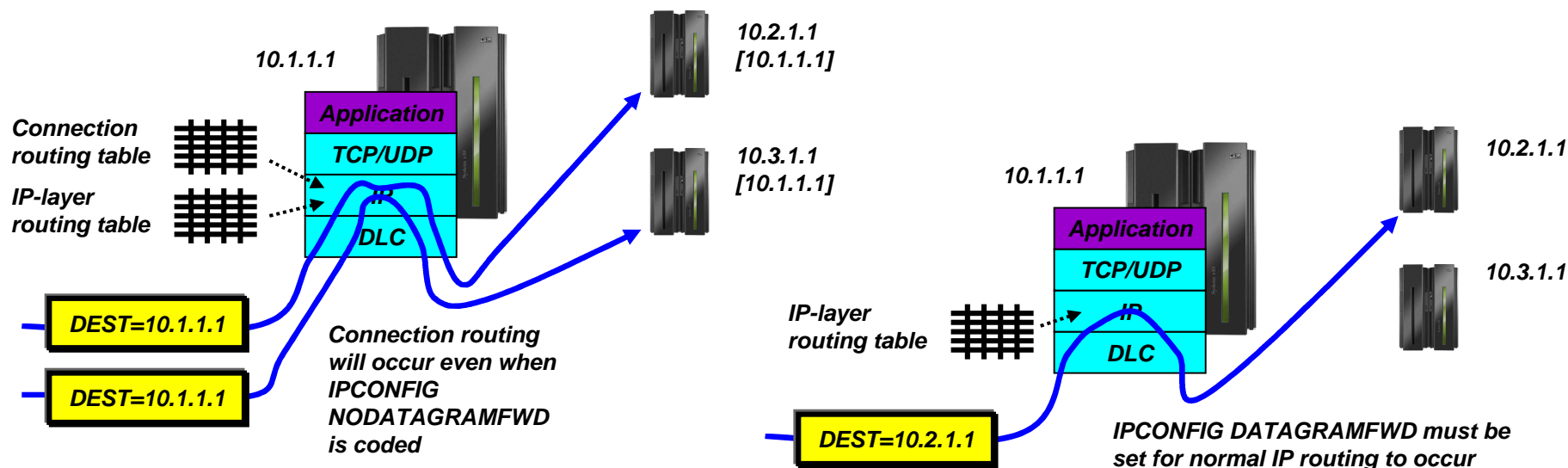
The role of XCF, ISTIQDIO HiperSockets, and external LAN interfaces in a z/OS Sysplex/subplex

- **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.

XCF connectivity

CPC1

DynamicXCF Interface — ISTIQDIO HiperSockets — DynamicXCF Interface

A HiperSockets LAN

z/OS LPAR1   OSA   z/OS LPAR2   OSA

CPC2

DynamicXCF Interface

z/OS LPAR3   OSA

External LAN – 1 GbE or 10 GbE
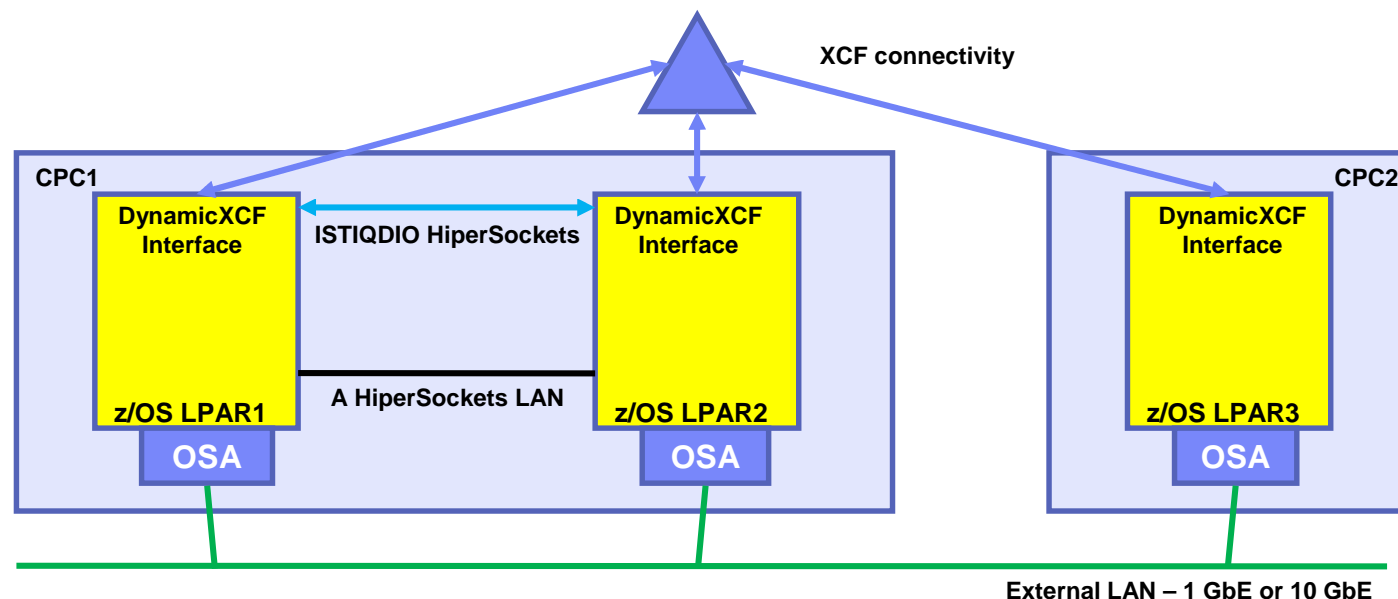
So what should I use for what type of routing?

- **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)
ISTIQDIO (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

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

GRE Encapsulation

IPv4 Delivery Header	20 Bytes
GRE Header	4 Bytes
Original IP Packet	nn Bytes

XCF connectivity

CPC1

SD

MTU: 32K

A HiperSockets LAN

z/OS LPAR1   z/OS LPAR2

OSA   OSA

CPC2

Appl.

z/OS LPAR3

OSA

MTU: 1492

IP Packet

IP-dev | GRE | IP Packet

MTU: 8892

External LAN – 1 GbE or 10 GbE
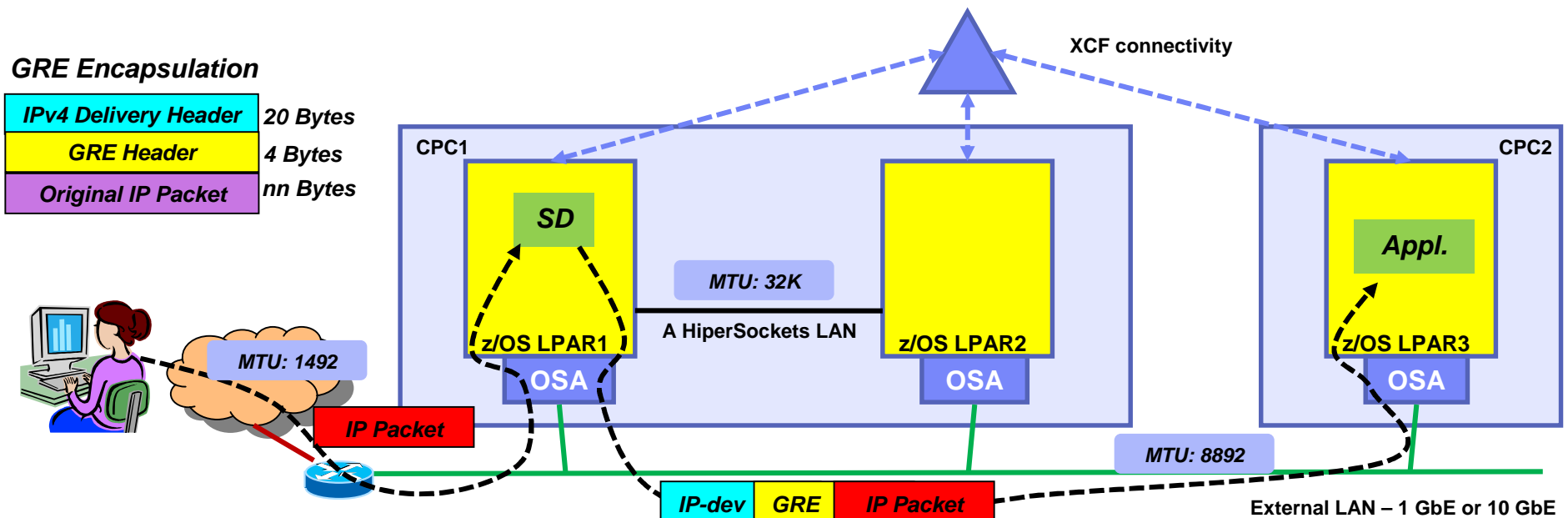
z/OS V1R12 - Enhance Packet Trace for Sysplex Distributor VIPAROUTE traffic

- 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

Local NMI Applications

TRACE

SMF EVENTS

QUERY

z/OS Communications Server

NMI
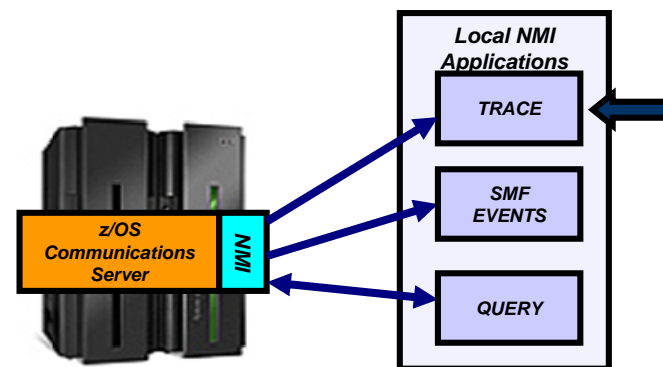
z/OS V1R11 QDIO and iQDIO routing accelerator

- **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)

Connection routing table

IP-layer routing table

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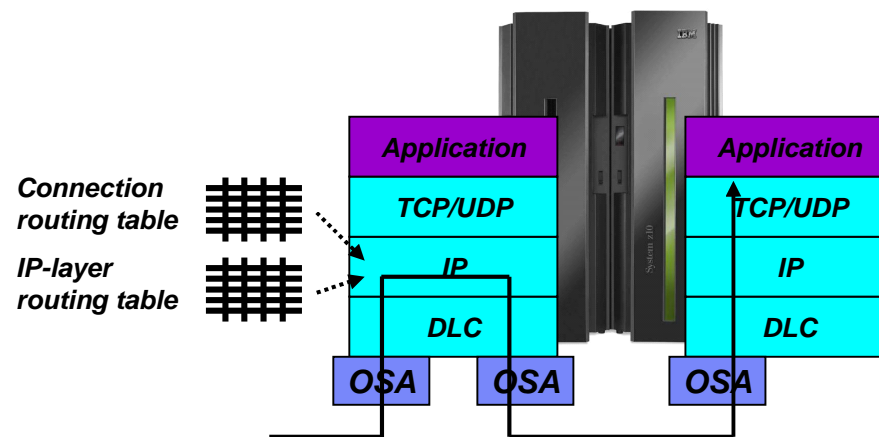
Shadow copies of selected entries in Connection routing table and IP-layer routing table

Application		Application
TCP/UDP		TCP/UDP
IP		IP
DLC		DLC
OSA	OSA	OSA
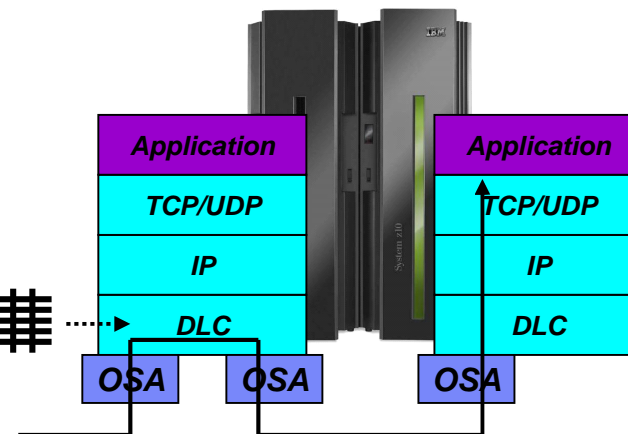
QDIO acceleration coexistence with IP filtering …

- 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		

Sysplex Distributor connection routing may benefit from QDIO/iQDIO accelerator
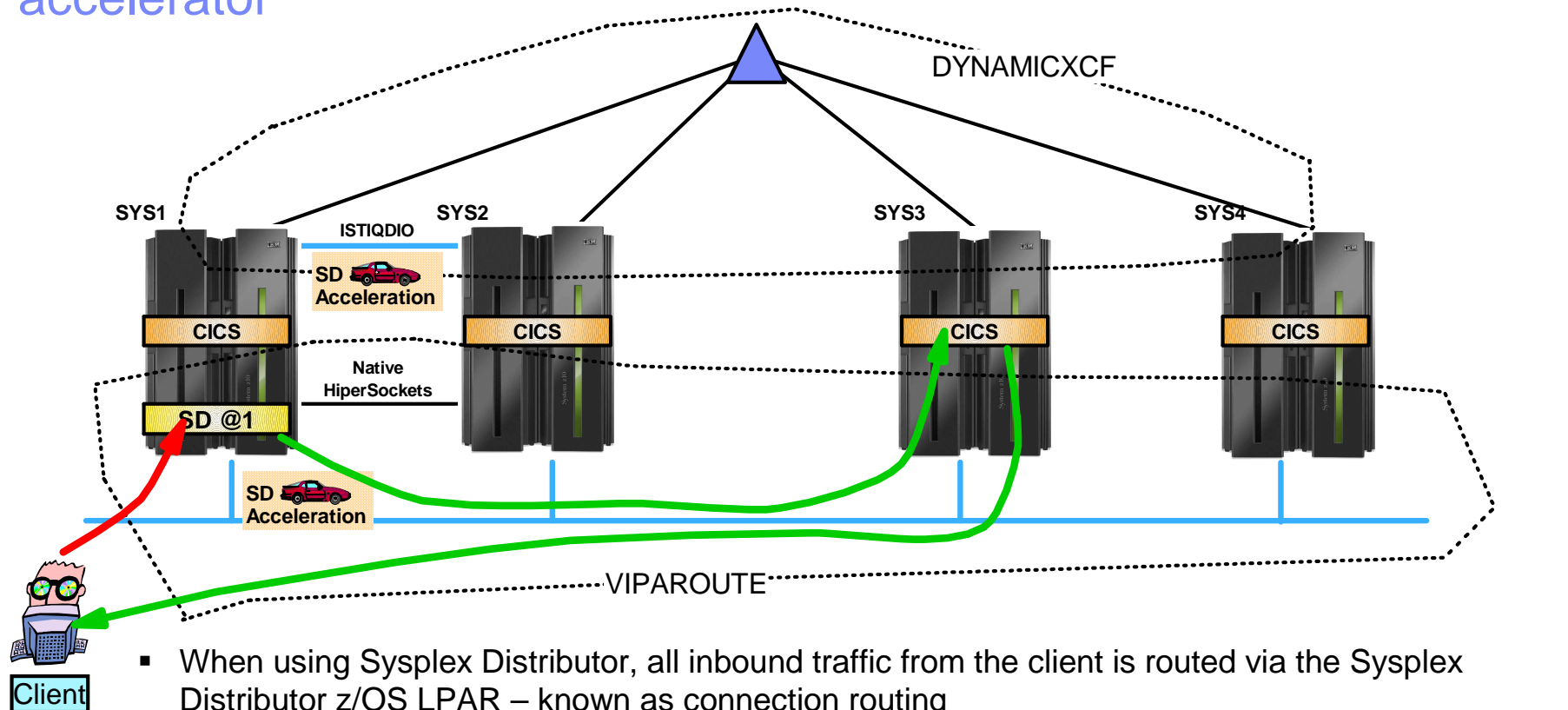
- ▪ When using Sysplex Distributor, all inbound traffic from the client is routed via the Sysplex Distributor z/OS LPAR – known as connection routing

  - – 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)

Sysplex Distributor accelerator performance

✓*Intended to benefit all existing Sysplex Distributor users*
✓*Measurements with Interactive workload (RR20)*
✓*Small data sizes (100 in, 800 out)*
✓*Percentages relative to no acceleration*
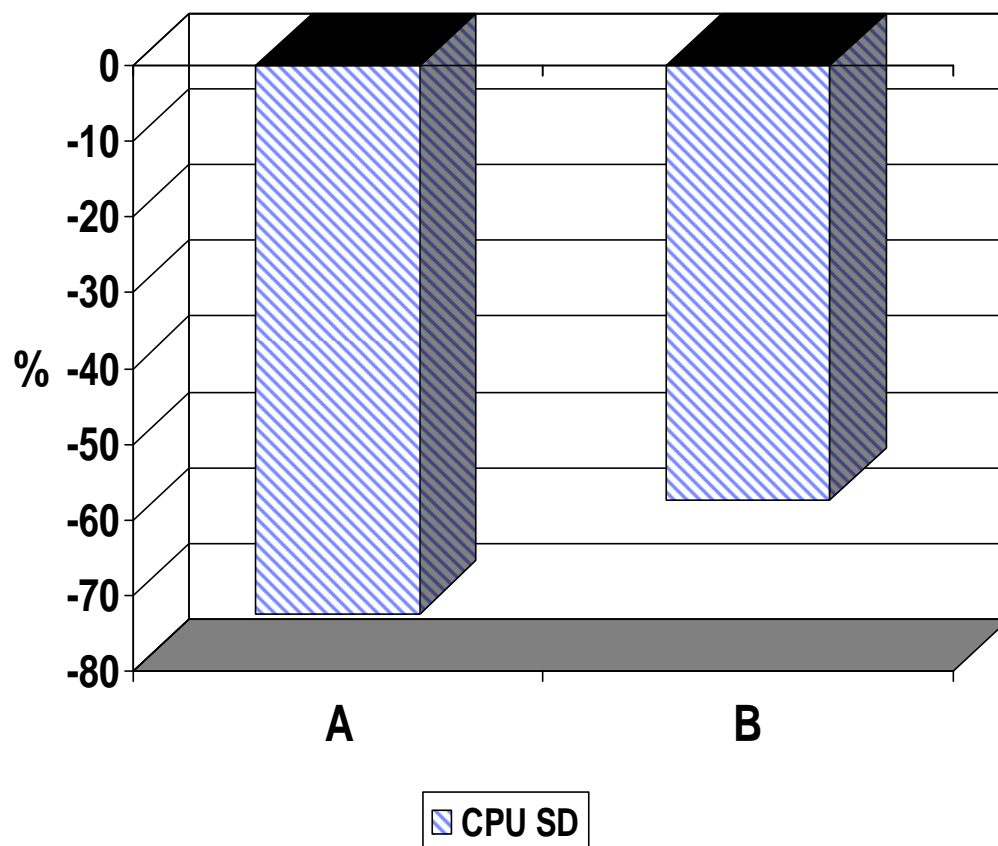
Configuration A – Three z10 LPARs with OSA Express 3 cards and HiperSockets between SD and server LPARs

z/OS 1 — Client
z/OS 2 — SD
HS
z/OS 3 — Server
1 GbE
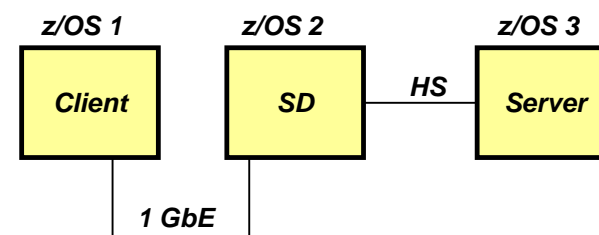
Configuration B – Three z10 LPARs with OSA Express 3 cards

z/OS 1 — Client
z/OS 2 — SD
z/OS 3 — Server
1 GbE
1 GbE

▨ **CPU SD**
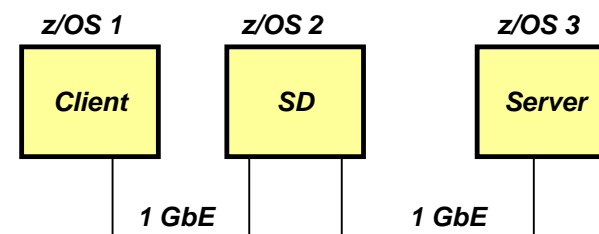
Note: The performance measurements discussed in this presentation are preliminary z/OS V1R11 Communications Server numbers and were collected using a dedicated system environment. The results obtained in other configurations or operating system environments may vary.

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

Site 1 CPC

z/OS   z/OS   z/OS

Site 2 CPC

z/OS   z/OS   z/OS

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

State-full

State-full

State-full

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.

Page 35

Sysplex Networking Technologies and Considerations

Networking Sysplex availability

IBM®

Sysplex autonomics extended with CSM storage constrained monitoring

EZBTCPxx

Dynamic XCF IP network

New in z/OS V1R12

New in z/OS V1R13

TCP/IP

OMPROUTE

VTAM

Network

Sick? Better remove myself from the IP Sysplex!

Feeling better? Maybe it's time to rejoin the IP Sysplex
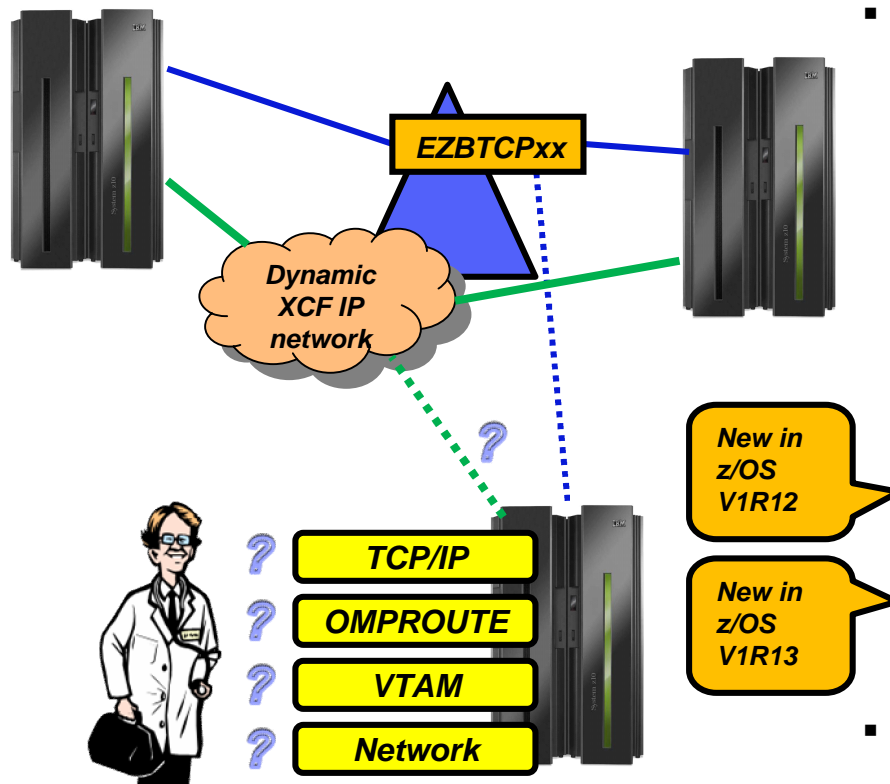
- **Monitoring:**
  - Monitor CS health indicators
    - Storage usage critical condition (>90%) - CSM, TCPIP Private & ECSA
      - For more than TIMERSECS seconds
  - Monitor dependent networking functions
    - OMPROUTE availability
    - VTAM availability
    - XCF links available
  - Monitor for 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)

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

VIPAROUTE TARGET CACHE

XCF @	Best route to
XCF2	*Route to IP@2*
XCF2	*Route to IP@3*

IP Layer routing table

Route lookup to populate VIPAROUTE target cache

XCF1   XCF2   XCF3

SD

IP@1   IP@2   IP@3
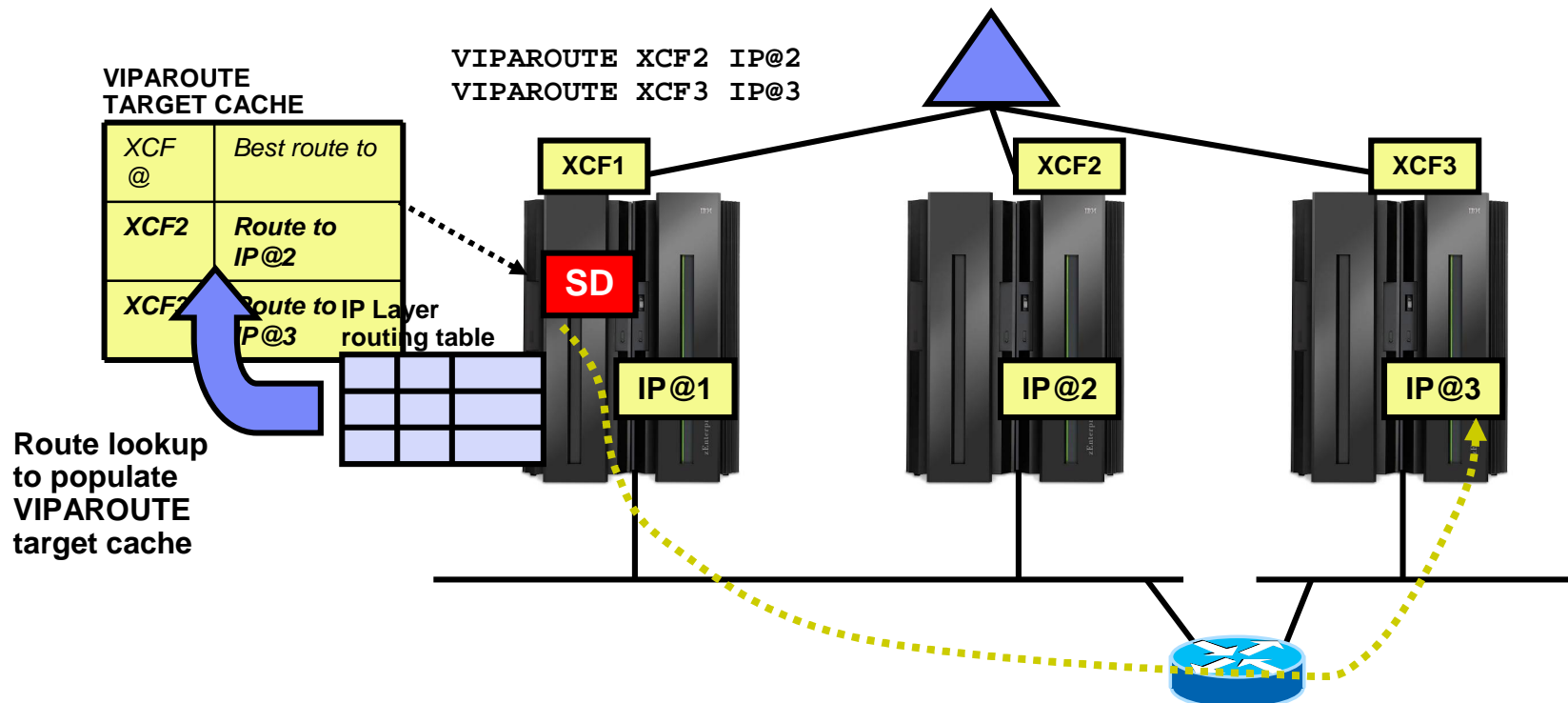
z/OS V1R11 storage shortages and OMPROUTE availability

- 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
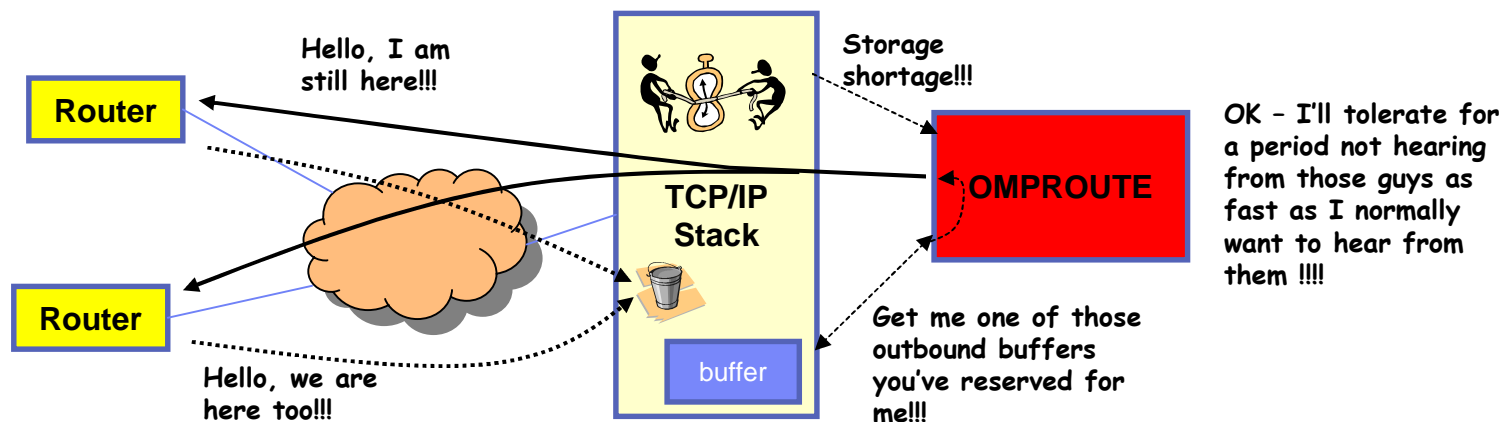
Sysplex Networking Technologies and Considerations

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

IBM ®

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

z/OS TCP/IP Stack

PortA **PortB**

QDIO **QDIO**

10.1.1.1 **10.1.1.2**
MAC: M1 **MAC: M2**

10.1.1.5

Inbound to 10.1.1.1   **Router-1**   Inbound to 10.1.1.2
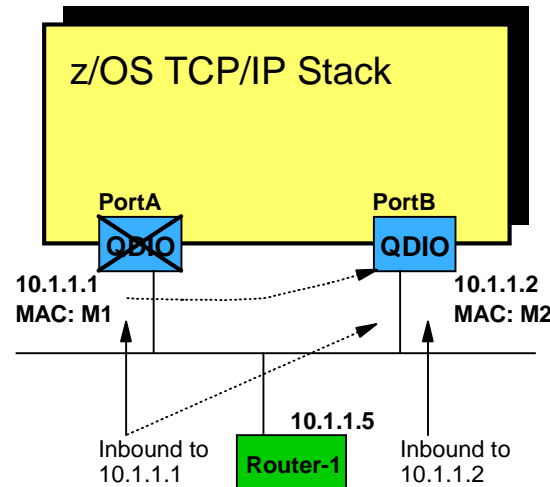
Router's initial ARP Cache

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

Some (restricted) support of dynamic VIPA without dynamic routing

z/OS-1

V1: 10.1.1.10

10.1.1.1
M1

10.1.1.2
M2

z/OS-2

V2: 10.1.2.1

10.1.1.3
M3

10.1.1.4
M4

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

10.1.1.0/24 subnet

Gratuitous ARP with 10.1.1.10

Gratuitous ARP with 10.1.2.1

10.1.1.6/24

Direct Delivery

Router

Indirect Delivery

Accept

Ignore

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)

1

Dest_IP= 10.1.1.10
Src_IP= 172.16.1.1

2

Dest_IP= 10.1.2.1
Src_IP= 172.16.1.1

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.

Sysplex Networking Technologies and Considerations

Network Subplex support

IBM®

Networking sub-plexing within a z/OS Sysplex

Networking Subplexing – Sysplex partitioning from a network perspective

DMZ SNA subplex	**Intranet SNA subplex**
VTAM VTAM	VTAM VTAM VTAM
DMZ IP subplex	**Intranet Primary IP subplex**
IP-1 IP-1	IP-1 IP-1 IP-1
	Research IP subplex
	IP-2 IP-2
	Development IP subplex
	IP-3 IP-3
LPAR1 LPAR2	LPAR3 LPAR4 LPAR5
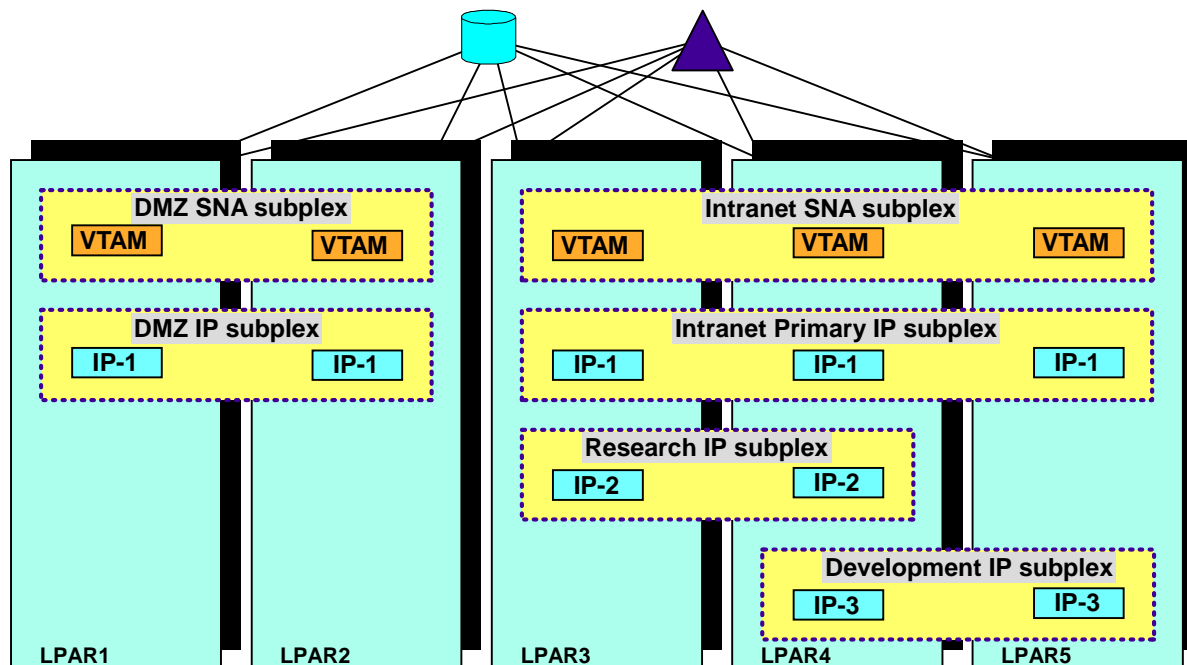
➢**Networking subplex scope:**
- ▸ VTAM Generic Resources (GR) and Multi-Node Persistent Session (MNPS) resources
- ▸ Automatic connectivity - IP connectivity and VTAM connectivity over XCF (including dynamic IUTSAMEH and dynamic HiperSockets based on Dynamic XCF for IP)
  - – HiperSockets VLANID support 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

➢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.

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.

Network security zone A

One z/OS Sysplex

OSA   OSA   OSA   OSA

LPAR1   LPAR2

TCPZONEA   TCPZONEA

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

HTTPA   HTTPA

ZONEA IP Sysplex

ZONEB IP Sysplex

No IP Sysplex awareness or connectivity!

HTTPB   HTTPB

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

TCPZONEB   TCPZONEB

OSA   OSA   OSA   OSA

Network security zone B

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 ….

Please fill out your session evaluation

- Sysplex Networking Technologies and Considerations

- Session # 14682

- 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

# Sysplex Networking Technologies and Considerations

Appendix

IBM®

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.
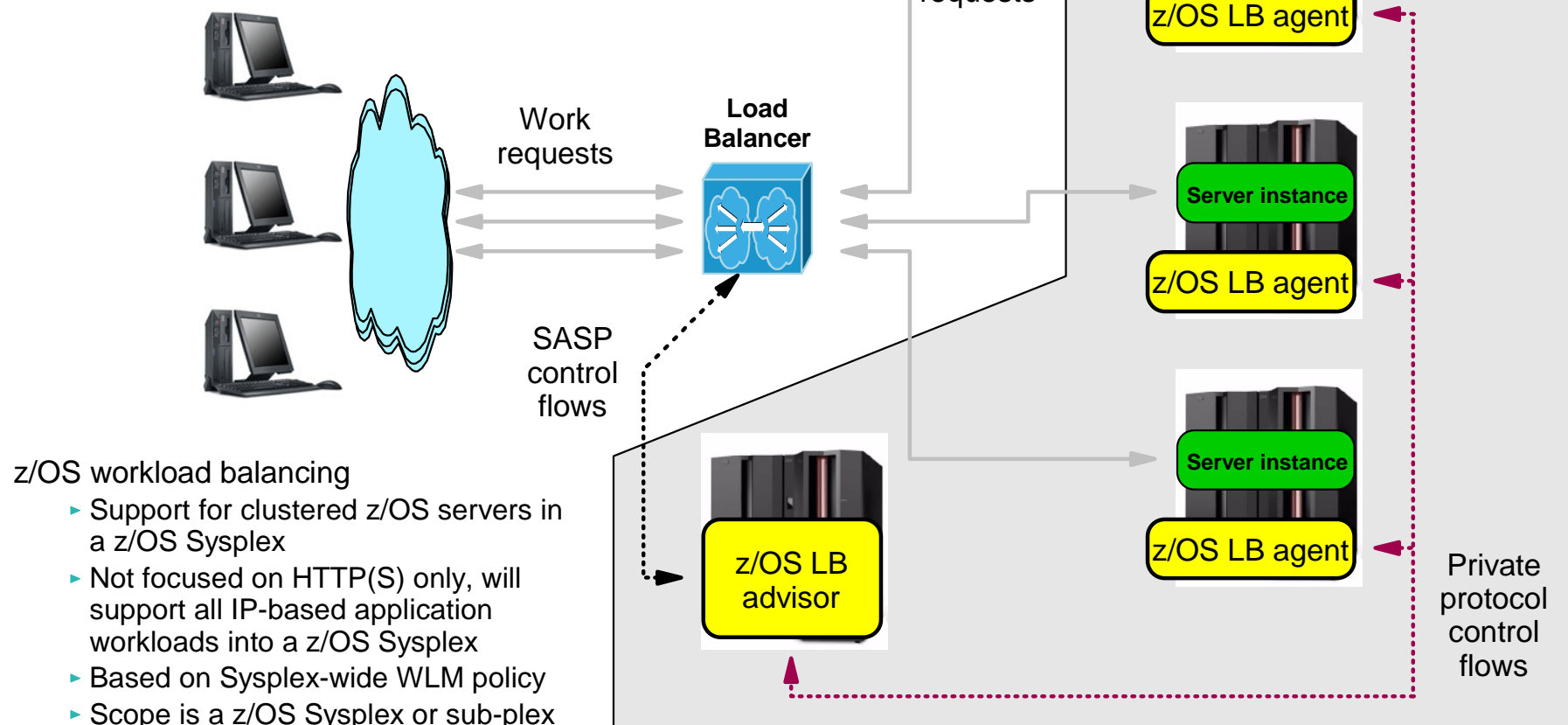
z/OS workload balancing

- ► 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

z/OS Sysplex

Server instance

z/OS LB agent

Work requests

Work requests

Load Balancer

SASP control flows

Server instance

z/OS LB agent

z/OS LB advisor

Server instance

z/OS LB agent

Private protocol control flows

Page 49

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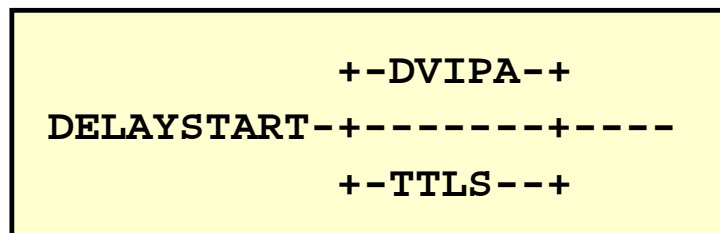
```
*EZD1212E TCPCS DELAYING SYSPLEX PROFILE PROCESSING - NO DYNAMIC ROUTES
 OVER MONITORED INTERFACES WERE FOUND
```

```
 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.

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