Using z/OS Communications Server for Optimized Workload Balancing to z/OS

Gus Kassimis – kassimis@us.ibm.com
IBM, Enterprise Networking Solutions
Raleigh, NC

Thursday, August 5, 2010: 4:30 PM-5:45 PM

Trademarks, notices, and disclaimers

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

- Advanced Peer-to-Peer Networking®
- AIX®
- alphaWorks®
- AnyNet®
- AS/400®
- BladeCenter®
- Candle®
- CICS®
- DataPower®
- DB2 Connect
- DB2®
- DRDA®
- e-business on demand®
- e-business (logo)
- e business(logo)®
- ESCON®
- FICON®

- GDDM®
- GDPS®
- Geographically Dispersed Parallel Sysplex
- HiperSockets
- HPR Channel Connectivity
- HyperSwap
- i5/OS (logo)
- i5/OS®
- IBM eServer
- IBM (logo)®
- IBM®
- IBM zEnterprise™ System
- IMS
- InfiniBand ®
- IP PrintWay
- IPDS
- iSeries
- LANDP®

- Language Environment®
- MQSeries®
- MVS
- NetView®
- OMEGAMON®
- Open Power
- OpenPower
- Operating System/2®
- Operating System/400®
- OS/2®
- OS/390®
- OS/400®
- Parallel Sysplex®
- POWER®
- POWER7®
- PowerVM
- PR/SM
- pSeries®
- RACF®

- Rational Suite®
- Rational®
- Redbooks
- Redbooks (logo)
- Sysplex Timer®
- System i5
- System p5
- System x®
- System z®
- System z9®
- System z10
- Tivoli (logo)®
- Tivoli®
- VTAM®
- WebSphere®
- xSeries®
- z9®
- z10 BC
- z10 EC

- zEnterprise
- zSeries®
- z/Architecture
- z/OS®
- z/VM®
- z/VSE

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

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

Notes:
- Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.
- IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.
- All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.
- This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.
- All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.
- Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.
- Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

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

Agenda

- ❑ **Introduction**

- ❑ **Network workload balancing overview**

- ❑ **Communications Server load balancing solutions and recent updates**

- ❑ **z/OS Sysplex support for external load balancers**

- ❑ **Summary**

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.

Optimized workload balancing to z/OS

Introduction

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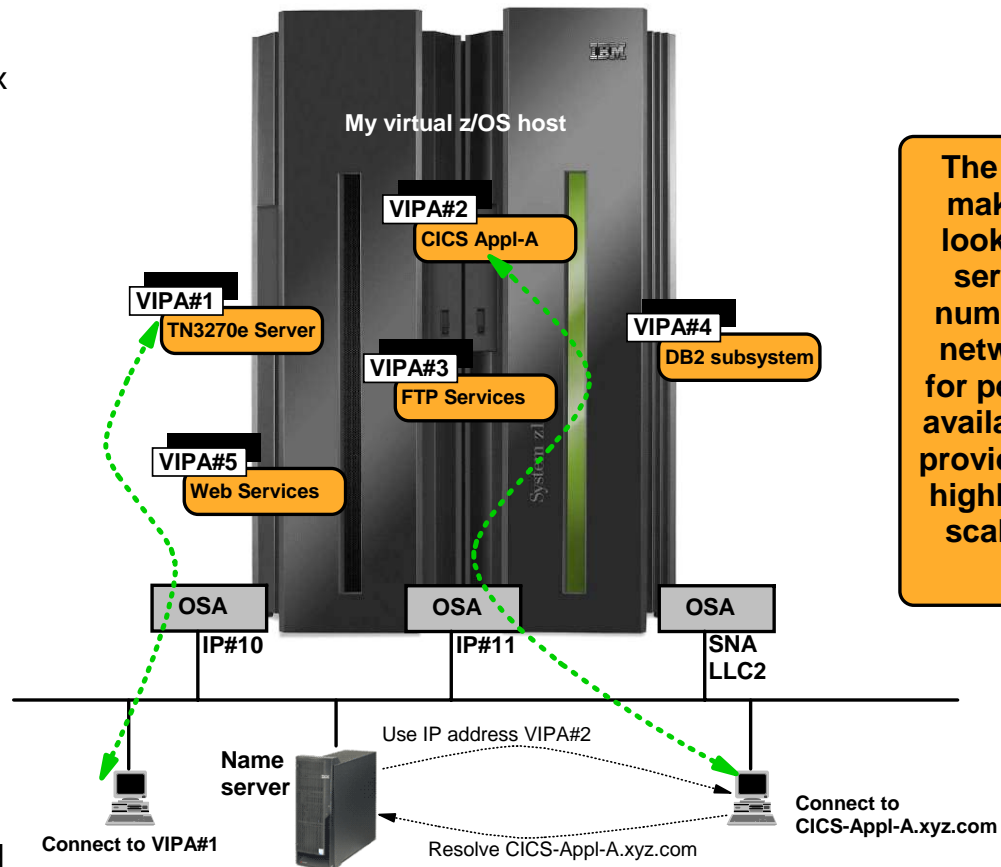
The objective is to make the Sysplex look like one large server that has a number of physical network interfaces for performance and availability - and that provides a number of highly available and scalable services.

Name server

Use IP address VIPA#2

Connect to CICS-Appl-A.xyz.com

Connect to VIPA#1

Resolve CICS-Appl-A.xyz.com

SNA and TCP/IP

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

A summary of the different types of z/OS VIPA addresses

- **Static VIPA**
  - Belongs to one TCP/IP stack. Manual configuration changes are needed to move it.
    - No dependencies on Sysplex functions – can be used in non-Sysplex LPARs
    - Required for certain functions such as Enterprise Extender
    - Beneficial for interface resilience, source IP addressing, etc.

- **Dynamic VIPA (DVIPA)**
  - **Stack-managed (VIPADEFINE/VIPABACKUP)**
    - Belongs to one TCP/IP stack, but backup policies govern which TCP/IP stack in the Sysplex takes it over if the primary TCP/IP stack leaves the Sysplex
    - Individual stack-managed dynamic VIPAs can be moved between primary and backup stacks using MVS operator commands
  - **Application-specific also known as bind-activated (VIPARANGE)**
    - Belongs to an application. Becomes active on the TCP/IP stack in the Sysplex where the application is started. Moves with the application.
  - **Command- or utility activated (VIPARANGE)**
    - Belongs to whatever TCP/IP stack in the Sysplex on which a MODDVIPA utility to activate the address has been executed.
    - Moves between TCP/IP stacks based on execution of the MODDVIPA utility.
  - **Distributed also known as a DRVIPA or sometimes DDVIPA (VIPADEFINE/VIPABACKUP + VIPADISTRIBUTE)**
    - Used with Sysplex Distributor as a cluster IP address that represents a cluster of equal server instances in the Sysplex.
    - From a routing perspective it belongs to one TCP/IP stack.
    - From an application perspective it is distributed among the TCP/IP stacks in the Sysplex where an instance of the server application is executing.

Why do I want to use virtual IP addresses instead of physical interface IP addresses?

- **Physical network interface resilience:**

  - Communication with a server host is unaffected by server physical network interface failures. As long as just a single physical network interface is available and operational on a server host, communication with applications on the server host will persist.

- **Application access independent of network topology:**

  - Separates network topology from server application topology - a VIPA address can be used to identify a server application instead of a physical network interface.

  - Allows network administrators to renumber physical network topology

    - no impact to end-user accessing server applications by IP address
    - no changes needed in DNS or hosts file configuration
    - no impact to firewall filtering rules

- **Single system image:**

  - Allows the Sysplex to be perceived as a single large server node, where VIPA addresses identify applications independently of which images in the Sysplex the server applications execute on.

  - Applications retain their identity when moved between images in a Sysplex.

  - Multiple instances of a server application can be accessed as one server.

IBM Software Group – Enterprise Networking Solutions

Optimized workload balancing to z/OS

Network workload balancing basics

IBM®

© Copyright International Business Machines Corporation 2010 and SHARE. All rights reserved.

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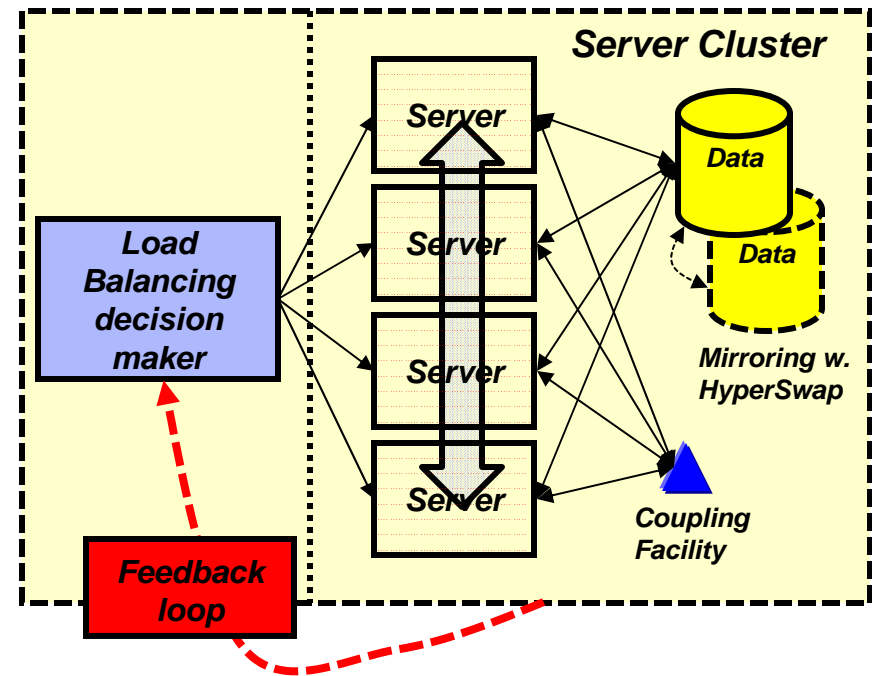
In order for the load balancing decision maker to meet those objectives, it must be capable of obtaining feedback dynamically, such as server instance availability, capacity, performance, and overall health.

Network workload balancing and the OSI layers – layer-3
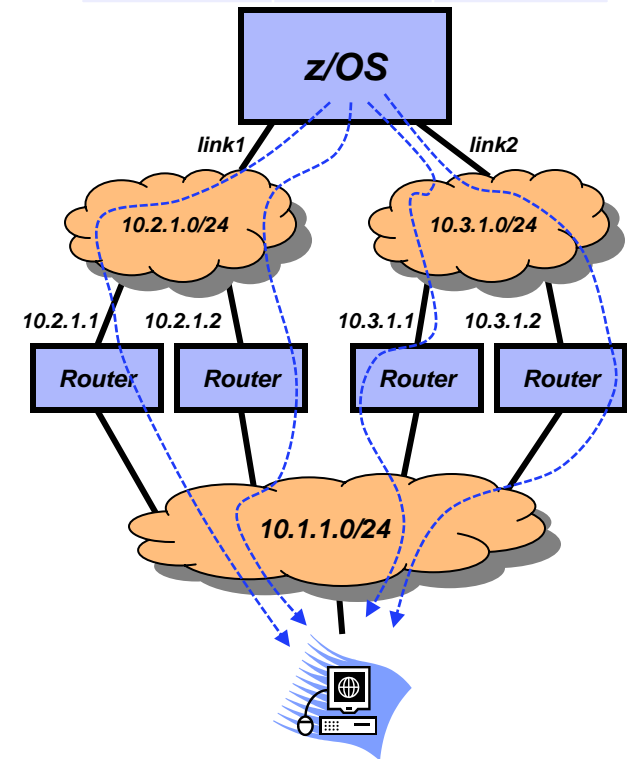
- IP packet transmission over multiple equal-cost routing table entries
  - Generally referred to as multi-path

Destination	Next Hop	Interface
10.1.1.0/24	10.2.1.1	Link1
10.1.1.0/24	10.2.1.2	Link1
10.1.1.0/24	10.3.1.1	Link2
10.1.1.0/24	10.3.1.2	Link2

- General methods for choosing the outbound link and next-hop router:

  - **Round robin**
    - Spraying packets over all equal-cost routes
    - z/OS supports this method (MULTIPATH PERPACKET)
    - Generally not recommended due to out-of-sequence packets

  - **All packets for a given connection**
    - Choose a route when the connection is established
      - When the router sees a SYN segment
    - z/OS supports this method (MULTIPATH PERCONNECTION)
    - Generally recommended method on z/OS

  - **All packets for a given destination IP address**
    - Choose a link the first time an IP packet to an IP address is being forwarded
    - Most routers use this method by default
      - Note: routers determine inbound multi-path processing to z/OS, not z/OS
    - Not very efficient since all packets destined for a specific IP address will use a single outbound interface on the router and first-hop router

- Use of IPSec security is transparent to multi-path support

Network workload balancing and the OSI layers – layer-4

- **For UDP**
  - Each UDP datagram for a given destination (destination IP address and port) is balanced
    - Always stateless
  - Not used very often
  - Could cause severe havoc for most UDP-based applications

- **For TCP**
  - When a TCP SYN segment for a given destination (destination IP address and port) is received, a load balancing decision is made
    - Always stateful – the load balancer remembers the decision so all IP packets belonging to that connection are forwarded to the server instance chosen for the connection
  - The load balancer needs a way to determine when the connection terminates to clear up the connection cache
    - External load balancers must be in the routing path for all outbound packets for the load-balanced connections
    - Sysplex Distributor uses a side-band channel (XCF messaging) to notify the load-balancing node (the distributing stack), when a connection has terminated
  - Use of SSL/TLS security is transparent to the load balancer
    - The load balancer only depends on TCP header information

- An IPSec VPN cannot traverse a layer-4 load balancer, but must be terminated on or before the load balancer
  - Potentially with back-to-back VPNs
  - Sysplex Distributor handles this via a function known as Sysplex Wide Security Associations (SWSA)

Network workload balancing and the OSI layers – layer-5 and above

- This type of load balancing is generally referred to as **contents-based load balancing or routing (CBR)** or, more recently, application delivery controllers (ADC)

- The load balancer in a sense becomes a proxy server or application layer gateway (ALG) when doing CBR
  - The load balancer must establish the full connection with the client and receive the initial inbound data
    - respond with a SYN/ACK
    - receive the final ACK
    - read the initial data from the client off the connection – typically an HTTP request header
      - In theory, CBR can be used for applications other than HTTP, but very rarely seen
    - make a load-balancing decision based on the content of the HTTP request header
  - At this point the load balancer establishes a new connection with the chosen server and "proxies" inbound data to the server and outbound from the server to the client
    - This may be done using two simple back-to-back connections in the load balancer (pure proxy)
    - Some load balancers do this at the TCP layer, where they after the initial sequences have been performed, "proxy" the TCP segments by swapping control information in the packet headers

- If the connections are secured with SSL/TLS, the load balancer also has to perform SSL offload
  - The SSL/TLS connection must terminate on or before the load balancer
  - A second SSL/TLS connection can, if needed, be used between the load balancer and the server

- An IPSec VPN cannot traverse a layer-4 load balancer, but must be terminated on or before the load balancer
  - Potentially with back-to-back VPNs

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.
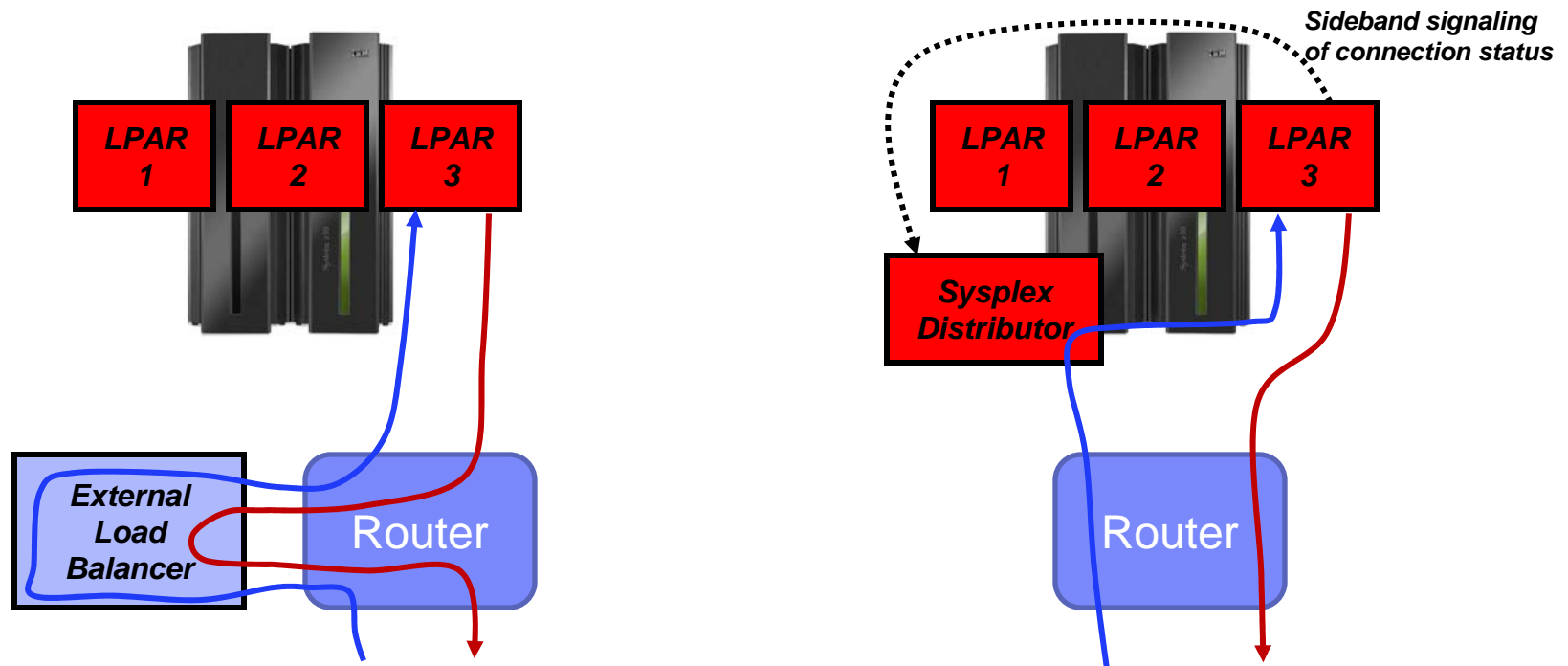
Optimized workload balancing to z/OS

z/OS Communications Server IP workload balancing update

IBM ®

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

App1 **App2**

PortSharing

SD

WLM

App3 **App4**

PortSharing
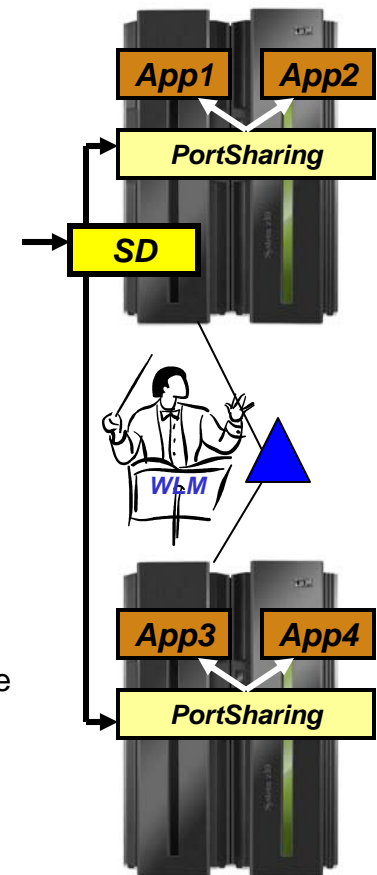
Sysplex Distributor distribution method overview

- 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

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*

Sysplex Distributor and affinities

- Since Sysplex Distributor is a layer-4 load balancer, it cannot base affinity on content of data, such as HTTP cookies, etc.

- For a given DDVIPA and port combination, Sysplex Distributor does support a timer-based affinity
  - Per client (source) IP address
  - Specified in seconds
  - Default is no affinity

- Can be used with all distribution methods (does not apply to HOTSTANDBY)

- Note: be aware of proxy servers
  - All connections will come from one source IP address (that of the proxy server)

Connection	Sysplex Distributor action
Connection 1	A server (SERVER1) is chosen based on the distribution method in use
Connection 2	Goes to SERVER1 – client already has affinity with it through an existing connection
Connection 3	Because the affinity timer that started after connection 2 ended has not yet elapsed, the original affinity to SERVER1 is still in effect, so this connection also goes to SERVER1
Connection 4	The affinity timer that was started after connection 3 ended, has expired, so a new server (SERVER2) can be chosen for this connection

Connect to 10.2.1.1 port 3002

SERVER 1   **SERVER 2**

Time

Connection 1
Affinity timer
Connection 2
Affinity timer
Connection 3
Affinity timer
Connection 4

Source IP address 10.1.1.1
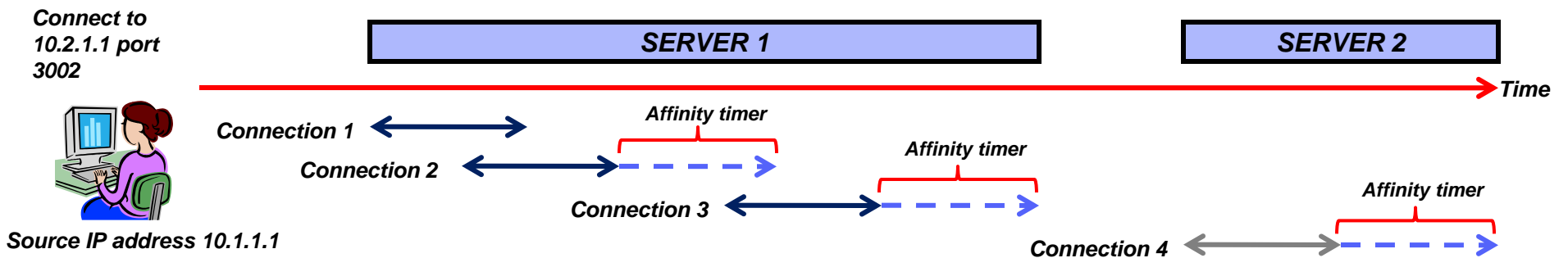
Why you may not always want WLM to "decide"

LPAR Capacity

LPAR1

LPAR2

LPAR1

LPAR2

SHAREPORT

LPAR1

LPAR2

Fence as spare capacity
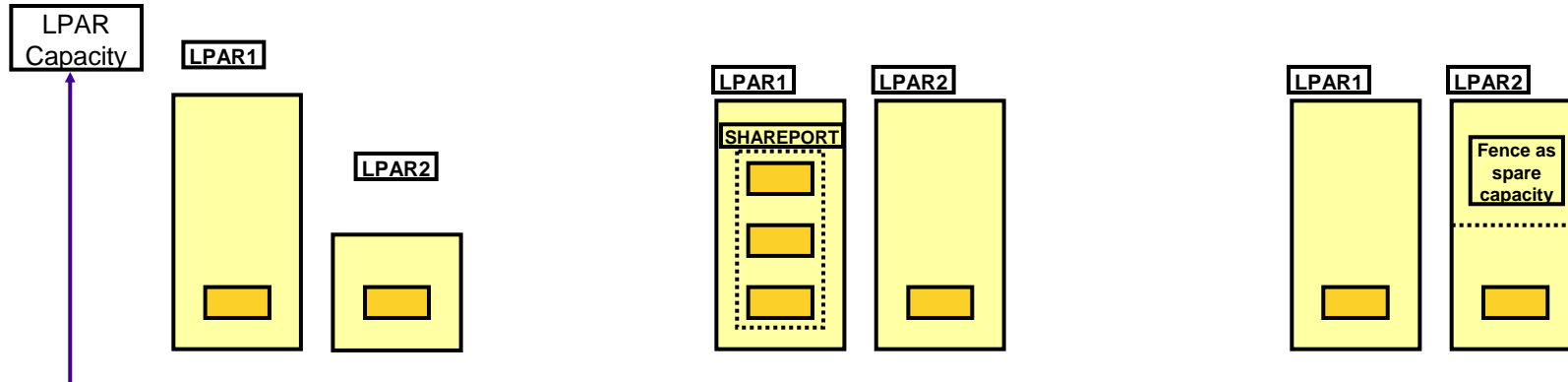
➢**Target systems vary significantly in terms of capacity (small systems along with larger systems).**

- ► WLM recommendations may favor the larger systems significantly. However, the target application may not scale well to larger systems, being unable to take full advantage of the additional capacity on the larger systems. The result can be that these types of servers when running on larger systems get inflated WLM recommendations and as a result they get overloaded with work.

- ► ServerWLM can help address this issue by incorporating the target application's Performance Index into the recommendation (BASEWLM does not)

➢**SHAREPORT is deployed, yet not all systems have the same number of SHAREPORT server instances (one has three the other has only one).**

- ► The simple RoundRobin or WLM recommendations do not change distribution based on the number of server instances on each target. RoundRobin distributes 1 connection per target stack regardless of the number of shareport server instances on that stack. WLM Server-specific weights from a target stack with multiple server instances reflect the average weight.

➢**Customers would like to reserve some capacity on certain systems for batch type of workloads that get injected into the system during specific time periods and which have specific time window completion requirements.**

- ► If that system is also a target for long running DDVIPA connections, WLM recommendations will allow that available capacity to be consumed and thereby potentially impact the completion of the batch jobs or vice versa (the connections on that system may suffer from a performance perspective when those jobs are running).

WEIGHTEDACTIVE method: how it works

- Weights to be configured on the VIPADISTRIBUTE statement per destination IP address

- Weights to balance active connections, not just incoming connections
  - Objective is to keep the number of active connections distributed according to the configured weights
    - More optimal than traditional round robin or weighted round robin algorithms
- This is the preferred method when not using WLM input as a decision criteria

LPAR1	LPAR2	LPAR3	LPAR4
		Fence as spare capacity	SHAREPORT

SD

| Weight 2 | Weight 1 | Weight 1 | Weight 3 |
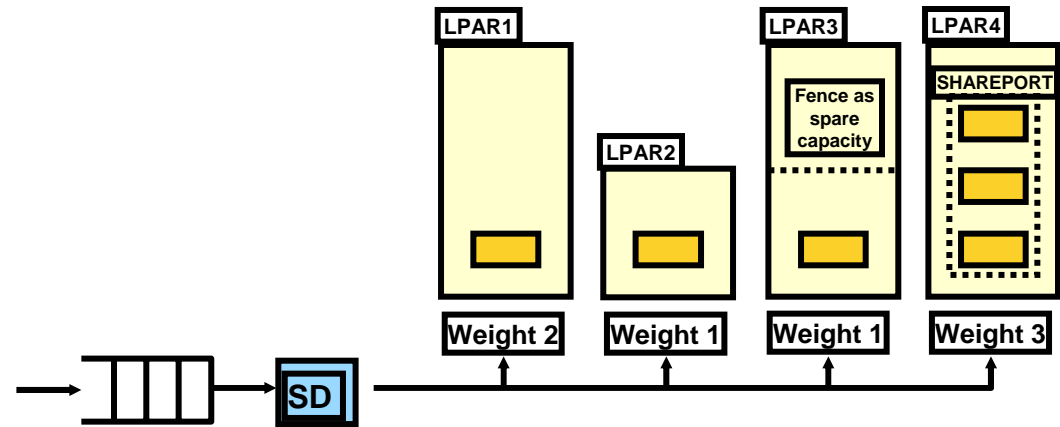
Case 1	Configured weights	Current number of active connections	Connection goal	Status
LPAR1	20	15	20	Below
LPAR2	10	10	10	On target
LPAR3	10	10	10	On target
LPAR4	30	30	30	On target

Case 2	Configured weights	Current number of active connections	Connection goal	Status
LPAR1	20	30	20	Above
LPAR2	10	10	10	On target
LPAR3	10	10	10	On target
LPAR4	30	20	30	Below

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
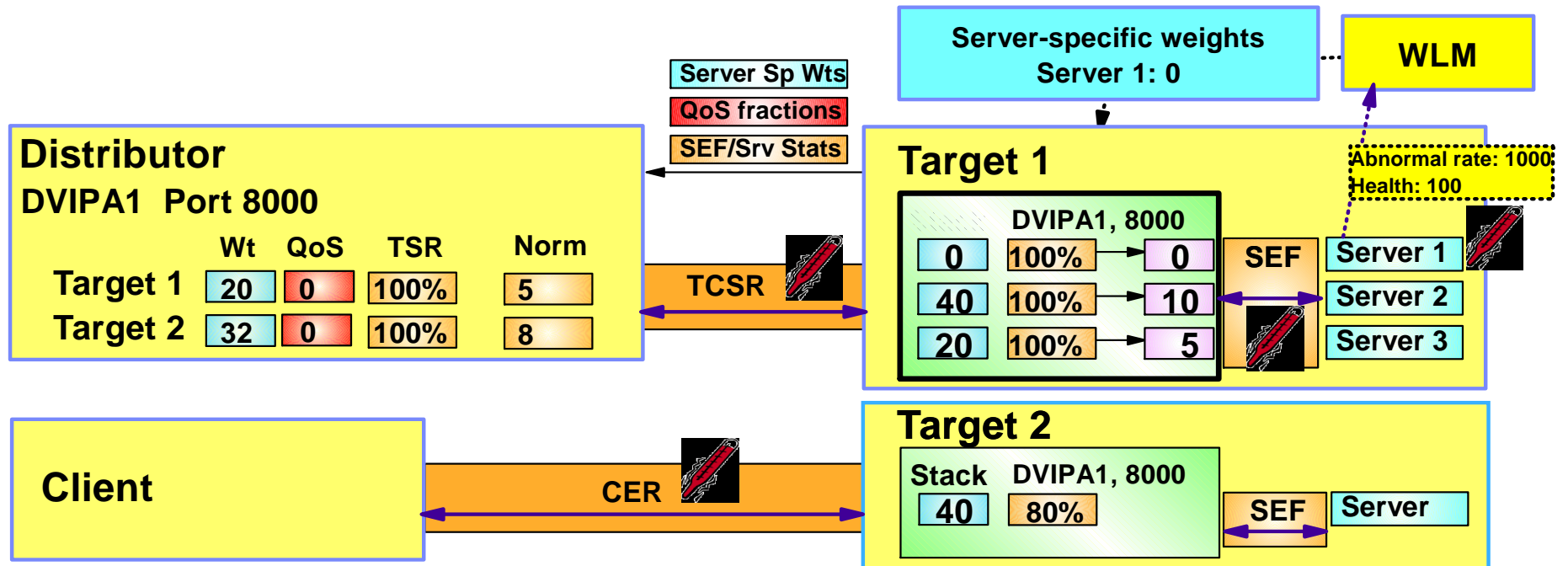
What impacts the final selection of a target server instance?

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

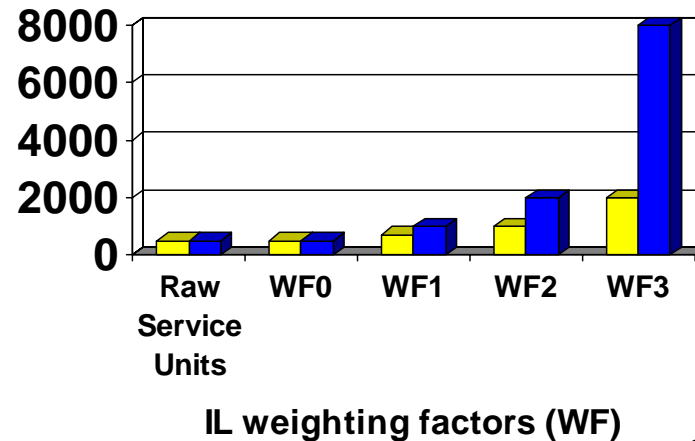 – In this example, LPAR2

New workload to run at Importance Level 2

Displaceable service units per importance level
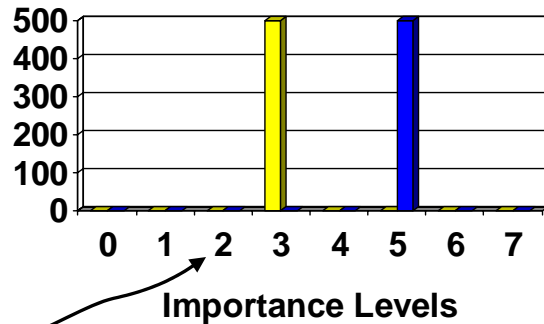
Adjusted displaceable service units

Importance Levels

IL weighting factors (WF)

☐ LPAR1 ■ LPAR2

☐ LPAR1 ■ LPAR2

z/OS V1R11

BASEWLM method: specialty processors - overview

- **When using WLM system weights**
  - WLM returns
    - The raw CP, zAAP, and zIIP system weights
  - Sysplex Distributor calculates proportional and composite weights using configured proportions
    - VIPADISTRIBUTE BASEWLM PROCTYPE CP 20 ZAAP 80 ZIIP 0
    - Composite weight is determined from the proportional weights

WLM

1 Raw weights: CP 30 ZAAP 60 ZIIP 60

PROCTYPE: CP 20 ZAAP 80 ZIIP 0

TCP/IP

2 Proportional weights: CP 6 ZAAP 48 ZIIP 0

TCP/IP

3 Composite weight: 54

TCP/IP Target Server Responsiveness: 90%

TCP/IP

Health-adjusted weight: 49
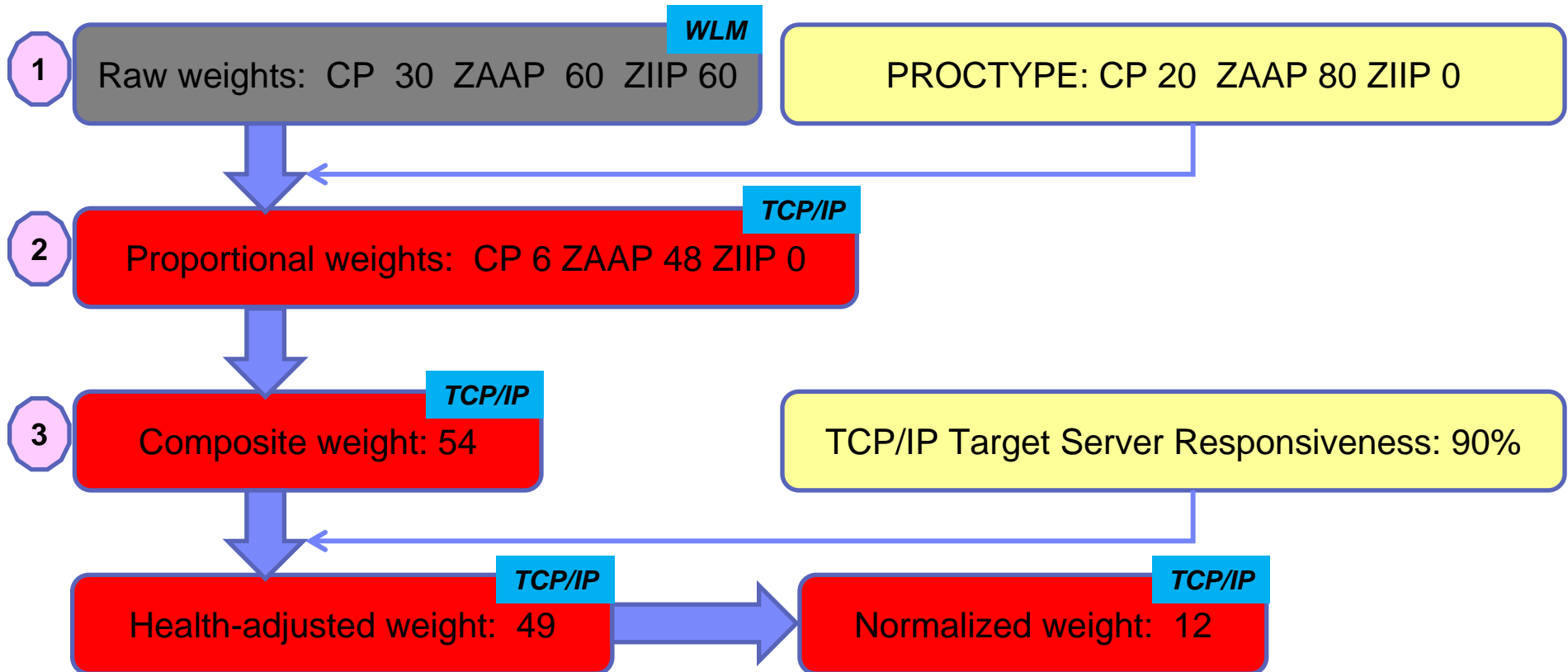
TCP/IP

Normalized weight: 12

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
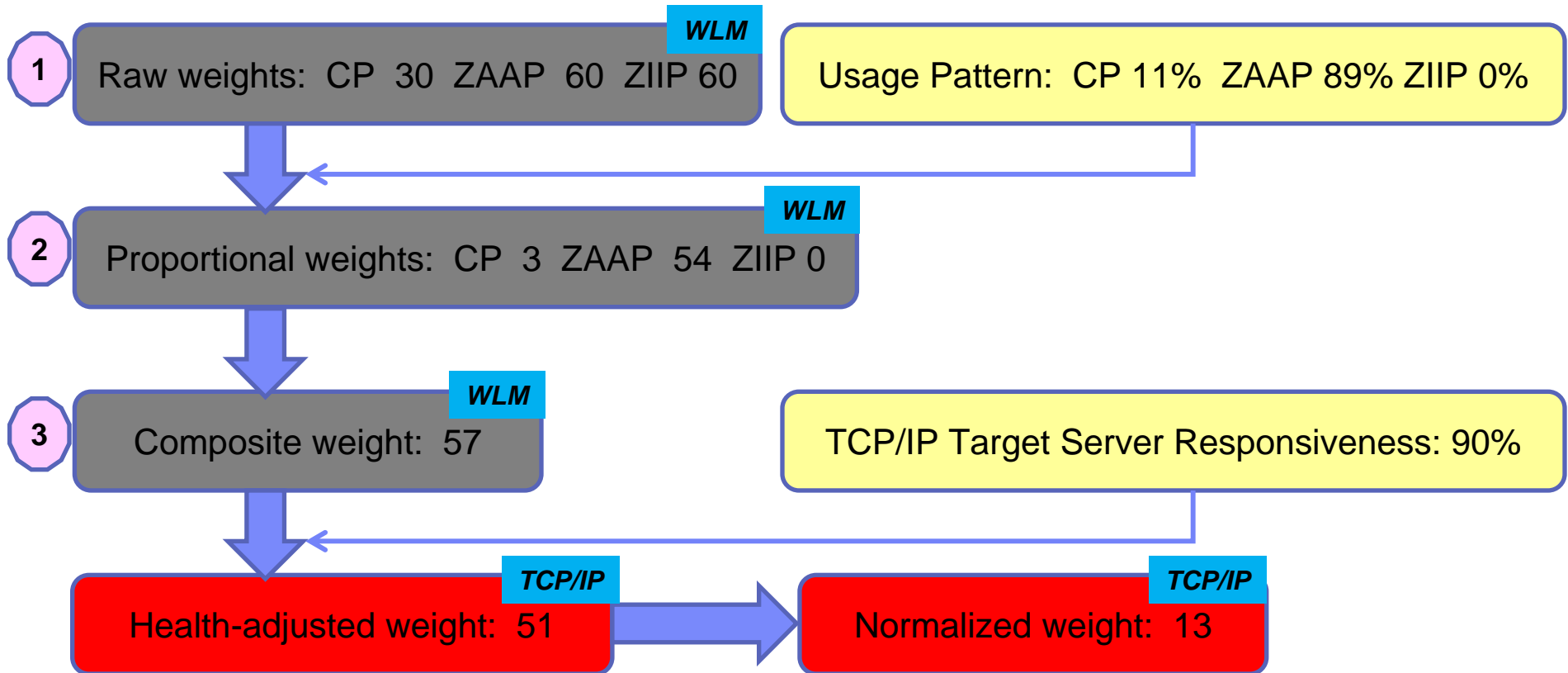
SERVERWLM method: zIIP and zAAP cross-over to CP

- 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

Displaceable Service Units

Application Workload Design

Relative weights of LPAR1 and LPAR2
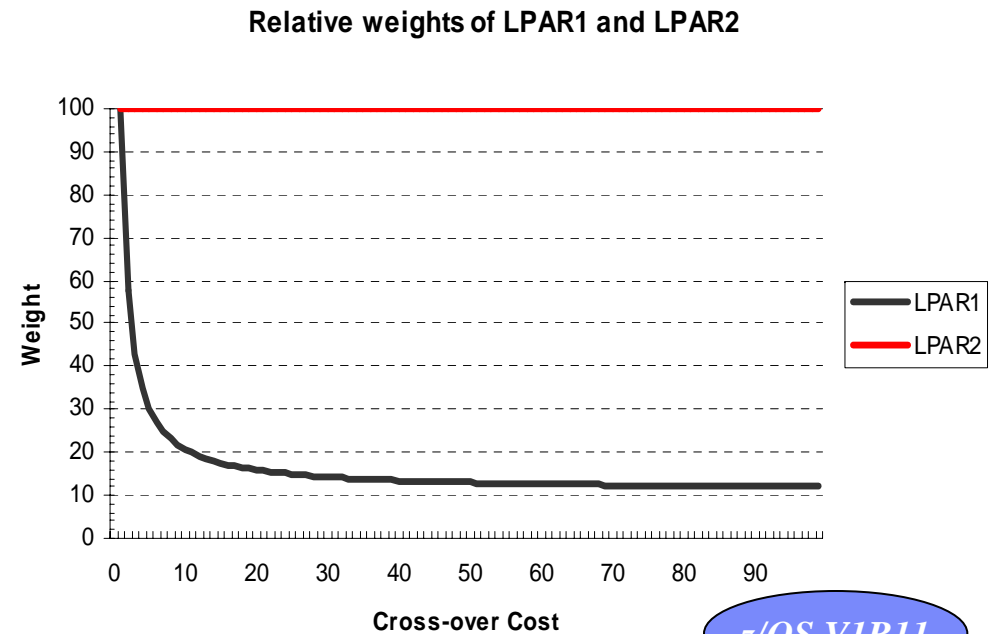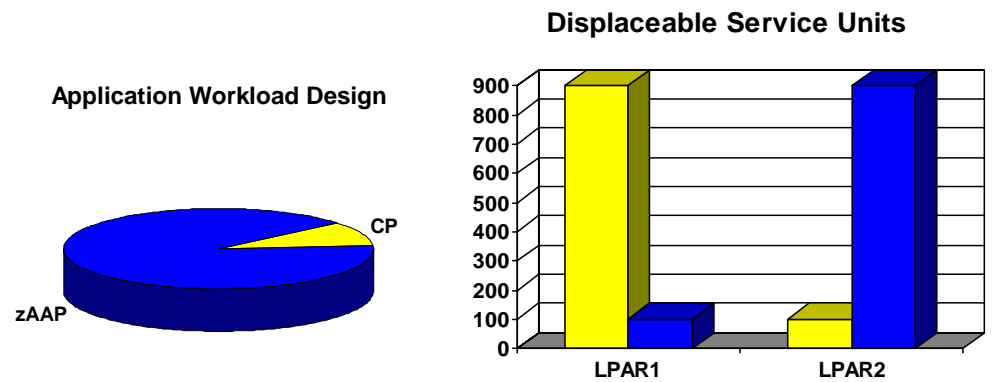
z/OS V1R11

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
```
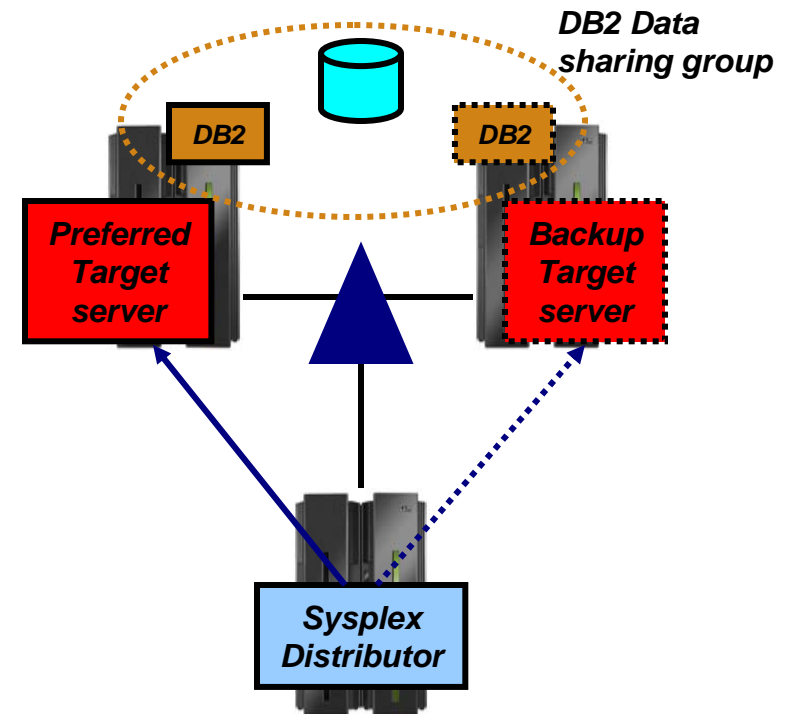
Sysplex Distributor hot standby support

z/OS V1R12

- Data Sharing provides for highly scalable and highly available configuration

  - Additional application instances and LPARs can be cloned and added to increase capacity and improve performance

  - But what if the workload can comfortably fit within a single LPAR?

    - Data Sharing becomes primarily an availability feature

    - A Hot Standby configuration would allow all workload to be routed to a single LPAR

      - Minimizing data sharing overhead!
      - While retaining high availability!

- This configuration is currently possible via Policy Agent

  - Using QoS policy:

    - ibm-policyGroupForLoadDistribution:TRUE

- But several users had requested a simpler mechanism for doing this via the TCP/IP profile

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:
```
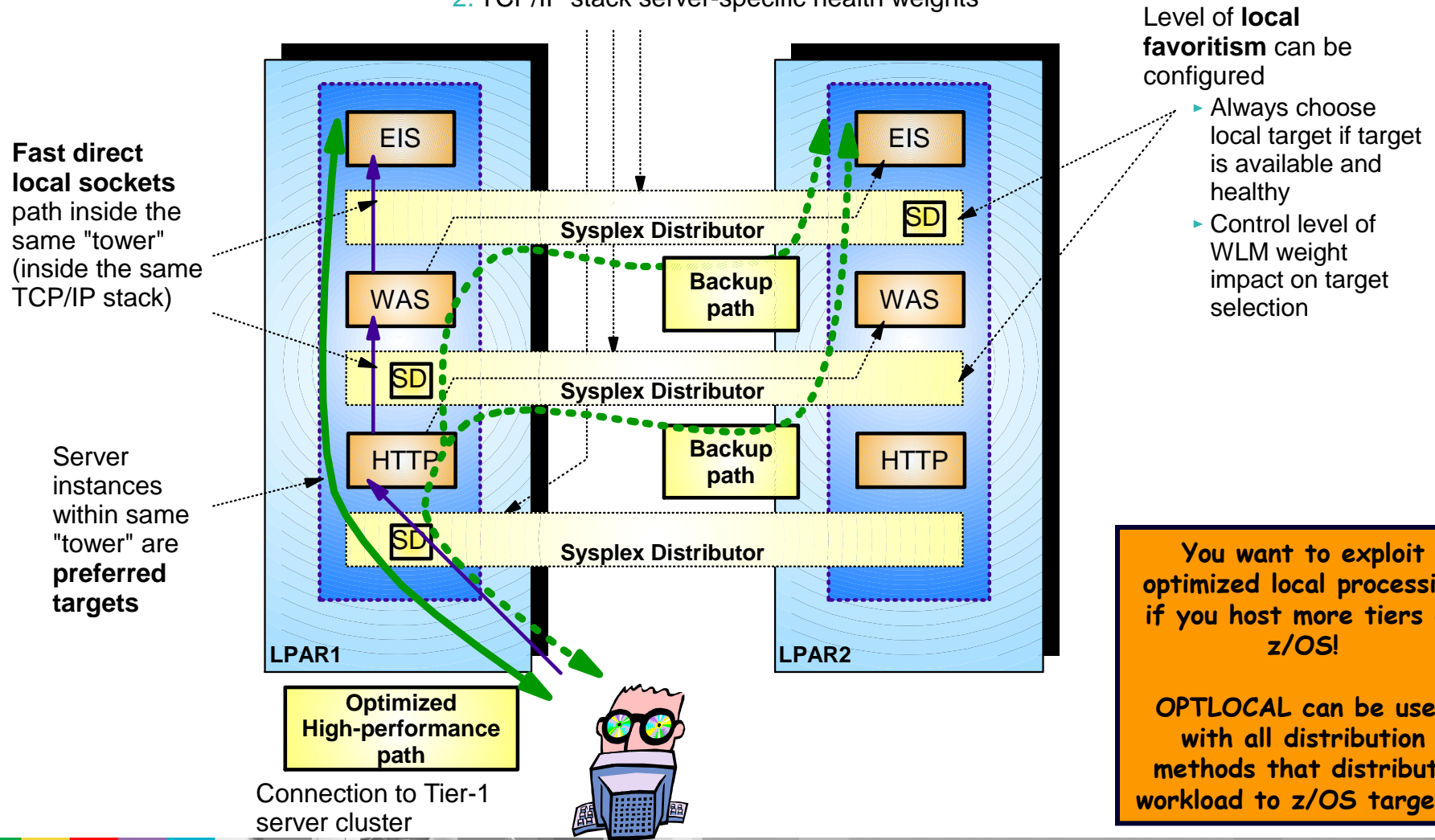
Sysplex Distributor hot standby support…

- 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
```
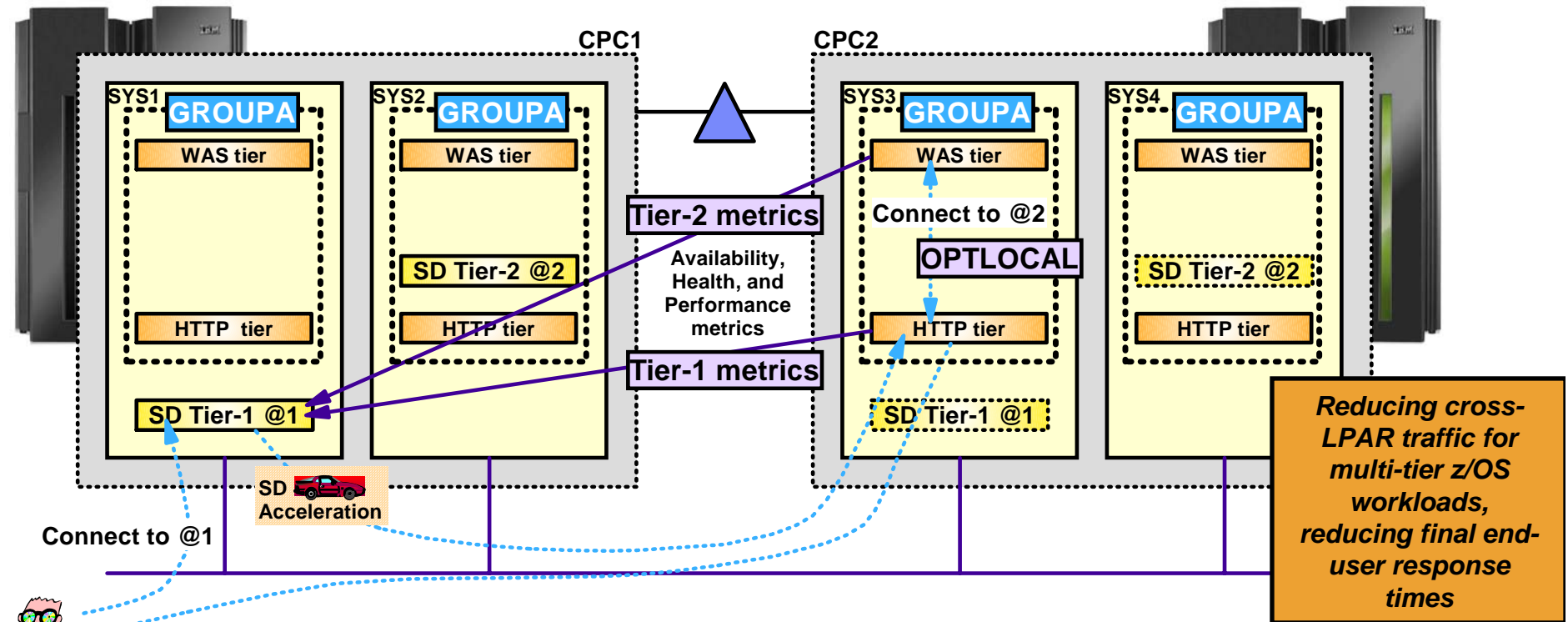
Optimized local (OPTLOCAL) – the basics (as it was originally implemented in z/OS V1R8)

1. WLM LPAR and server-specific performance weights
2. TCP/IP stack server-specific health weights

Level of **local favoritism** can be configured

- ► Always choose local target if target is available and healthy
- ► Control level of WLM weight impact on target selection

Fast direct local sockets path inside the same "tower" (inside the same TCP/IP stack)

Server instances within same "tower" are **preferred targets**

EIS

Sysplex Distributor

SD

Backup path

WAS

SD

Sysplex Distributor

Backup path

HTTP

SD

Sysplex Distributor

EIS

SD

WAS

HTTP

LPAR1

LPAR2

Optimized High-performance path

Connection to Tier-1 server cluster

You want to exploit optimized local processing if you host more tiers on z/OS!

OPTLOCAL can be used with all distribution methods that distribute workload to z/OS targets.

Optimized local (OPTLOCAL) and grouping of multiple z/OS tiers

CPC1   **CPC2**

SYS1   **GROUPA**   **SYS2**   **GROUPA**   **SYS3**   **GROUPA**   **SYS4**   **GROUPA**

WAS tier   **WAS tier**   **WAS tier**   **WAS tier**

Tier-2 metrics

Connect to @2

Availability, Health, and Performance metrics

SD Tier-2 @2   **OPTLOCAL**   **SD Tier-2 @2**

HTTP tier   **HTTP tier**   **HTTP tier**   **HTTP tier**

Tier-1 metrics

SD Tier-1 @1   **SD Tier-1 @1**

SD Acceleration

Connect to @1

Reducing cross-LPAR traffic for multi-tier z/OS workloads, reducing final end-user response times
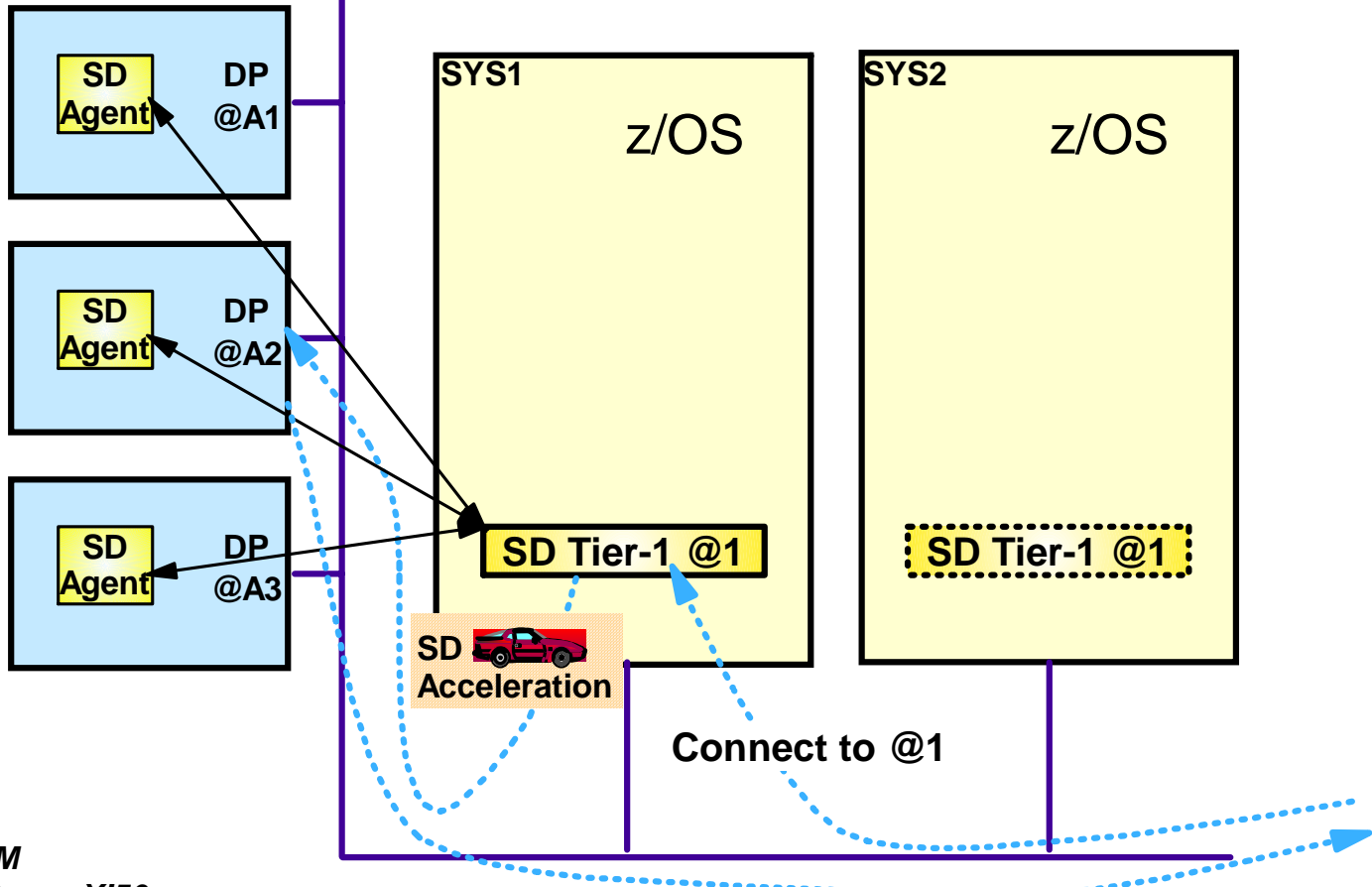
- **Sysplex Distributor in z/OS V1R11 adds support for grouping two tiers of z/OS servers together**
- **Sysplex Distributor will for such tiered workload include availability, health, and performance metrics of both the tier-1 z/OS server and the tier-2 z/OS server on a target LPAR when determining which LPAR the tier-1 connection should go to.**
  - This will increase the likelihood of such connections gaining the performance benefits of optimized local support, such as use of fast local sockets.

z/OS V1R11
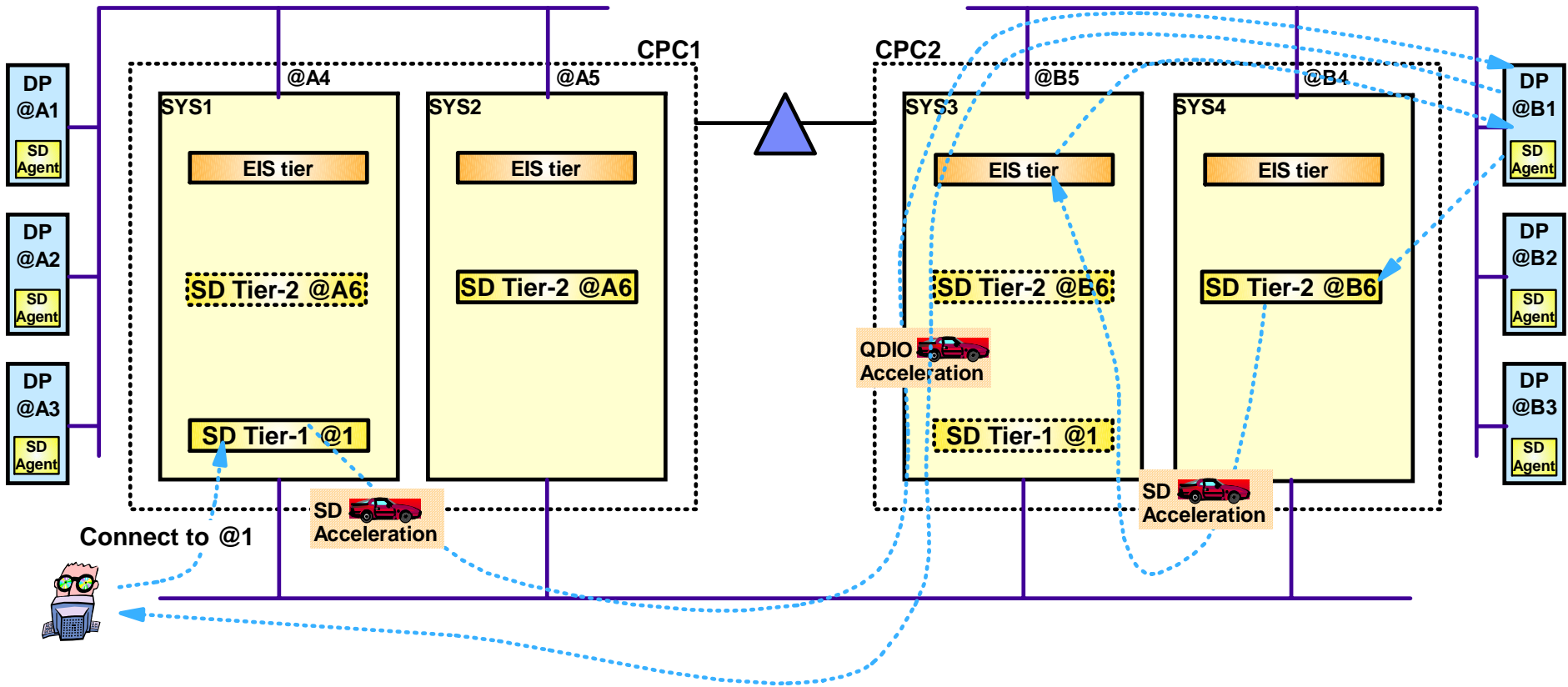
DataPower integration with Sysplex Distributor

z/OS V1R11

DataPower
appliances

SD Agent | DP @A1

SD Agent | DP @A2

SD Agent | DP @A3

SYS1
z/OS

SD Tier-1 @1

SD Acceleration

SYS2
z/OS

SD Tier-1 @1

Connect to @1

Note: Requires IBM WebSphere DataPower XI50 with firmware version V8.3.1

DataPower integration - multi site/CPC with private data network

CPC1

@A4   @A5

SYS1   SYS2

EIS tier   EIS tier

SD Tier-2 @A6   SD Tier-2 @A6

SD Tier-1 @1

CPC2

@B5   @B4

SYS3   SYS4

EIS tier   EIS tier

SD Tier-2 @B6   SD Tier-2 @B6

QDIO Acceleration

SD Tier-1 @1

SD Acceleration

DP @A1 — SD Agent
DP @A2 — SD Agent
DP @A3 — SD Agent

DP @B1 — SD Agent
DP @B2 — SD Agent
DP @B3 — SD Agent

Connect to @1

SD Acceleration

Note: Requires IBM WebSphere DataPower XI50 with firmware version V8.3.1

z/OS V1R11

Optimized workload balancing to z/OS

z/OS Sysplex support for external load balancers

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.
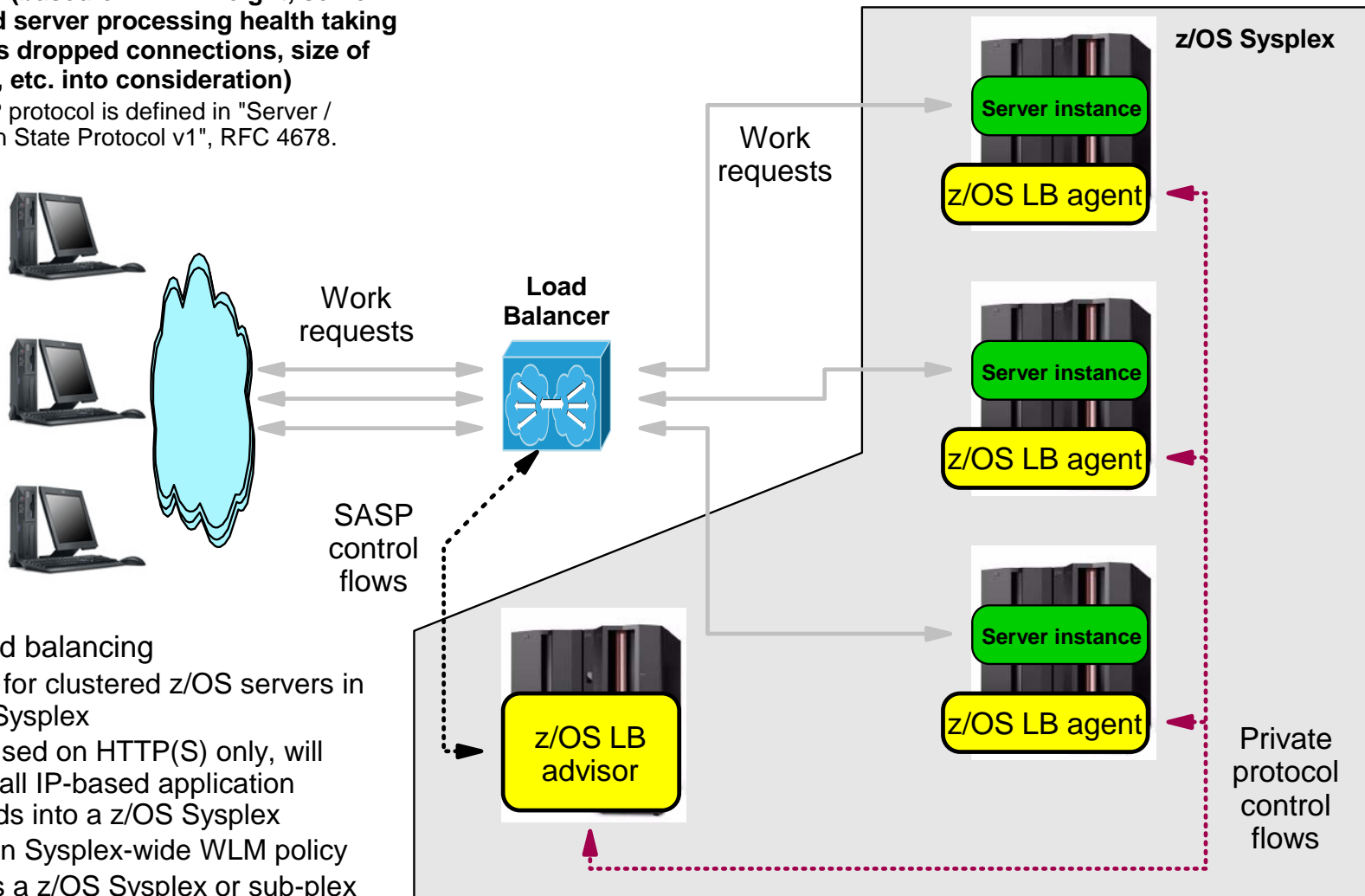
z/OS workload balancing

▶ 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

Server instance

z/OS LB agent

SASP control flows

Server instance

z/OS LB agent

z/OS LB advisor

Private protocol control flows

Optimized workload balancing to z/OS

Summary

Sysplex-internal or external load balancer?

Internal (SD)

SD

z/OS LPAR | z/OS LPAR | z/OS LPAR

OSA

inbound data path | outbound data path

Switch

External (F5, Cisco CSM, etc.)

z/OS LBA

z/OS LPAR | z/OS LPAR | z/OS LPAR

OSA

inbound data path | outbound data path

Control path (SASP connection)

Switch | Load balancer node
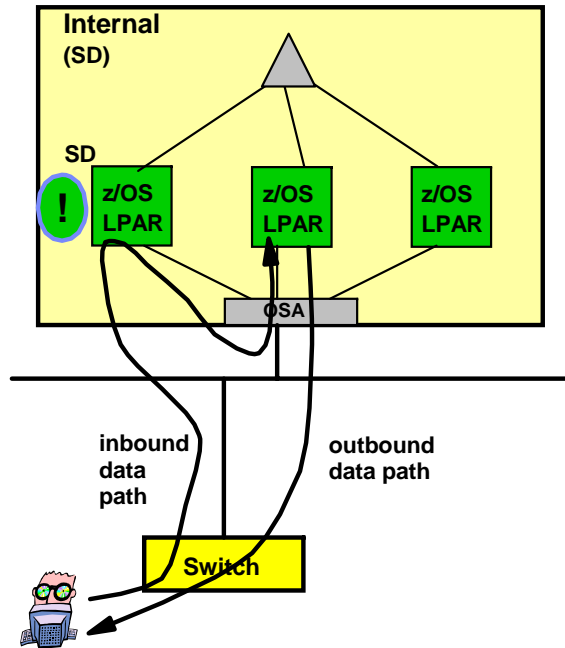
➤ **Sysplex Distributor has realtime information available...**
- ▸ **more timely capacity information**
- ▸ **QoS from Service Policy Agent**
- ▸ **application-independent server availability**

➤ **No problems with shared OSA adapters; no intermediate routers**

➤ **Uses Sysplex resources for routing of inbound traffic**
- ▸ **Inbound traffic funneled through single point**
- ▸ **Routing stack uses System z MIPs for inbound routing**
  - – **z/OS V1R11 significantly reduces this overhead and cost**
- ▸ **Routing may be done via shared LAN or XCF**
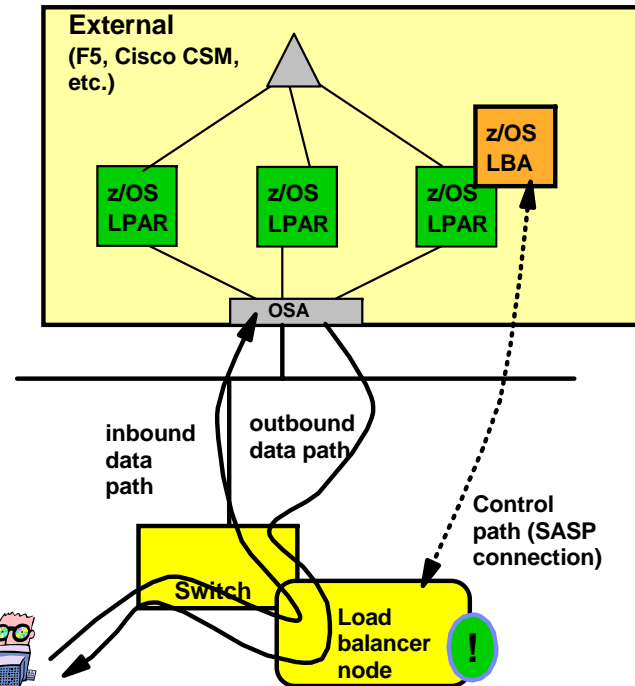
➤ **Specialized routing hardware may be more cost-effective**

➤ **May be configured for no single point of traffic flow**

➤ **z/OS Load Balancing Advisor offers sysplex insight (server and stack availability and WLM information)**
- ▸ **If external load balancer supports the SASP protocol**
- ▸ **Not many do: Cisco CSM, F5's BigIP are among the exceptions**
  - – **Cisco's new ACE switch does not support SASP**

➤ **Without SASP support, the external load balancer will need to use application-specific health probes ("application ping") to detect availability of servers, and it will have no WLM insight.**

➤ **There are problems with shared OSA adapters or other intermediate routers based on how the load balancer is configured (generally server NATing works OK)**

➤ **Both inbound and outbound traffic must pass through load balancing node**

A few final words about a z/OS Sysplex load balancing strategy

- DNS/WLM as workload balancing can no longer be used
  - z/OS V1R8 provides a replacement technology for the dynamic DNS registration/deregistration functions of the old DNS/WLM technology.
  - DNS/WLM usage for load balancing should be replaced with Sysplex Distributor
    - In z/OS V1R11, the DNS/WLM support has been removed

- Where HTTP workload is to be balanced based on content of HTTP requests, an outboard load balancer that supports contents inspection must be deployed
  - If HTTPS workload is to be included, the load balancing node must be accompanied by an SSL/TLS offload technology
  - Can be combined with a cache appliance for improved performance

- UDP workload balancing must be deployed using an outboard load balancer - SD does not support UDP balancing

- Sysplex Distributor is the recommended technology for balancing the remaining TCP connection into the z/OS Sysplex environment
  - SD has more real-time information available than outboard load balancers - even with outboard load balancers using the SASP protocol
    - Better load balancing quality of service
  - The desire to have full z/OS administrative control over TCP workload management in the Sysplex is often a major criteria in deciding to use Sysplex Distributor
  - With the SD accelerator function in z/OS V1R11, the added cost (in terms of latency and CPU usage) of routing inbound data through the SD node, is significantly reduced
    - Unlike outboard load balancers, SD allows response flows to go directly from the target server to the client
    - Most outboard load balancers require inbound and outbound flows to be routed through the load balancer

For more information

URL	Content
http://www.twitter.com/IBM_Commserver	IBM Communications Server Twitter Feed
http://www.facebook.com/IBMCommserver	IBM Communications Server Facebook Fan Page
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.ibm.com/software/network/commserver/z_lin/	IBM Communications Server for Linux on System z
http://www.ibm.com/software/network/ccl/	IBM Communication Controller for Linux on System z
http://www.ibm.com/software/network/commserver/library/	IBM Communications Server library
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

For pleasant reading ….

# *z/OS V1R10 added subplex support to the z/OS load balancing advisor*
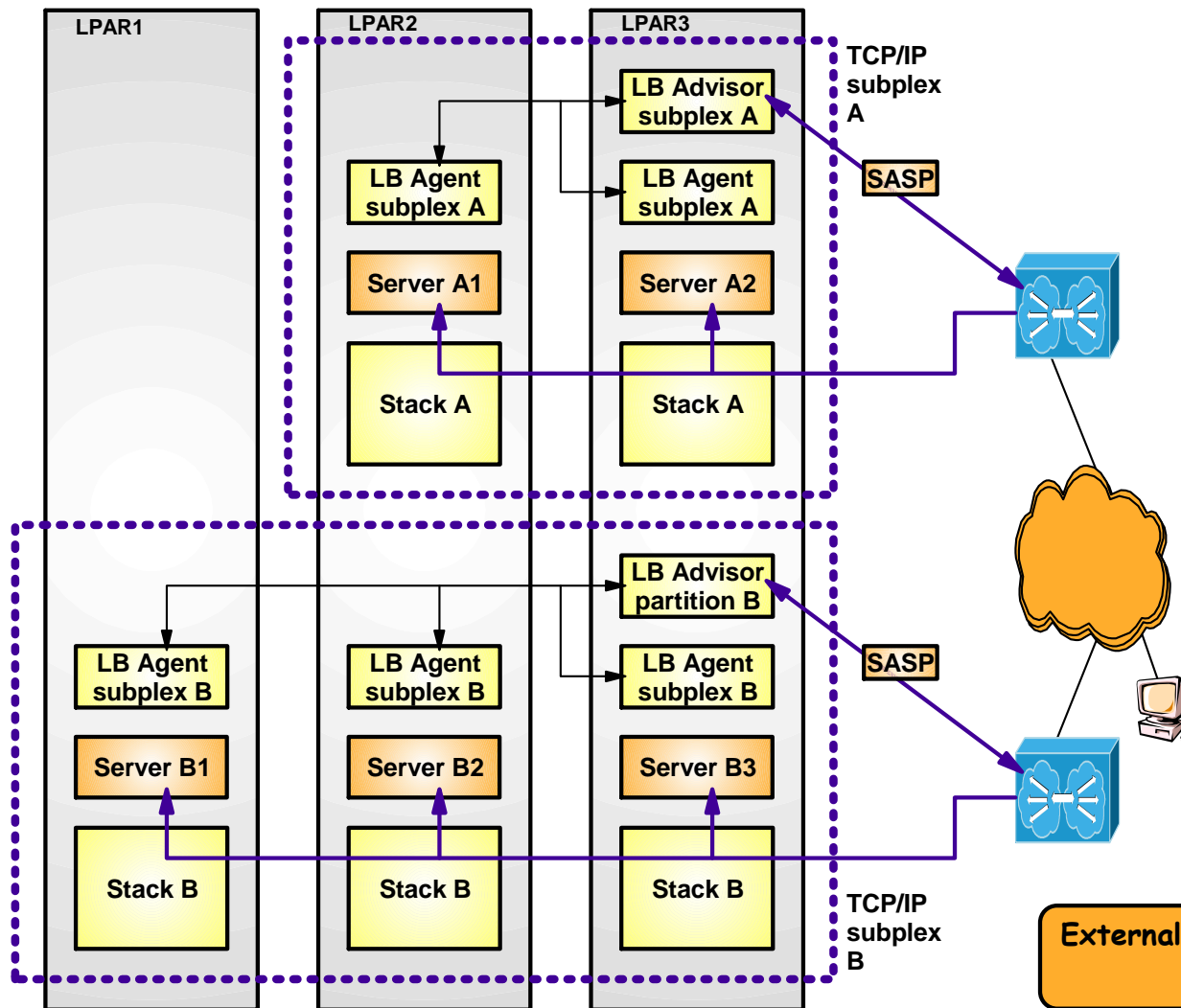
➢ **One advisor address space per TCP/IP subplex**
  ► Explicit configuration option

➢ **One agent per LPAR per TCP/IP subplex**
  ► Explicitly configured

➢ **Some configuration guidance will be required:**
  ► Only specify server IP addresses and ports that belong to the subplex for which the LBA advisor/agent belongs

External load balancing to servers in a z/OS TCP/IP subplex
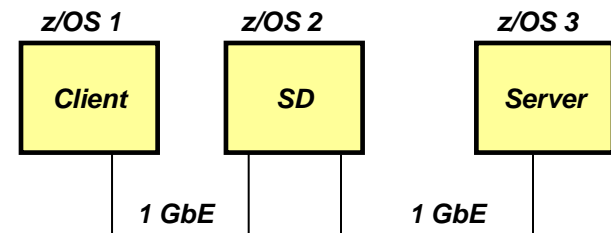
Sysplex Distributor accelerator performance

✓ *Intended to benefit all existing Sysplex Distributor users*
✓ *Measurements with Interactive workload*
✓ *Small data sizes (100 in, 800 out)*
✓ *Percentages relative to no acceleration*

Legend: **CPU SD** **CPU server**
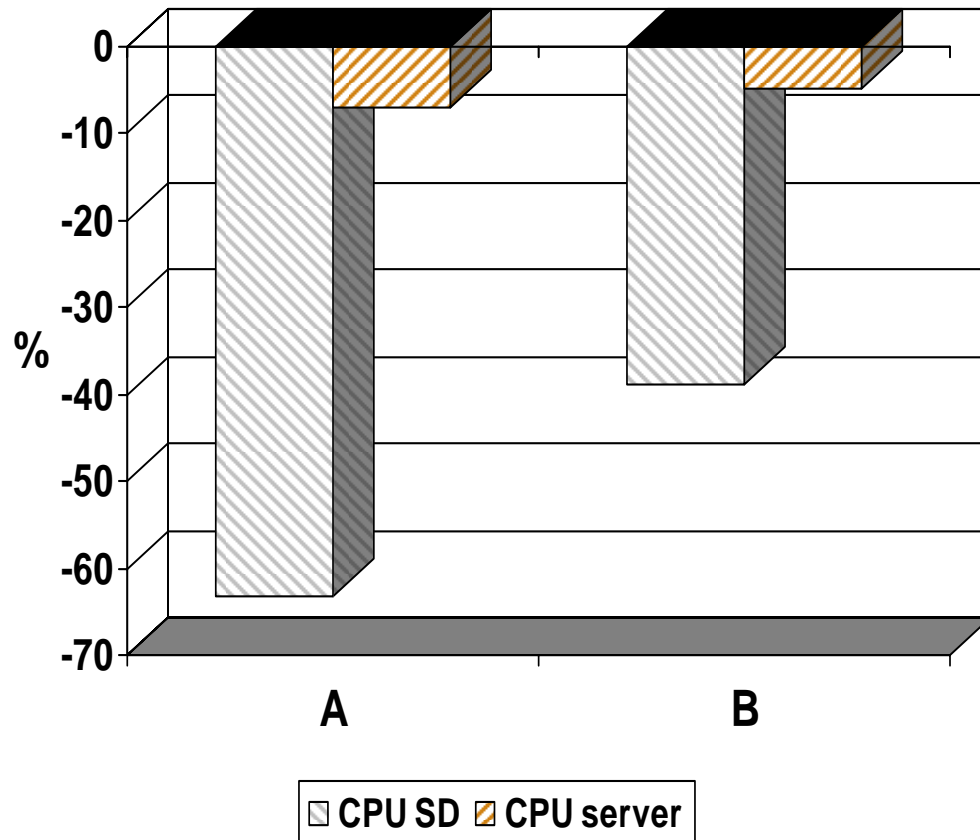
Configuration A – Three z10 LPARs with OSA Express 3 cards and HiperSockets between SD and server LPARs

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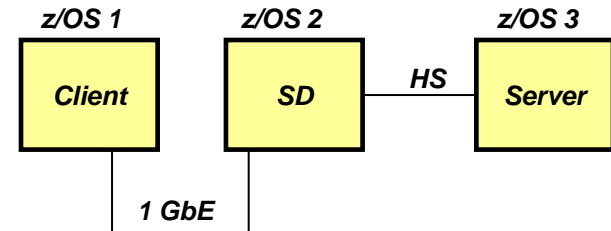
Configuration B – Three z10 LPARs with OSA Express 3 cards

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

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.