SHARE
EDUCATE • NETWORK • INFLUENCE

Sysplex Networking Technologies and Considerations

Session 26069 (Winter SHARE 2020)

Mike Fitzpatrick – mfitz@us.ibm.com
Gus Kassimis – kassimis@us.ibm.com

Agenda

- Sysplex overview
- Workload balancing considerations
- Intra-Sysplex connectivity
- Networking Sysplex availability
- Network availability in a flat network environment
- Network subplexing

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.

SHARE
EDUCATE • NETWORK • INFLUENCE

Sysplex Networking Technologies and Considerations

Sysplex Overview

Goal of Sysplex is to improve availability of business services

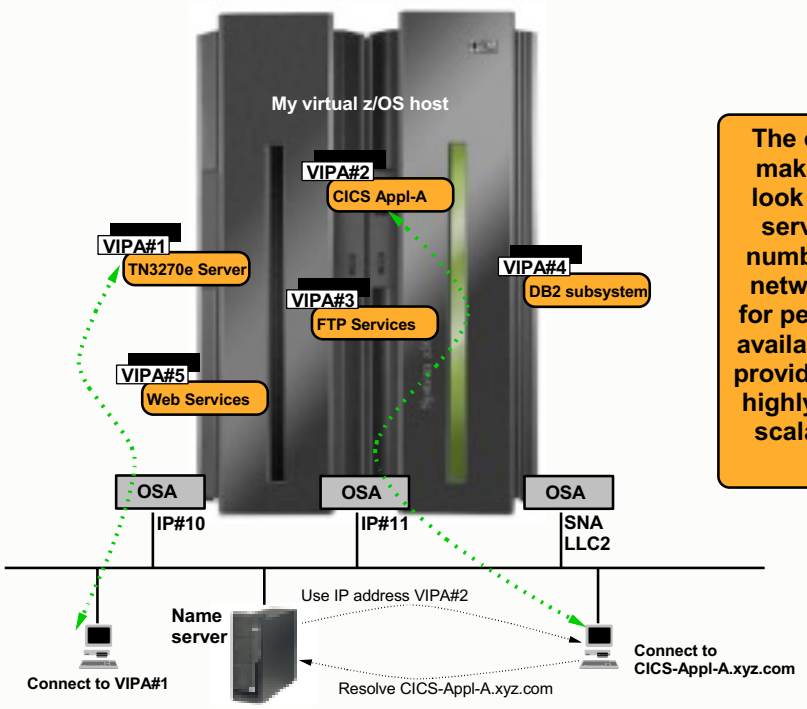
SHARE
EDUCATE • NETWORK • INFLUENCE

- How can we mask failures so that critical business services appear to be highly available?
- How can we scale depending on workload peaks?
- Requires the need for a technology that can provide:
  - Near continuous availability in the face of both planned and unplanned outages
  - High-performance, scalable, read/write access to shared data from multiple systems
  - Rapid, flexible response to changing workload demands
  - Automatic, dynamic workload balancing across systems
  - Incremental, granular growth with near linear scalability

High Availability extends to Communications Server

Hide the complexity of a Sysplex by providing a single system image view

- **The goals of the Parallel Sysplex cluster environment are:**
  - Application location independence
  - Ability to shift application workload between LPARs
  - Application single system image view from the network
  - Application capacity on-demand
  - Component failures do not lead to application failure
- **To gain these benefits:**
  - Carefully design 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

Types of virtual IP addresses on z/OS

- **Static VIPAs**
  - Assigned to a single TCP/IP stack
  - Protects against a network interface failure as VIPA can be reached via a redundant network interface
  - Used by Enterprise Extender
- **Stack-specific dynamic VIPAs**
  - Assigned to a single TCP/IP stack
  - Unlike static VIPAs, can be dynamically moved to another TCP/IP stack within the Sysplex
  - Created using VIPADEFINE statement
  - VIPABACKUP statements to indicate where it can move if primary z/OS system is unavailable
- **Application-specific dynamic VIPAs**
  - A virtual IP address that represents a specific server application instance
  - Dynamically created/deleted when server application's socket binds or closes
  - Can also be created using ioctl() API or by using the MODDVIPA utility)
  - Dynamic VIPA moves with server application if restarted on another TCP/IP stack within the Sysplex
  - Used for defining zCX interface
  - Defined using VIPARANGE statement (on primary system and all other systems the application can move to)
- **Sysplex-specific dynamic VIPAs**
  - Known as distributed dynamic VIPAs
  - Intra-Sysplex load balancing performed across the multiple server application instances that are using the dynamic VIPA
  - Used by Sysplex Distributor
  - Created using the VIPADISTRIBUTEIBUTE statement

SHARE
EDUCATE • NETWORK • INFLUENCE

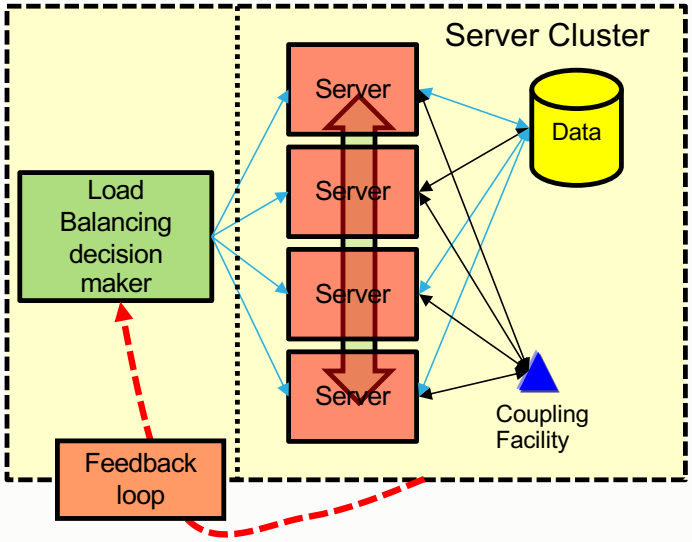
Sysplex Networking Technologies and Considerations

Workload Balancing Considerations

What are the main objectives of network workload balancing?

- **Performance**
  - Workload management across a cluster of server application instances
  - One server application instance on a single hardware node may not be sufficient to handle all the workload requests
- **Availability**
  - As long as one server application instance is up-and-running, the "service" is available
  - Individual server application instances and associated hardware components may fail without impacting overall "service" availability
- **Capacity management & horizontal growth**
  - Transparently add/remove server application instances and/or hardware nodes to/from the pool of server applications 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

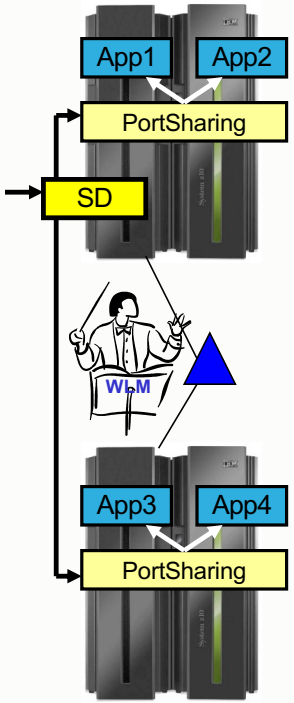
All server application instances must be able to provide the same basic service. In a z/OS Sysplex that means the applications must be Sysplex-enabled and be able to share data across all LPARs in the Sysplex.



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

z/OS IP network workload balancing overview

- **Two main technologies:**
  - Sysplex Distributor
  - Port sharing
- **Sysplex Distributor**
  - Sysplex Distributor performs layer-4 load balancing
    - Makes routing decision when it sees an inbound TCP SYN segment destined for one of the distributed dynamic VIPA IP address & port combinations it load balances
  - Sysplex Distributor uses MAC-level forwarding when connection routing takes place over XCF or HiperSockets links
  - Sysplex Distributor uses GRE headers when connection routing takes place over any other 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 TCP/IP stack advertises routing ownership for the distributed dynamic VIPA
  - All outbound packets from the server application instances can take whatever route is most optimal from the server application instance node back to the client
- **Port sharing**
  - PORTSHARING can be used within a TCP/IP stack to distribute connections among multiple server application address spaces within that TCP/IP stack
    - SHAREPORT – TCP/IP Server Efficiency Factor (SEF) value used to perform a weighted round robin distribution to server application instances for new connections
    - SHAREPORTWLM – WLM input is used to select a server application instance for a new connection



Sysplex Distributor distribution method overview

- **z/OS target applications 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 application instances' percentage of active connections match specified weights

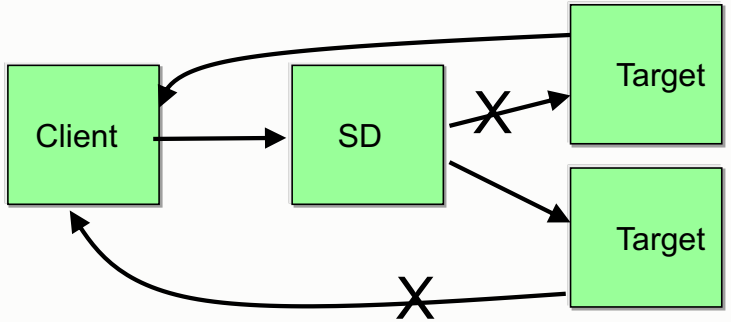
- **z/OS target applications using WLM recommendations**
  - BASEWLM
    - Based on LPAR level CPU capacity/availability and workload importance levels
  - SERVERWLM
    - Similar to BASEWLM but takes into account WLM service class and how well individual server applications are performing (i.e. meeting 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***

Sysplex Distributor distribution method overview …

- **Newer distribution methods:**
  - HOTSTANDBY
    - Incoming connections are distributed to a primary server application instance and only rerouted to a backup server application instance (the “hot standby”) when the primary server application instance is not ready, unreachable, or unhealthy

Sysplex Distributor built-in awareness of abnormal conditions

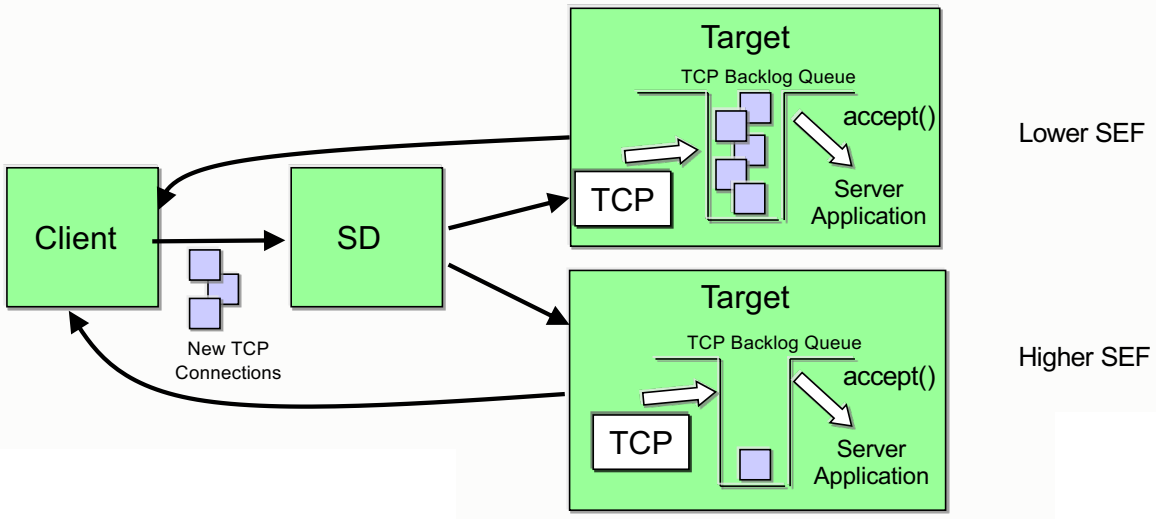
- **TSR – Target Server Responsiveness**
  - How healthy is the target system and application from an SD perspective? A percentage, 0-100%
  - Comprised of several individual health metrics:
    - **TCSR** – Target Connectivity Success Rate (percentage: 100 is good, 0 is bad)
      - Are connections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



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

Sysplex Distributor built-in awareness of abnormal conditions…

- **TSR – Target Server Responsiveness (continued)**
  - **SEF** – Server Efficiency Fraction (percentage: 100 is good, 0 is bad)
    - Is the target server application keeping up with new connections in its backlog queue?
    - Is the new connection arrival rate higher than the application accept rate? (Is backlog growing over time?)
    - How many connections in the TCP backlog queue? How close to maximum backlog queue depth? Were any new connections dropped because the backlog queue depth was exceeded?
    - Is the server application hung? (Is server application not accepting any connections?)
    - Is the number of half-open connections on the backlog queue growing? (Similar to CER)



Middleware/Application issues (“Storm Drain Problem”)

- TCP/IP and WLM are not aware of all problems experienced by load balancing targets (middleware/applications)
  - 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 not completing successfully
- This is sometimes referred to as the *Storm Drain Problem*
  - The server application is favored by WLM since it is using very little CPU capacity
  - As workloads increase, the server application is favored more and more over other server applications
  - All this work goes "down the drain"

Improving WLM awareness of Application Health

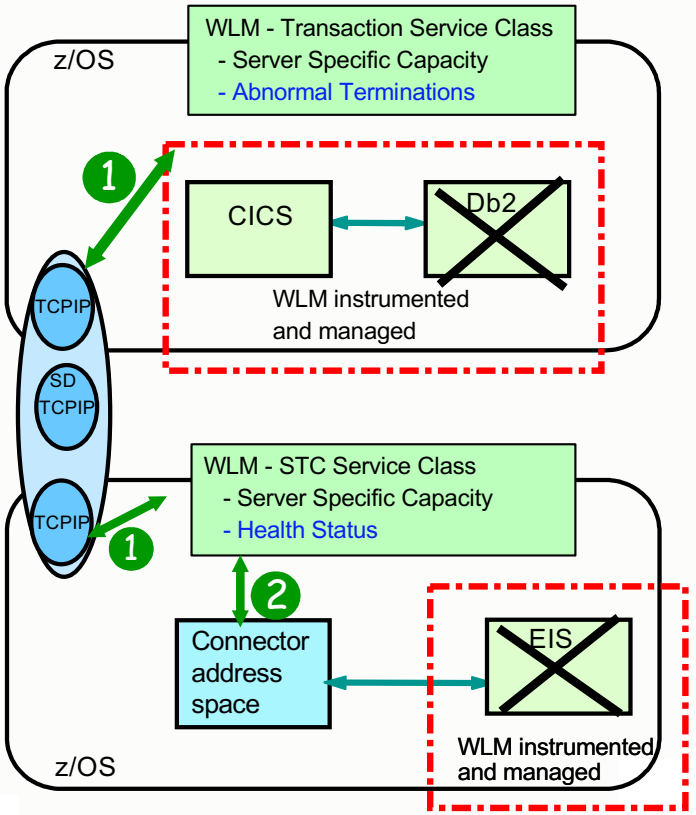
Server Scenarios

1 IWM4SRSC WLM Service

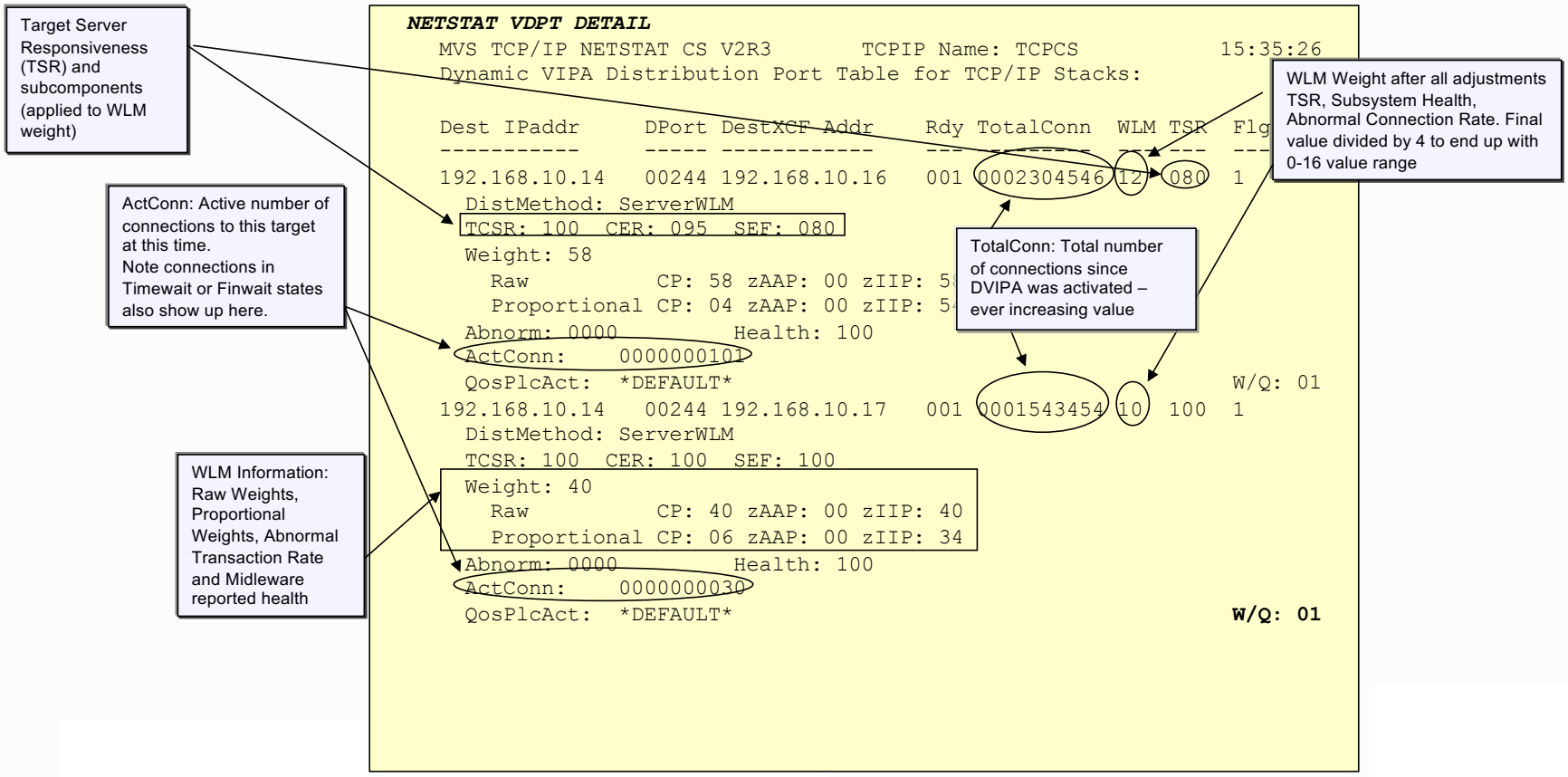
- Used by Sysplex Distributor to obtain WLM recommendations
- Abnormal Termination information: Reported by 1st-tier server application when transactions can not complete because back-end resources are not available
  - WLM uses this information to reduce the recommendation for unhealthy server applications

2 IWM4HLTH WLM Service

- Allows address spaces which are not instrumented with WLM to set a health status which is also returned by IWM4SRSC
- The ServerWLM recommendations are reduced when the health is <100%
- Exploited by CICS, CICS Transaction Gateway, Db2, LDAP, IMS, and WebSphere Application Server



Using Netstat VDPT Display to monitor Sysplex Distributor



What impacts the final selection of a target server instance?

Technology	Target LPAR displaceable capacity as seen by WLM	Application performance as seen by WLM	Application self-perceived health (reported to WLM)	Application TCP/IP perceived health (TSR value)	QoS perceived network performance (QoS fraction)
SD ROUNDROBIN	No	No	No	Yes (only if TSR=0)	No
SD WEIGHTEDACTIVE	No	No	Yes	Yes	No
SD BASEWLM	Yes	No	No	Yes	Yes
SD SERVERWLM	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
SD HOTSTANDBY	No	No	Yes	Yes	No
PORT SHAREPORT	No	No	No	Yes (only SEF)	No
PORT SHAREPORTWLM	No	Yes	Yes	Yes (only SEF)	No

SERVERWLM method: What is displaceable LPAR capacity?

- LPAR capacity that is currently being used for less important work than what Sysplex Distributor wants to send to the LPAR

- An example:
  - New work will run at Importance level 2
  - Which LPAR is best (LPAR 1 or LPAR 2)?
    - They both have 500 service units of displaceable work
    - Typically, both would be considered equally good targets
  - Sysplex Distributor can take importance level of displaceable workload into consideration
    - LPAR2 will be preferred since the importance level of the work being displaced is lower than the work that would be displaced on LPAR1

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

IL 0: High importance
IL 7: Low importance

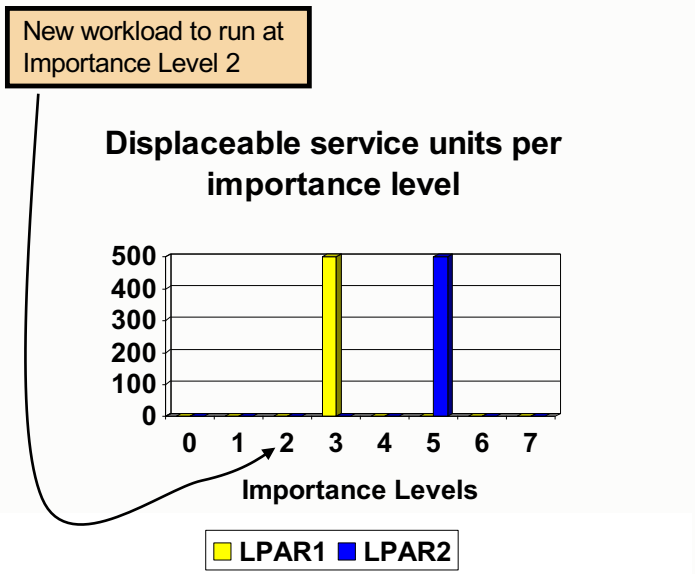
LPAR1	
IL	SUs
0	0
1	0
2	0
3	500
4	0
5	0
6	0
7	0

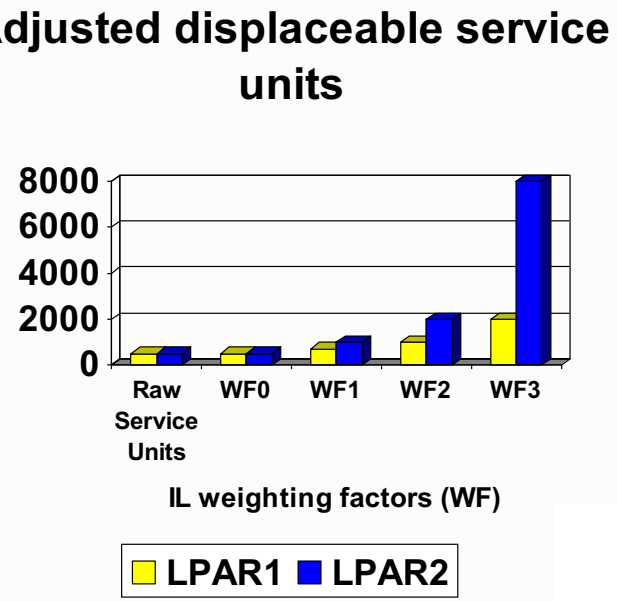
LPAR2	
IL	SUs
0	0
1	0
2	0
3	0
4	0
5	500
6	0
7	0

SERVERWLM method:

How much should importance levels influence the workload distribution?

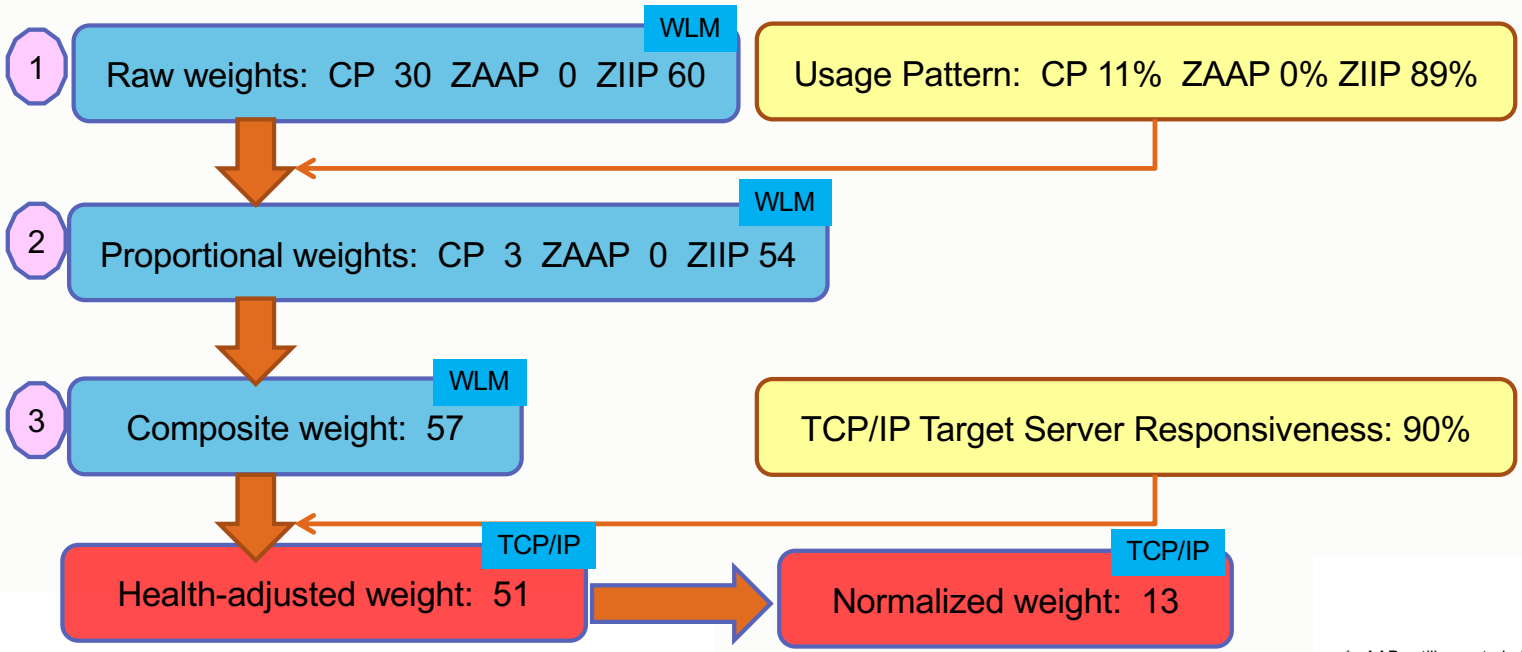
- Importance level weighting factor of zero (IL0) means no consideration for importance level of existing work
- Importance level weighting factors of one through three (IL1 through IL3) will gradually shift new workloads towards LPARs with the lowest importance level work to displace
  - In this example, LPAR2





SERVERWLM method: specialty processors - overview

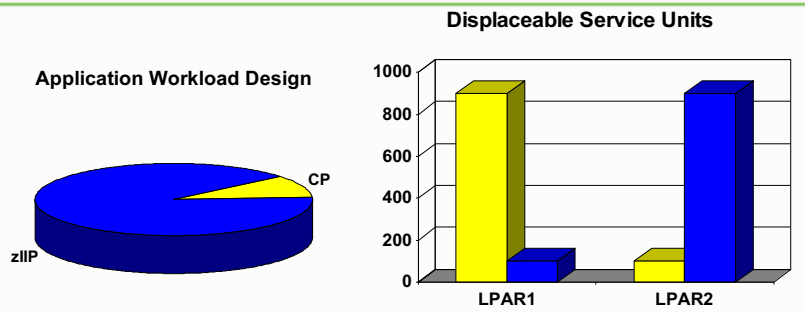
- **When using WLM server-specific weights**.
  - WLM returns
    - The raw CP, zAAP*, and zIIP system weights
    - Proportional weights – raw weight modified by actual server usage pattern as observed by WLM
    - Composite weight

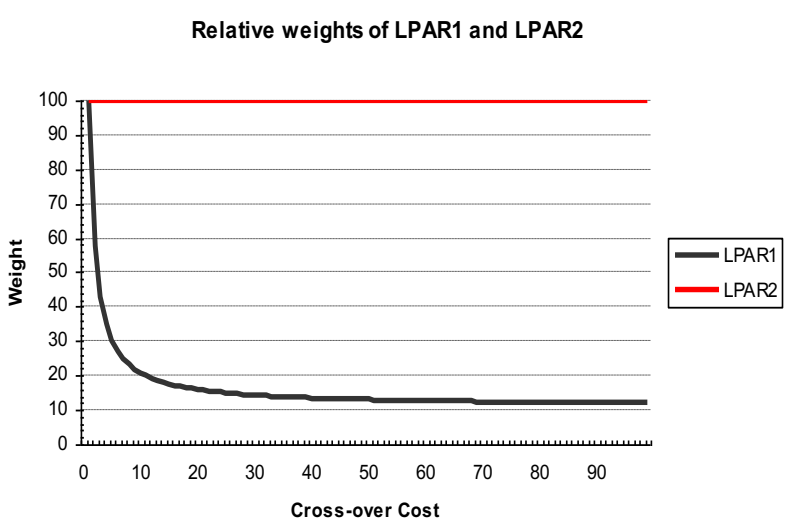


* zAAPs still reported although no longer available

SERVERWLM method: zIIP cross-over to CP

- Application designed to use 10% CP and 90% zIIP
- LPAR1 and LPAR2 are targets
  - LPAR1 has 900 CP SUs and 100 zIIP SUs that can be displaced
  - LPAR2 has 100 CP SUs and 900 zIIP SUs that can be displaced
- Without a cross-over cost, the two targets are equally good to receive new workload
- 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: no penalty for cross-over
  - 100: maximum penalty for cross-over





SERVERWLM method: Configuring and displaying the options

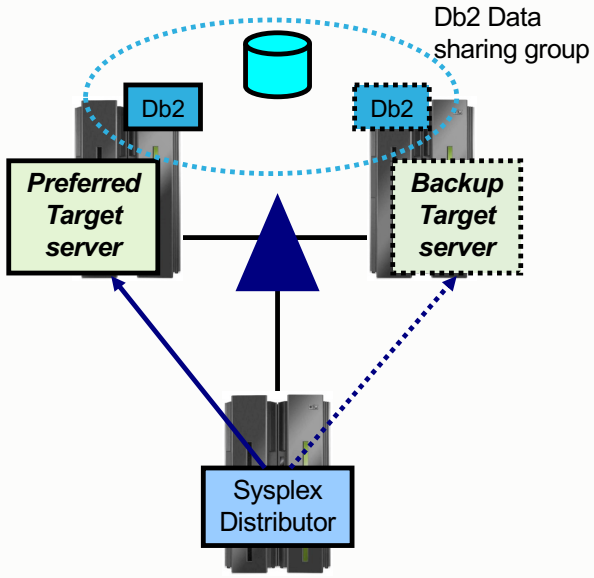
- The new configuration parameters are only valid when server-specific recommendations are being used
- These parameters can affect performance
  - Importance Level values range from 0 (no impact) to 3 (aggressive weighting).
    - Guideline – use Moderate (ILWEIGHTING 1) value initially.
  - Crossover cost values range from 1 (no impact) to 100 (crossover cost very expensive).
    - Guideline – Use a low PROXCOST initially.

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

```
NETSTAT VIPADCFG DETAIL
VIPA Distribute:
  Dest:        192.168.10.11..8000
    DestXCF:   ALL
      SysPt:   No  TimAff: No  Flg: ServerWLM
      OptLoc:  No
      ProcXCost:
        zAAP: 001  zIIP: 020
      ILWeighting: 1
```

Sysplex Distributor: Hot Standby support

- Data sharing provides for highly scalable and highly available configuration
  - Additional server application instances and LPARs can be cloned and added to increase capacity and improve performance
  - But what if the work can comfortably fit within a single LPAR?
    - Data sharing becomes primarily an availability feature
    - A Hot Standby configuration would allow all work 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 customers had requested a simpler mechanism for doing this via the TCP/IP profile



Sysplex Distributor: Hot Standby support…

- Configure a single target server application to receive all new connection requests
  - Other target server applications are active but not receiving any new connection requests
  - Automatically route traffic to a backup target server application when the active target server application becomes unavailable
- Enable using a new HOTSTANDBY distribution method
  - One preferred target
    - AUTOSWITCHBACK option - switch to the preferred target whenever it becomes available (not used if original switch was for health problems)
  - And one or more backup targets ranked in order of preference
  - A target is not available when:
    - Server application not ready -OR-
    - Route to target TCP/IP stack 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 V2R3       TCPIP Name: TCPCS          13:35:14
Dynamic VIPA Information:

  ......
  VIPA Distribute:
    Dest:      192.168.1.1..8020
      DestXCF: 192.168.0.1
      DistMethod: HotStandby     SrvType: Preferred
      SysPt:   No   TimAff: No    Flg:
    Dest:      192.168.1.1..8020
      DestXCF: 192.168.0.2
      DistMethod: HotStandby     SrvType: Backup  Rank: 200
      SysPt:   No   TimAff: No    Flg:
    Dest:      192.168.1.1..8020
      DestXCF: 192.168.0.3
      DistMethod: HotStandby     SrvType: Backup  Rank: 100
      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 V2R3       TCPIP Name: TCPCS           14:18:18
Dynamic VIPA Destination Port Table for TCP/IP stacks:
Dest:         192.168.1.1..8020
  DestXCF:    192.168.0.1
  TotalConn: 0000000000  Rdy: 000  WLM: 10  TSR: 100
  DistMethod: HotStandby           SrvType: Preferred
  Flg: Backup
Dest:         192.168.1.1..8020
  DestXCF:    192.168.0.2
  TotalConn: 0000000000  Rdy: 001  WLM: 10  TSR: 100
  DistMethod: HotStandby           SrvType: Backup
  Flg: Backup
Dest:         192.168.1.1..8020
  DestXCF:    192.168.0.3
  TotalConn: 0000000000  Rdy: 001  WLM: 10  TSR: 100
  DistMethod: HotStandby           SrvType: Backup
  Flg: Active
```

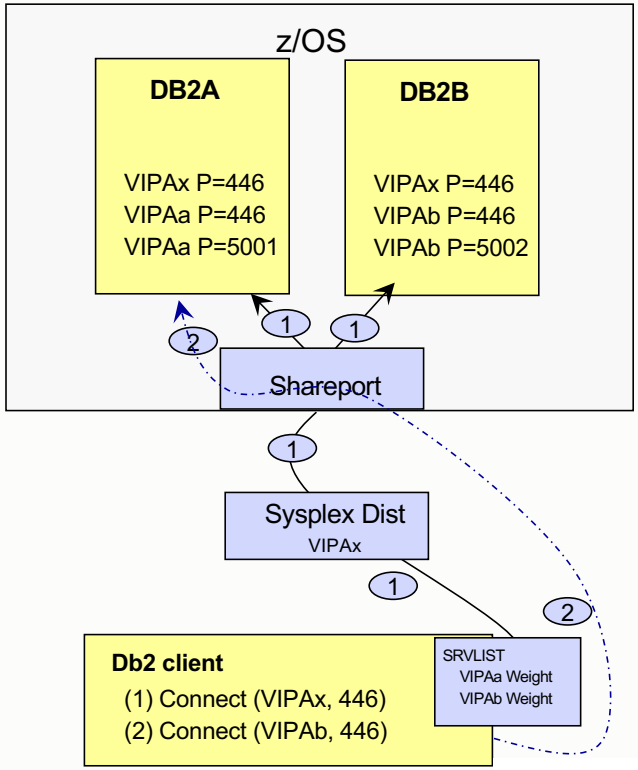
Sysplex Distributor Polling Intervals

- By default Sysplex Distributor queries WLM every 60 seconds
  - During this same interval, Sysplex Distributor also calculates 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
      - ***Should be configured on all*** TCP/IP stacks in the Sysplex environment (distributor and targets)
        - Each TCP/IP stack implements its own timer

Dynamic VIPA affinity support – Problem Scenario

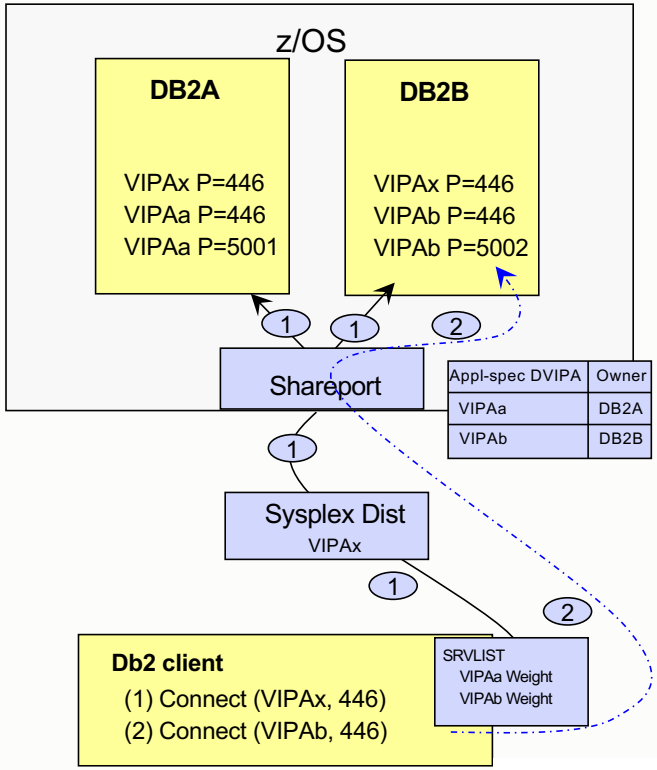
- Multiple Db2 members deployed in the same z/OS image
  - Currently 2 methods for creating Dynamic VIPAs to represent each member (member-specific DVIPAs)

1. BIND-specific approach: Using the TCP/IP profile PORT BIND specific support, each Db2 member binds its listening socket to a specific member DVIPA
   - ***All is well,*** incoming connections get routed to the appropriate Db2 member
2. INADDR_ANY approach: Dynamic VIPAs configured via Db2 BSDS
   - Db2 programmatically creates the DVIPAs during initialization and ***binds listening sockets to* INADDR_ANY**
     - Allows each Db2 member to be reached using any IP address (including the IPv4 and IPv6 DVIPAs, Db2 Location Alias IP addresses)
     - However if multiple Db2 members in the same z/OS system use the BSDS method, ***incorrect routing of incoming connections is possible,*** resulting in suboptimal load balancing



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 differently 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 (using existing SHAREPORT method)
      - 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**
  - Socket APIs (SIOCSVIPA and SIOCSVIPA6 IOCTLs)
  - MODDVIPA utility program
- **Allows sysplexWLB connection using a BSDS DVIPA to be routed to the intended Db2 member when multiple Db2 members coexist on the same TCP/IP stack affinity**
  - Requires new Db2 exploitation support (APAR PI08208 available on Db2 V10 and Db2 V11)



SHARE
EDUCATE • NETWORK • INFLUENCE

Sysplex Networking Technologies and Considerations

Intra-Sysplex Connectivity

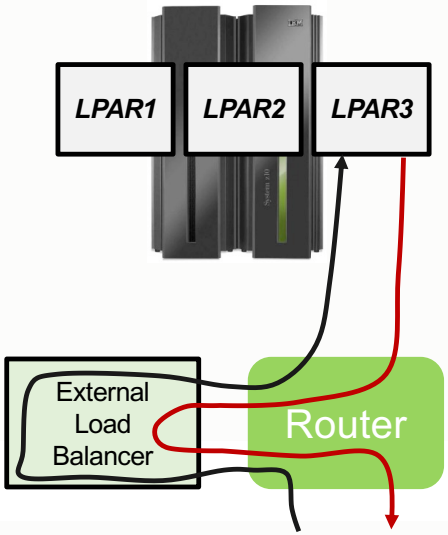
Load Balancing - Inbound and outbound routing paths

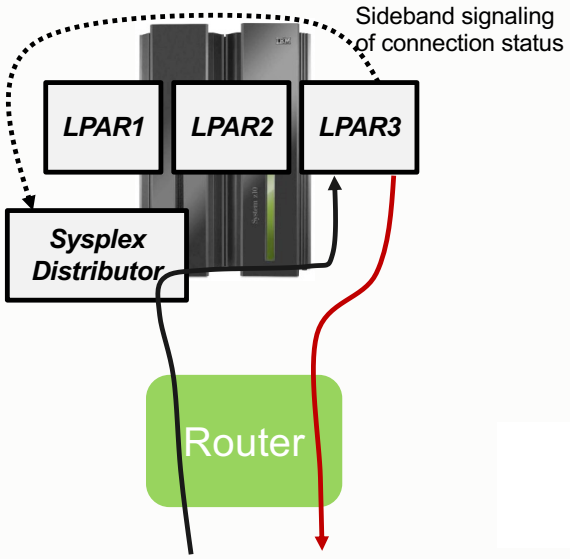
Most external load balancers

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

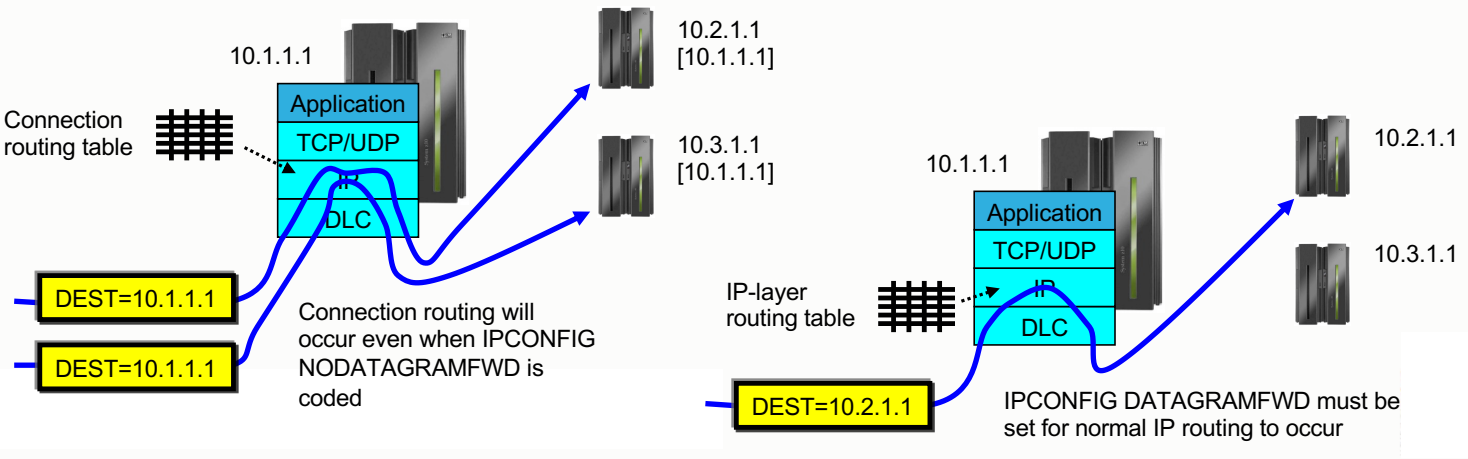
Sysplex Distributor does not use NAT; it uses MAC-level forwarding for the inbound flows, which requires the target server applications be on a directly connected network (XCF network), or the use of GRE (VIPAROUTE). Sysplex Distributor does not need to be in the outbound path, so no control of outbound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





Two types of intra-Sysplex 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 and 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

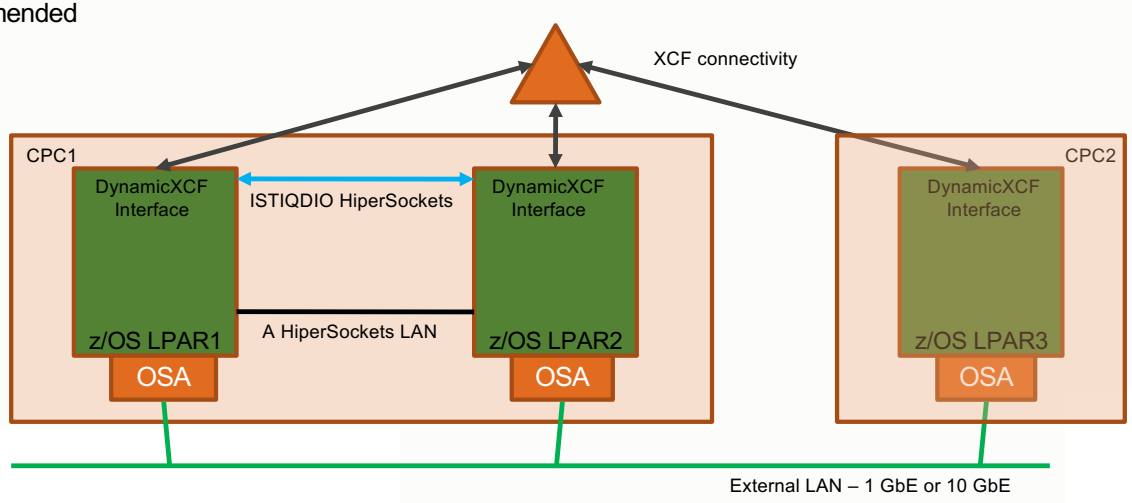


Role of XCF, HiperSockets, and external interfaces in Sysplex

- **XCF**
  - All XCF control messaging between stacks in a Sysplex always go via XCF messages
  - DynamicXCF Sysplex Distributor connection routing (but only if no VIPAROUTE defined)
  - If configured for static or dynamic routing, normal IP routing between LPARs – not recommended
- **HiperSockets**
  - If ISTIQDIO is defined (in VTAM), DynamicXCF Sysplex Distributor connection routing between LPARs on same CPC uses this interface instead of XCF
  - Considered part of the DynamicXCF network interface – no separate DEVICE/LINK or INTERFACE definitions required
- **External interface or a manually defined HiperSockets**
  - If VIPAROUTE defined, then used for Sysplex Distributor 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 TCP/IP stacks.

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

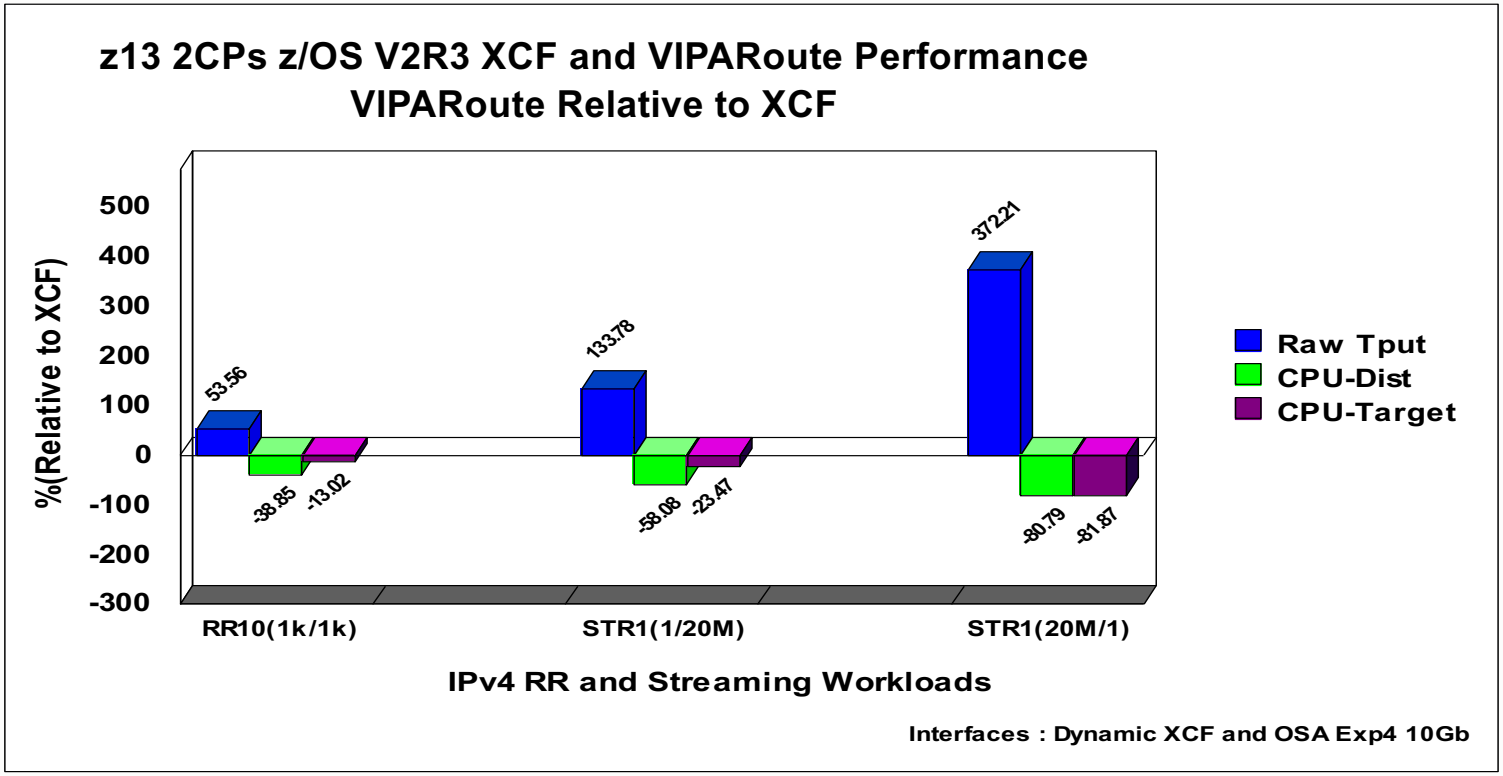


So what should be used for what type of routing?

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

	Exchange control messages between stacks in a Sysplex	Sysplex Distributor connection routing (forwarding inbound packets for distributed connections)	General IP routing between stacks in a Sysplex
XCF messaging	Always	Yes - If no VIPAROUTE specified	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 TCP/IP stacks on same CPC only.	Can be used (not recommended since XCF will be used to reach TCP/IP stacks on other CPCs)
All other connectivity options between stacks in a Sysplex (OSA, HiperSockets, MPC, etc.)	Never	Yes - If VIPAROUTE is defined	Always

VIPAROUTE vs XCF – Performance Comparison

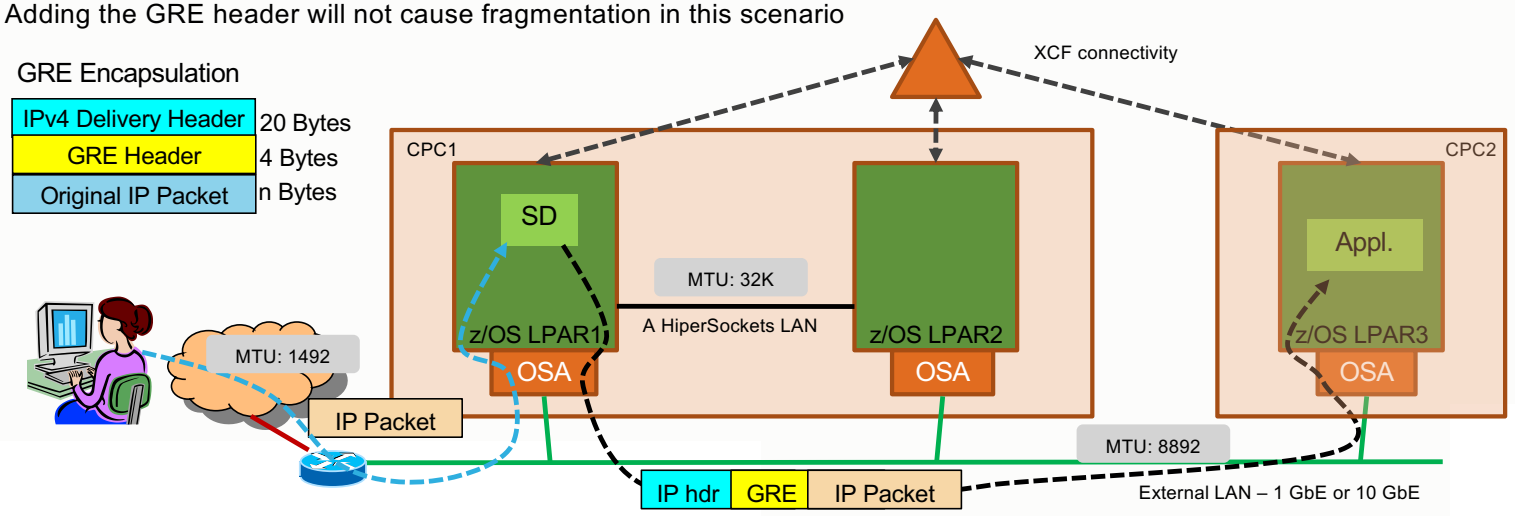


Significant throughput benefits for both request/response (RR) and streaming data (STR) patterns
Significant reduction in networking CPU overhead on both Sysplex Distributor and Target LPARs

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.

VIPAROUTE and MTU size considerations

- When VIPAROUTE is used, the distributing TCP/IP stack adds a GRE header to the original IP packet before forwarding to the target TCP/IP 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



VIPAROUTE fragmentation avoidance

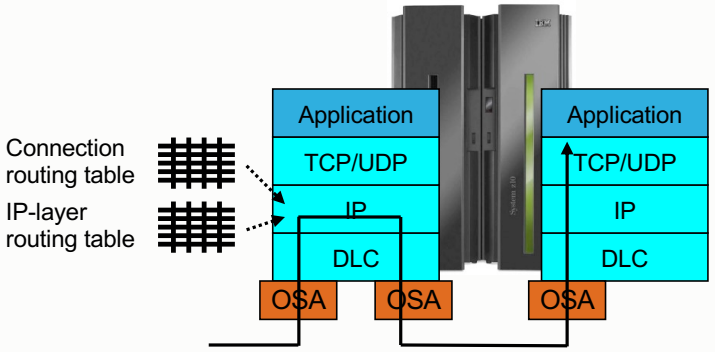
- VIPAROUTE is used by many customers to offload Sysplex Distributor forwarded traffic from XCF links
  - When used in combination with QDIO Accelerator for Sysplex Distributor, can result in dramatically reduced overhead for Sysplex Distributor forwarding
- Fragmentation is still a concern for some 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 firewalls)
  - Jumbo frames can cause fragmentation for outbound traffic routed through standard network (1492 MTU)
  - Fragmentation can cause significant performance degradation

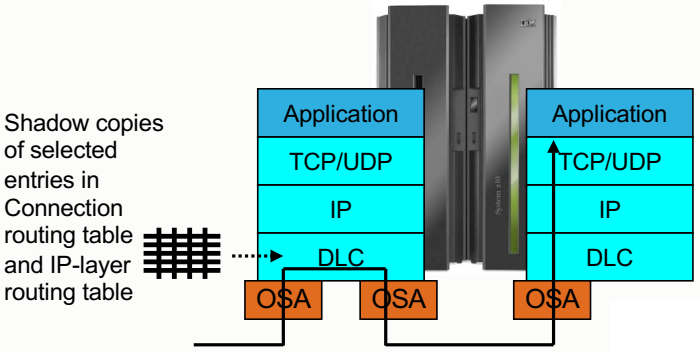
VIPAROUTE fragmentation avoidance…

- 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 allows client TCP/IP stacks to build packets that account 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 TCP/IP stack if it is also a target and VIPAROUTE is defined
    - Option *ALL* – Enables adjusted MSS for all connections using a dynamic VIPA (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 TCP/IP stack specific option - if the default is not taken, then it must be configured on all TCP/IP stacks in the sysplex

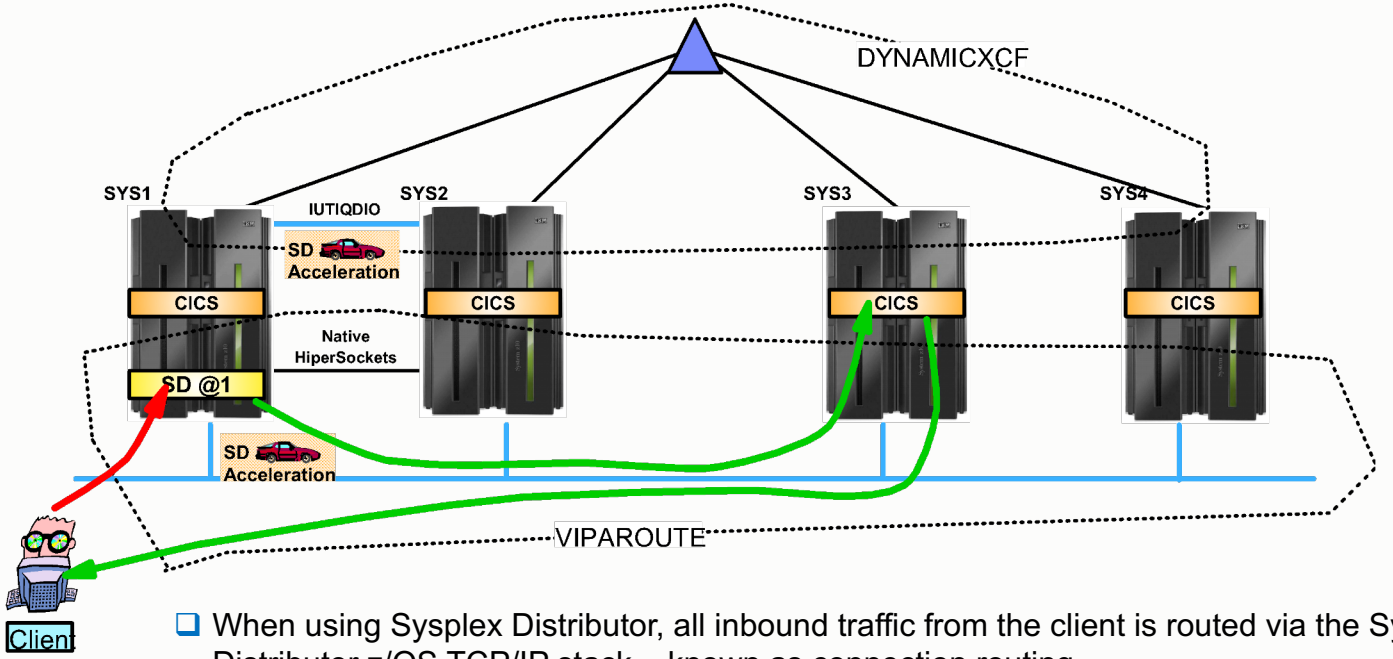
QDIO routing accelerator

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





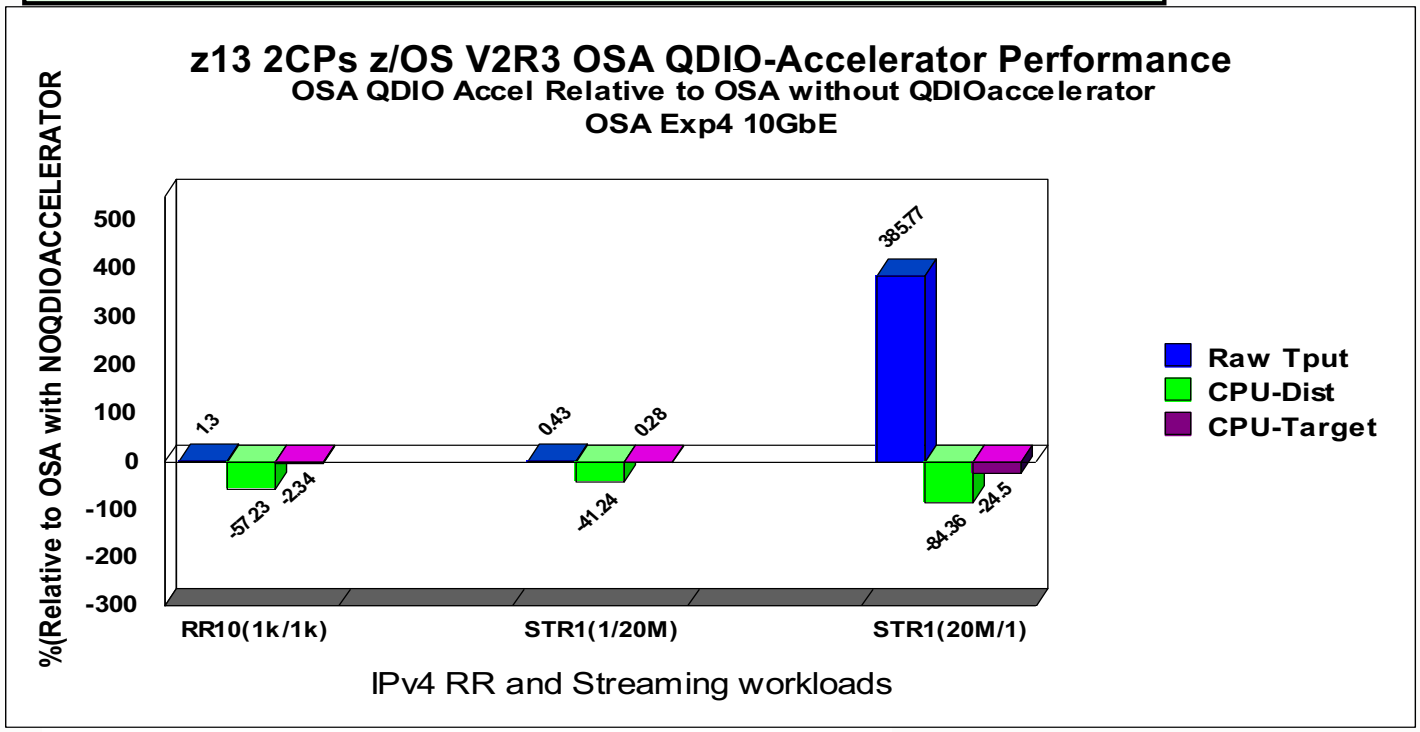
Sysplex Distributor connection routing benefits from QDIO Accelerator



- When using Sysplex Distributor, all inbound traffic from the client is routed via the Sysplex Distributor z/OS TCP/IP stack – known as connection routing
  - Outbound traffic goes directly back to the client
- When inbound packets to Sysplex Distributor is over QDIO (OSA) or iQDIO (HiperSockets), Sysplex Distributor will perform accelerated connection routing when outbound route uses 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

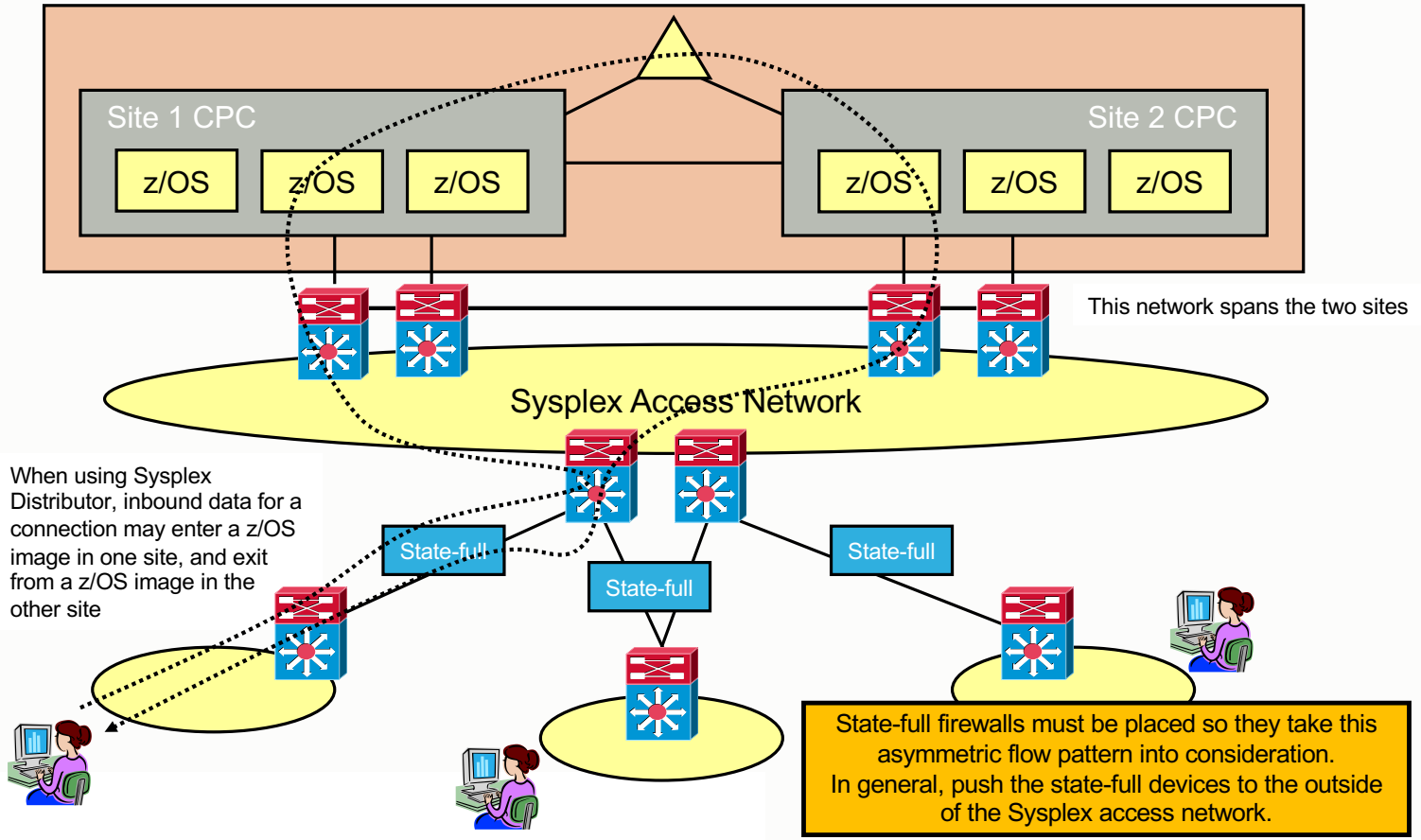
Sysplex Distributor Accelerator Performance

- ✓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 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.

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

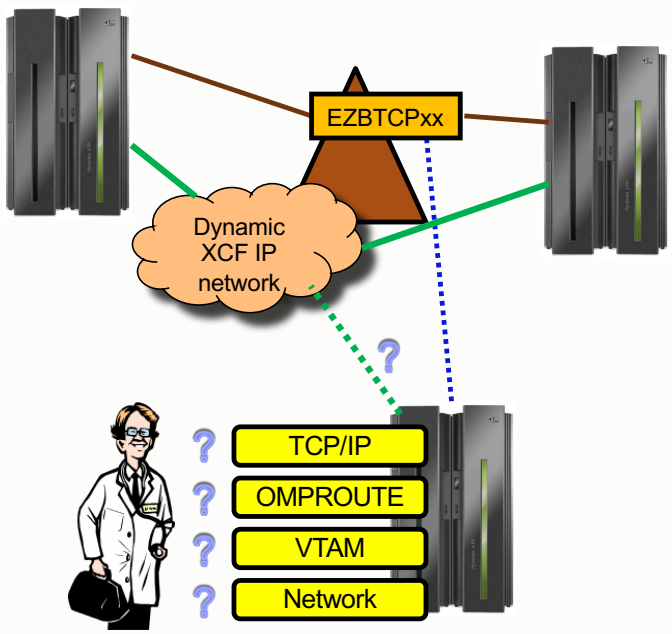


SHARE
EDUCATE • NETWORK • INFLUENCE

Sysplex Networking Technologies and Considerations

Network Sysplex Availability

Sysplex autonomics

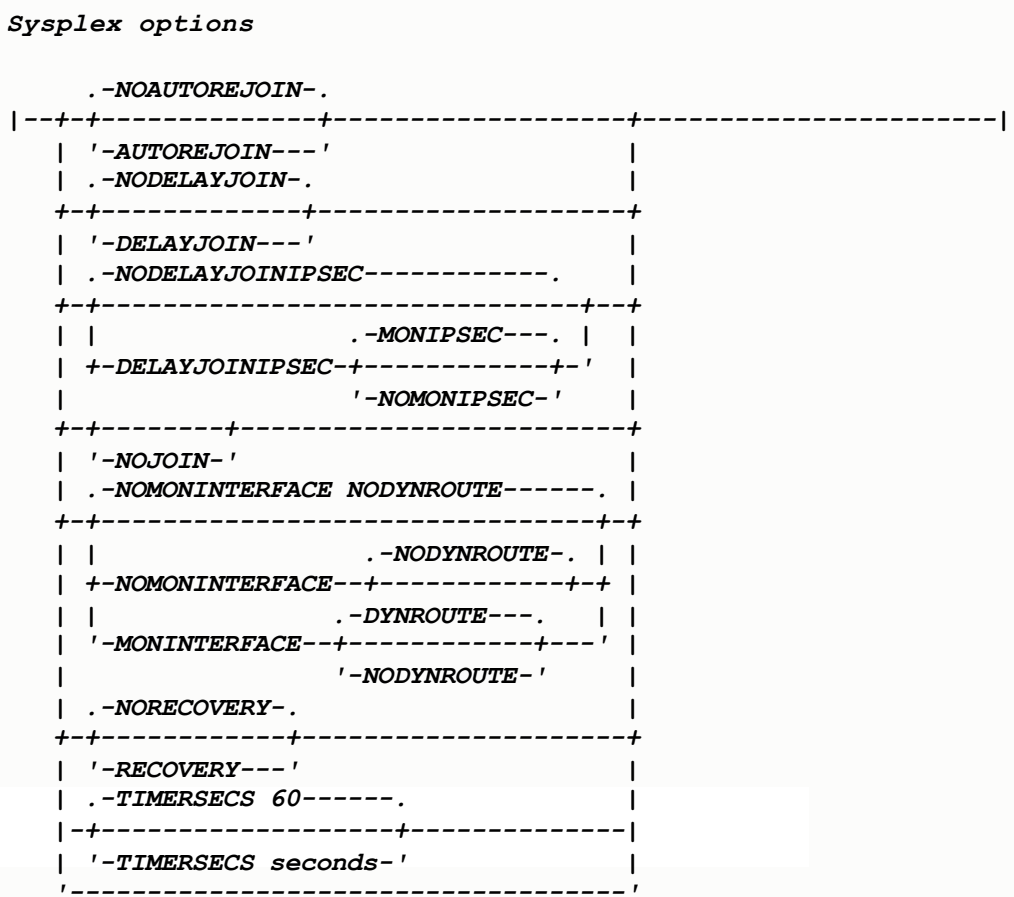


Sick? Better remove myself from the IP Sysplex!

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

- **Monitoring:**
  - Monitor Communications Server health indicators
    - Storage usage critical condition (>90%) - CSM, TCP/IP Private, and ECSA
  - Monitor dependent networking functions
    - OMPROUTE availability
    - VTAM availability
    - XCF links available
  - Monitor for abends in Sysplex-related TCP/IP stack components
    - Selected internal components that are vital to Sysplex processing
  - Monitor for repetitive internal abends in non-Sysplex related TCP/IP stack components
    - 5 times in less than 1 minute
  - Selected network interface availability and routing
  - Detect when CSM FIXED, CSM ECSA, or CSM HVCOMMON has been constrained (>80% utilization) for multiple monitoring intervals
  - IPSec infrastructure active and operational
- **Actions:**
  - Remove the stack from the TCP/IP Sysplex (manual or automatic)
    - Retain the current Sysplex configuration data in an inactive state when a TCP/IP stack leaves the Sysplex
  - Reactivate the currently inactive Sysplex configuration when a TCP/IP stack rejoins the Sysplex (manual or automatic)

GLOBALCONFIG SYSPLEXMONITOR Options

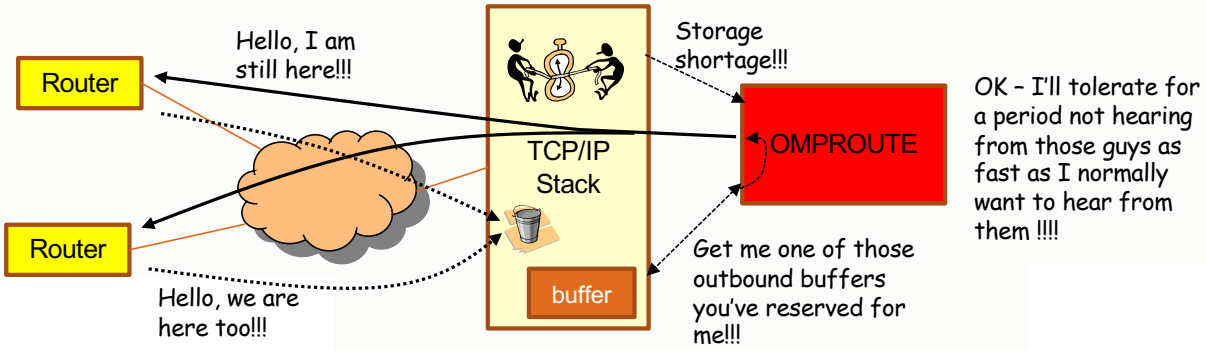


New in V2R4!

DELAYJOINIPSEC

OMPROUTE availability during storage shortages

- OMPROUTE and the TCP/IP stack work together to make OMPROUTE more tolerant of storage shortage conditions:
  - TCP/IP stack informs OMPROUTE of 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



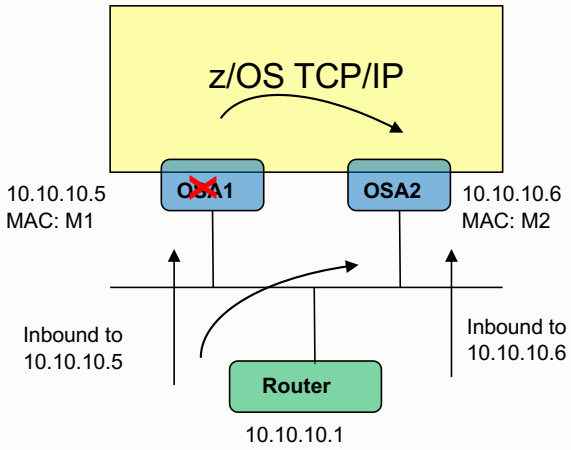
SHARE
EDUCATE • NETWORK • INFLUENCE

Sysplex Networking Technologies and Considerations

Network Availability in a flat network environment

Interface resilience without dynamic routing

- At least two adapters attached to same network
- Adapters should be different physical interfaces for real availability



Router's initial ARP cache

IP Address	MAC Address
10.10.10.5	M1
10.10.10.6	M2

Router's ARP cache after move

IP Address	MAC Address
10.10.10.5	M2
10.10.10.6	M2

Example: OSA1 fails or is stopped

1. z/OS TCP/IP moves 10.10.10.5 to OSA2
2. z/OS TCP/IP issues gratuitous ARP for 10.10.10.5 with MAC M2
3. Downstream routers on same subnet will update their ARP caches
4. Routers now route packets for both 10.10.10.5 and 10.10.10.6 to OSA2

SHARE
EDUCATE • NETWORK • INFLUENCE

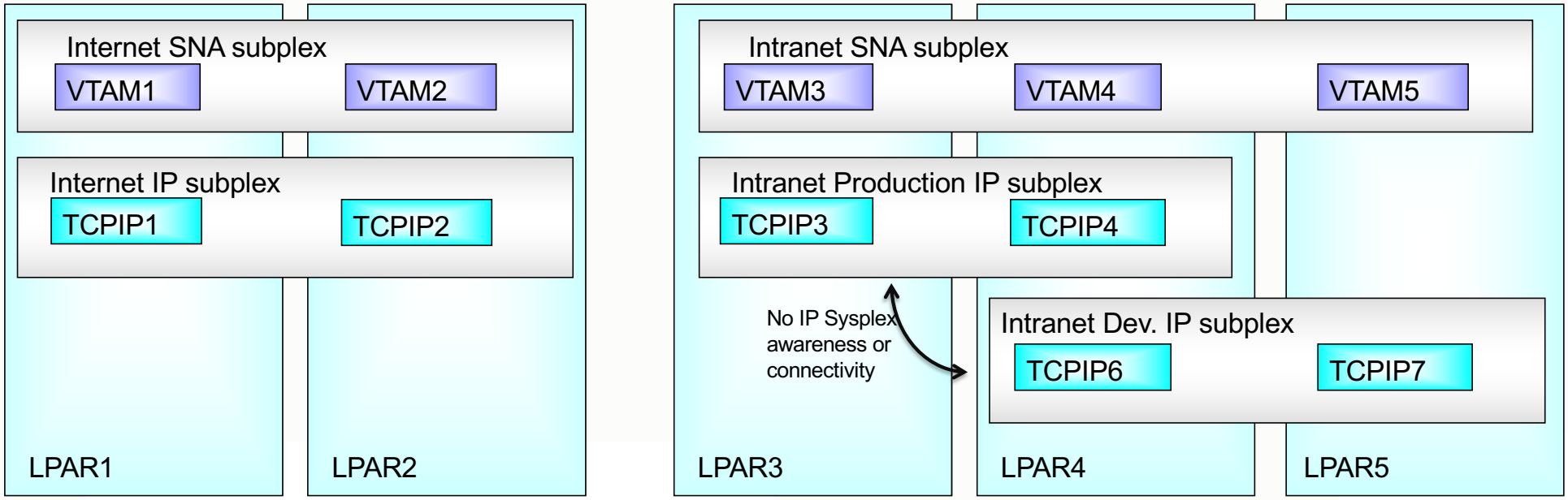
Sysplex Networking Technologies and Considerations

Network subplexing

Networking sub-plexing within a z/OS Sysplex

Subplexing allows network isolation of multiple secure areas

- TCP/IP Sysplex networking:
  - Establish IP Connectivity only to stacks in subplex; dynamic IP address discovery only within a subplex
- Coupling Facility resources:
  - VTAM generic resources, Sysplex ports resources, SWSA resources all defined at subplex level



SHARE
EDUCATE • NETWORK • INFLUENCE

Sysplex Networking Technologies and Considerations

Questions?

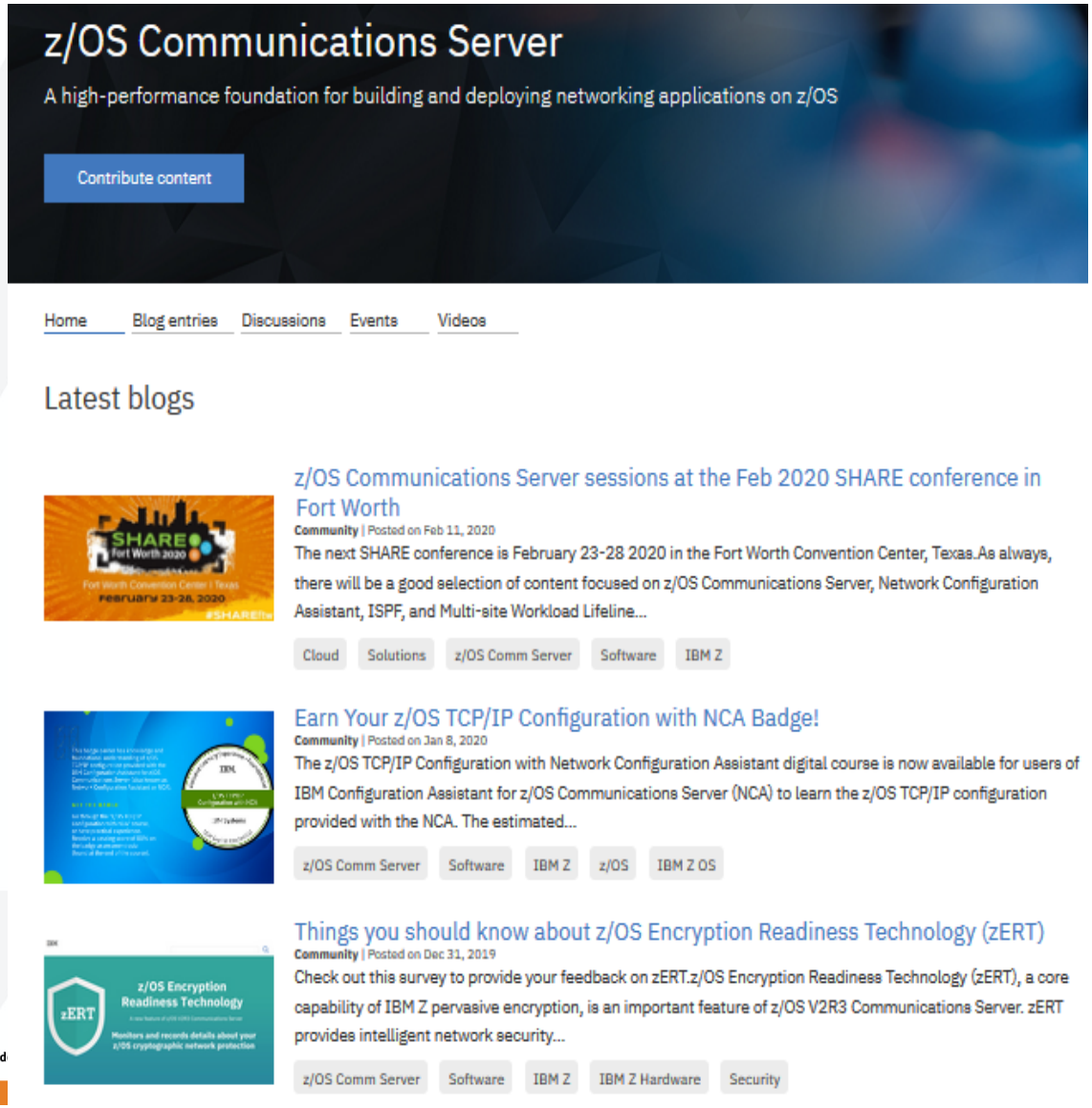
Join Us on IBM Community!

IBM developerWorks platform will be sunset soon.

With **a new and improved platform**, we will continue to provide rich and up-to-date technical content on the IBM Community page, including blogs, videos, and events. We will also migrate some critical entries to the new IBM Community page.

Join us at our new home:
https://www.ibm.com/community/z/software/comm-server/

Scan the QR code to visit the z/OS Communications Server home page on IBM Community.



Online Course and Digital Badge

Networking on z/OS - Foundations

The online course on Networking on z/OS – Foundations is available: https://ibm.biz/zosnetworkingcourse.

After completing this course, you will have foundational understanding of networking on z/OS, including general networking concepts, TCP/IP and SNA communication protocols, network security, and network operations.

If you pass the final assessment of this course, you will earn an IBM Open Badge: https://ibm.biz/zosnetworkingbadge.

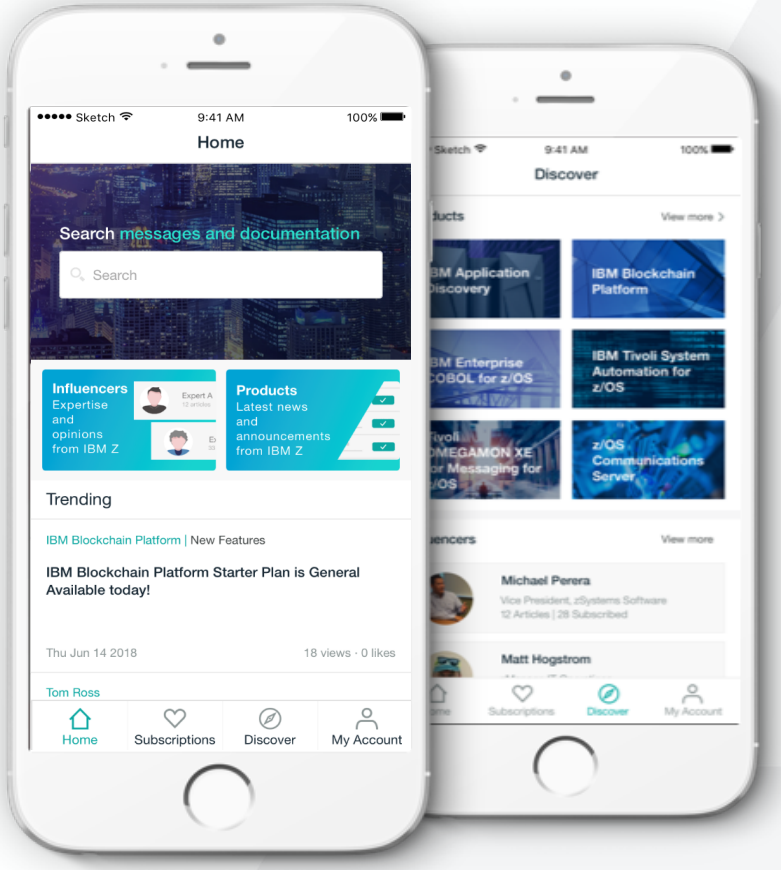
z/OS TCP/IP Configuration with NCA

The online course on TCP/IP profile configuration with the Network Configuration Assistant is available: http://ibm.biz/NCATCPIPcourse.

After completing this course, you will be able to use the NCA to create and manage TCP/IP profiles, including import of existing profiles, dynamic reconfiguration of running stacks, and configuration for disaster recovery and planned outages.

If you pass the final assessment of this course, you will earn an IBM Open Badge: http://ibm.biz/NCAbadge.

IBM Doc Buddy – Your one-stop Z mobile content portal



- **Find mainframe documentation in no time**
  - Integrate mainframe product docs and error messages for 66 products from 141 releases, including z/OS Communications Server
- **Gain insights from best technical and business leaders**
  - Latest product blogs, videos, presentations and white papers for important mainframe products and trending IT business insights
- **Share great content to peers**
  - Connect with your peers via technical content sharing
- **Support 8600+ IBM Z users**
- **Review rating: 4.21/5**

iOS

Android

Your feedback is important!

SHARE
EDUCATE • NETWORK • INFLUENCE

Submit a session evaluation for each session you attend:

SHARE mobile app -or- www.share.org/evaluation