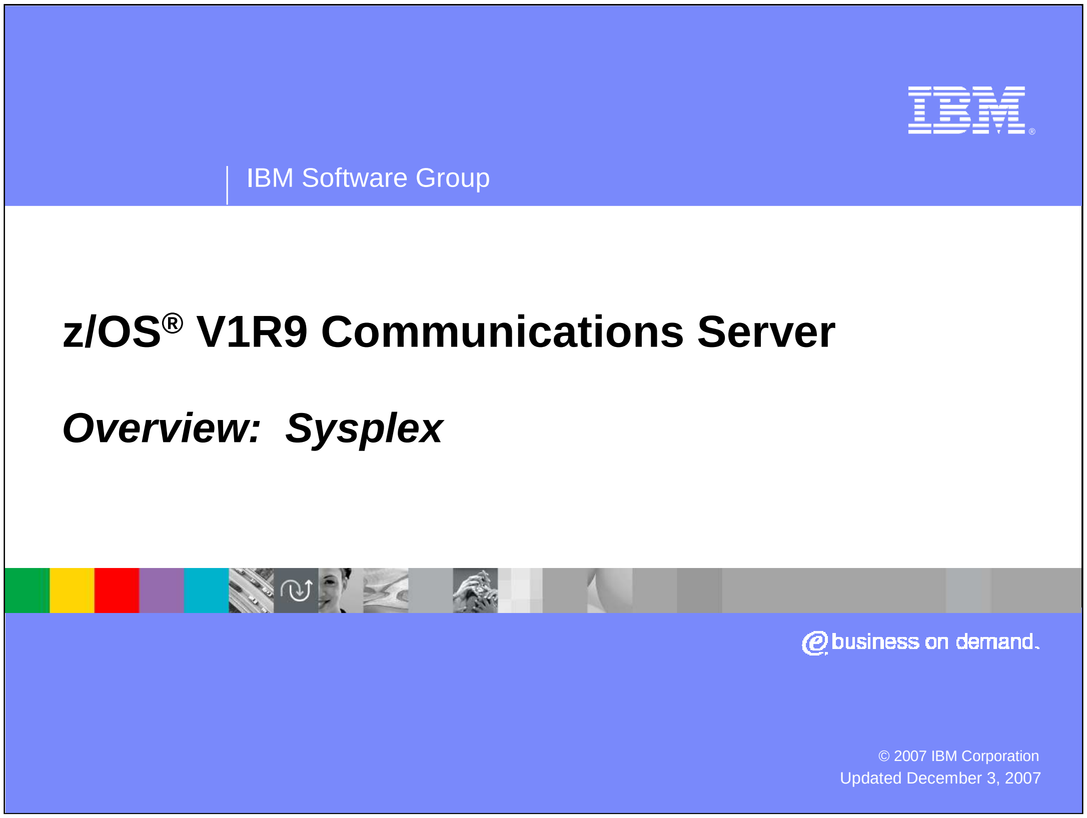


IBM Software Group

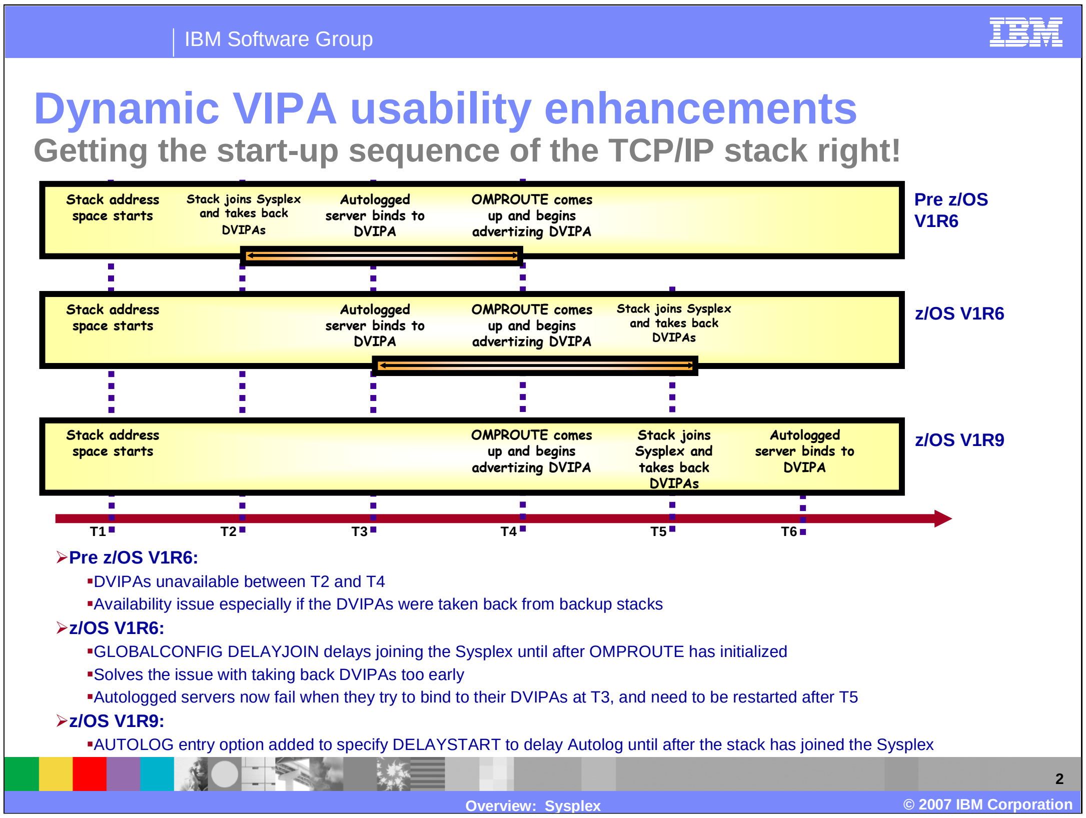
z/OS® V1R9 Communications Server

Overview: Sysplex

business on demand.

Updated December 3, 2007

This presentation is an overview of the Sysplex enhancements for z/OS V1R9 Communications Server.



IBM Software Group

Dynamic VIPA usability enhancements

Getting the start-up sequence of the TCP/IP stack right!

Pre z/OS V1R6

Stack address space starts

Stack joins Sysplex and takes back DVIPAs

Autologged server binds to DVIPA

OMPROUTE comes up and begins advertizing DVIPA

z/OS V1R6

Stack address space starts

Autologged server binds to DVIPA

OMPROUTE comes up and begins advertizing DVIPA

Stack joins Sysplex and takes back DVIPAs

z/OS V1R9

Stack address space starts

OMPROUTE comes up and begins advertizing DVIPA

Stack joins Sysplex and takes back DVIPAs

Autologged server binds to DVIPA

T1 T2 T3 T4 T5 T6

- **Pre z/OS V1R6:**
  - DVIPAs unavailable between T2 and T4
  - Availability issue especially if the DVIPAs were taken back from backup stacks
- **z/OS V1R6:**
  - GLOBALCONFIG DELAYJOIN delays joining the Sysplex until after OMPROUTE has initialized
  - Solves the issue with taking back DVIPAs too early
  - Autologged servers now fail when they try to bind to their DVIPAs at T3, and need to be restarted after T5
- **z/OS V1R9:**
  - AUTOLOG entry option added to specify DELAYSTART to delay Autolog until after the stack has joined the Sysplex

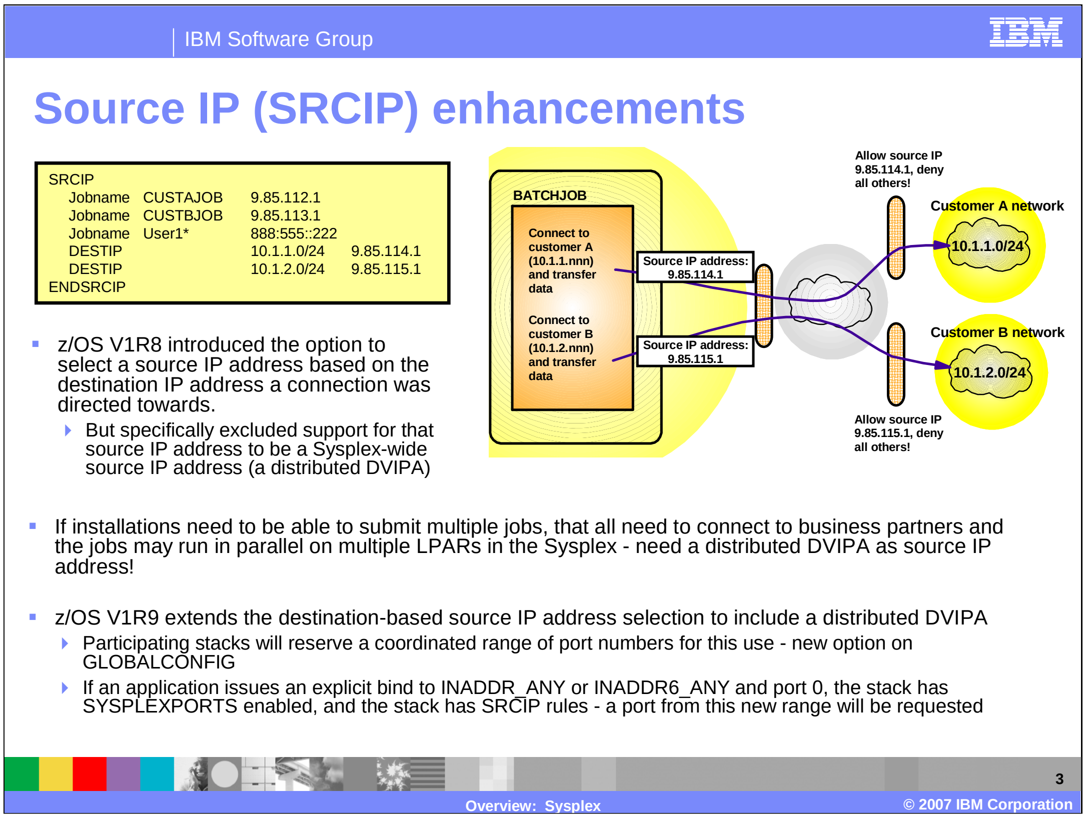
2

Overview: Sysplex

© 2007 IBM Corporation

AUTOLOG profile statement specifies procedures to be automatically started after TCP/IP is started, and monitored at regular intervals. DELAYJOIN is another configuration parameter. It is specified on the GLOBALCONFIG profile statement. When DELAYJOIN is specified, TCP/IP will not join the sysplex group until OMPROUTE is active . Since the stack's dynamic VIPA configuration is not processed until after the stack has joined the sysplex group, this delay in joining the sysplex group ensures that OMPROUTE will be active and ready to advertise dynamic VIPAs when they are created on this stack. OMPROUTE and the other AUTOLOGed procedures will be started at the same time. AUTOLOGed procedures that bind to dynamic VIPAs may fail due to the delay in joining the sysplex and creating these DVIPAs.

In z/OS V1R9 Communications Server, the starting of applications that bind to a dynamic VIPA can be delayed until TCP/IP has joined the sysplex and created DVIPAs.



IBM Software Group

Source IP (SRCIP) enhancements

```
SRCIP
   Jobname  CUSTAJOB      9.85.112.1
   Jobname  CUSTBJOB      9.85.113.1
   Jobname  User1*        888:555::222
   DESTIP                 10.1.1.0/24     9.85.114.1
   DESTIP                 10.1.2.0/24     9.85.115.1
ENDSRCIP
```

BATCHJOB
Connect to customer A (10.1.1.nnn) and transfer data
Connect to customer B (10.1.2.nnn) and transfer data
Source IP address: 9.85.114.1
Source IP address: 9.85.115.1
Allow source IP 9.85.114.1, deny all others!
Customer A network
10.1.1.0/24
Customer B network
10.1.2.0/24
Allow source IP 9.85.115.1, deny all others!

- z/OS V1R8 introduced the option to select a source IP address based on the destination IP address a connection was directed towards.
  - But specifically excluded support for that source IP address to be a Sysplex-wide source IP address (a distributed DVIPA)
- If installations need to be able to submit multiple jobs, that all need to connect to business partners and the jobs may run in parallel on multiple LPARs in the Sysplex - need a distributed DVIPA as source IP address!
- z/OS V1R9 extends the destination-based source IP address selection to include a distributed DVIPA
  - Participating stacks will reserve a coordinated range of port numbers for this use - new option on GLOBALCONFIG
  - If an application issues an explicit bind to INADDR_ANY or INADDR6_ANY and port 0, the stack has SYSPLEXPORTS enabled, and the stack has SRCIP rules - a port from this new range will be requested

3

Overview: Sysplex

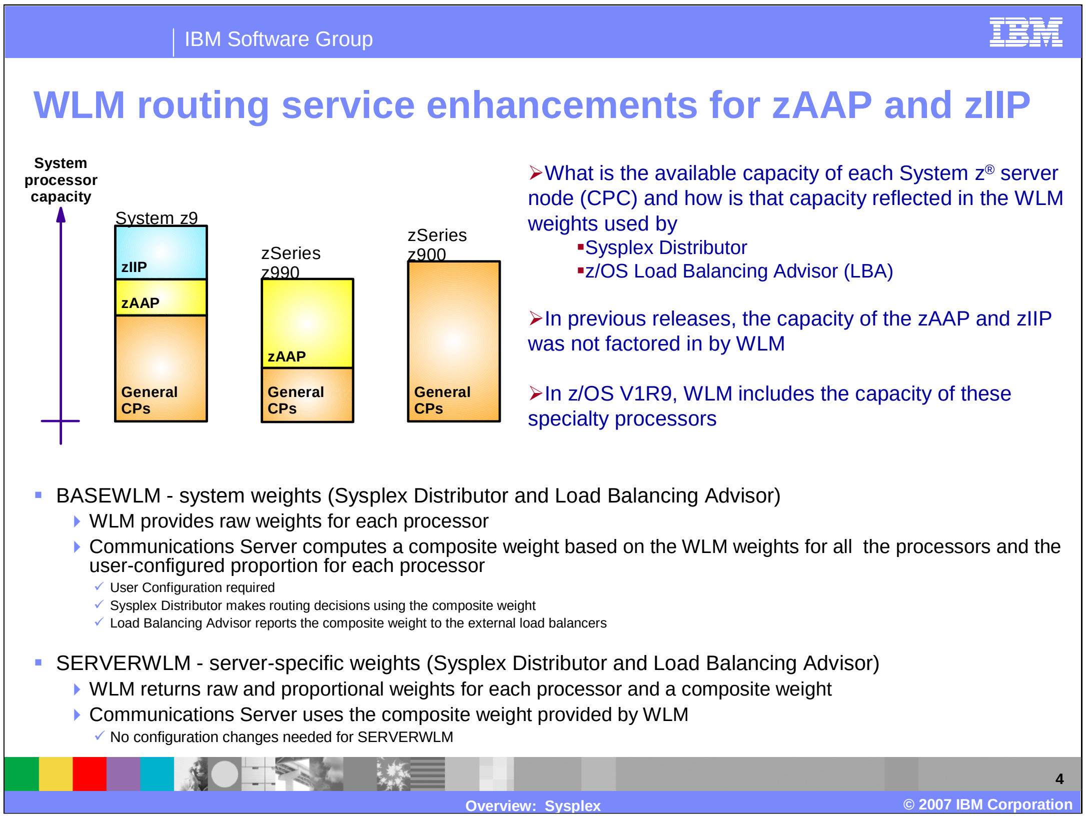
© 2007 IBM Corporation

In z/OS V1R8 Communications Server there is a restriction on the type of IP address that may be specified on a SRCIP block DESTINATION rule. Specifically, it could not be a distributed DVIPA. Because distributed DVIPAs may be active on many stacks in the sysplex, source ports that are allocated for connections from these DVIPAs must be coordinated across the sysplex. If they were not, an application on node A in the sysplex, connecting to a destination IP address and port using a specific distributed DVIPA might choose the same port as an application on node B, using the same distributed DVIPA to connect to the same destination IP address and port. This would result in two connection requests with the same 4-tuple (of source IP address, source port, destination IP address, destination port) being sent to the same destination. To prevent this, allocation of source ports (known as sysplexports) for distributed DVIPAs is coordinated using the EZBEPORTvvtt structure, which establishes an allocated source port pool for each specific distributed DVIPA.

A problem also occurs when an application uses an explicit BIND to INADDR_ANY or IN6ADDR_ANY and port 0 before issuing a CONNECT. The BIND protocols require that a port be assigned at this time. However, since the CONNECT has not yet been issued, TCP/IP does not know the destination address, so it would not know to allocate the port from the sysplex port pool associated with the matching DESTINATION rule's source IP address, if that source IP address were a distributed DVIPA.

A range of ephemeral ports can be designated that will be assigned uniquely across the sysplex. If an application explicitly binds to INADDR_ANY or IN6ADDR_ANY and port 0, a port from this range will be assigned. This ensures that when the TCP connection is later established if a distributed DVIPA (DDVIPA) is chosen as the source address because of a SRCIP DESTINATION rule, that the source port/DDVIPA combination will be unique throughout the sysplex. This removes the z/OS V1R8 Communications Server restriction which disallowed configuring a distributed DVIPA to be used as a source address on a DESTINATION statement in the SRCIP statement block.

This is intended for use in conjunction with SYSPLEXPORTS DDVIPAs. However, these ports are not associated with a specific DVIPA, but instead managed by the Coupling Facility and TCP/IP stacks to ensure that they are unique across the sysplex for all IP addresses.


System processor capacity
System z9
zIIP
zAAP
General CPs
zSeries z990
zAAP
General CPs
zSeries z900
General CPs

IBM Software Group

WLM routing service enhancements for zAAP and zIIP

➢What is the available capacity of each System z® server node (CPC) and how is that capacity reflected in the WLM weights used by

- Sysplex Distributor
- z/OS Load Balancing Advisor (LBA)

➢In previous releases, the capacity of the zAAP and zIIP was not factored in by WLM

➢In z/OS V1R9, WLM includes the capacity of these specialty processors

- BASEWLM - system weights (Sysplex Distributor and Load Balancing Advisor)
  - WLM provides raw weights for each processor
  - Communications Server computes a composite weight based on the WLM weights for all the processors and the user-configured proportion for each processor
    - ✓ User Configuration required
    - ✓ Sysplex Distributor makes routing decisions using the composite weight
    - ✓ Load Balancing Advisor reports the composite weight to the external load balancers
- SERVERWLM - server-specific weights (Sysplex Distributor and Load Balancing Advisor)
  - WLM returns raw and proportional weights for each processor and a composite weight
  - Communications Server uses the composite weight provided by WLM
    - ✓ No configuration changes needed for SERVERWLM

4

Overview: Sysplex

© 2007 IBM Corporation

The zSeries platform has specialty processors that can be deployed and exploited by targeted workloads on z/OS. This includes support for:

zAAP (zSeries Application Assist Processor) - these processors can be used for JAVA application workloads on z/OS (including workloads running under WAS).

zIIP (System z Integrated Information Processor) – they can be used for z/OS DB2 related workloads, such as z/OS DB2 workload initiated over the network (that is, using the DRDA protocol) and DB2 BI (Business Intelligence) workloads (that is. complex queries). It can also be used for IPSec workload.

The Sysplex Distributor and the z/OS Load Balancing Advisor (LBA) supports two types of distribution using WLM recommendations. Before z/OS V1R9 Communications Server, the capacity of specialty processors, such as zAAP and zIIP, was not factored in by WLM when calculating the weights that are used by the CS functions.

BaseWLM weights are based on a comparison of target Systems in the sysplex.

How much CP capacity is available on each system?

When all systems in the sysplex are running at or near 100% utilization, WLM will assign the higher weights to the systems with the largest amounts of lower importance work (systems with the most displaceable capacity).

ServerWLM weights are based on a comparison of target servers within the same service class

How well is a server meeting the goals of its service class?

How much displaceable capacity is available on this system for new work based on the importance of this service class?

When ServerWLM is being used WLM will return server-specific weights for each processor.

•Processor’s Raw weights - zIIP, zAAP, and CP weights

•Processor’s Proportional weights - raw processor weights modified by actual usage by this server (that is. Raw processor weight * Server usage proportion)

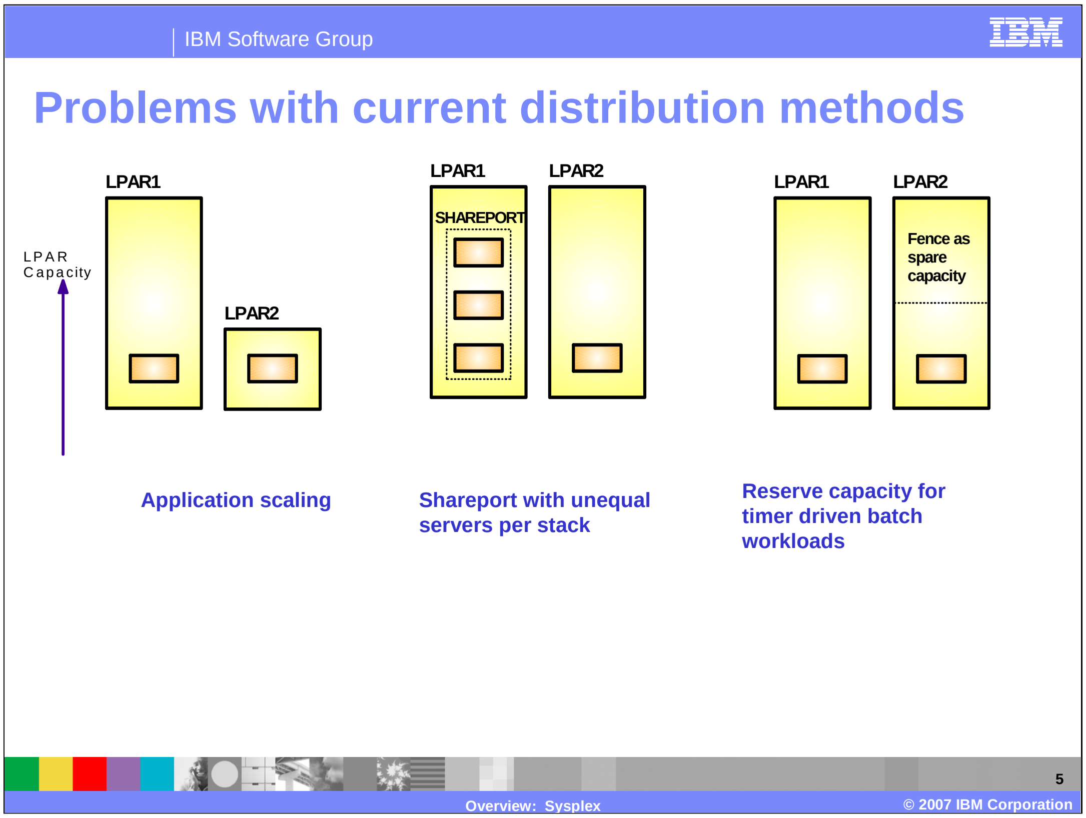
•Composite weight - Sum of all the processor’s proportional weights.

This composite weight returned by WLM is used by sysplex distributor and LBA.

When BaseWLM is used WLM will return system weights for each processor.

•Raw processor weights - zIIP, zAAP, and CP

Communications Server determines a processor’s proportional fraction by dividing the configured proportion for a processor by the sum of all configured processor proportions. It then determines each processor’s proportional weight using the raw weight received from WLM (that is. Processor Raw weight * Processor Proportional fraction). Communications Server computes the Composite weight, which is the sum of all the processor’s proportional weight, and that is used by sysplex distributor and LBA.



IBM Software Group

Problems with current distribution methods

Application scaling

Shareport with unequal servers per stack

Reserve capacity for timer driven batch workloads

5

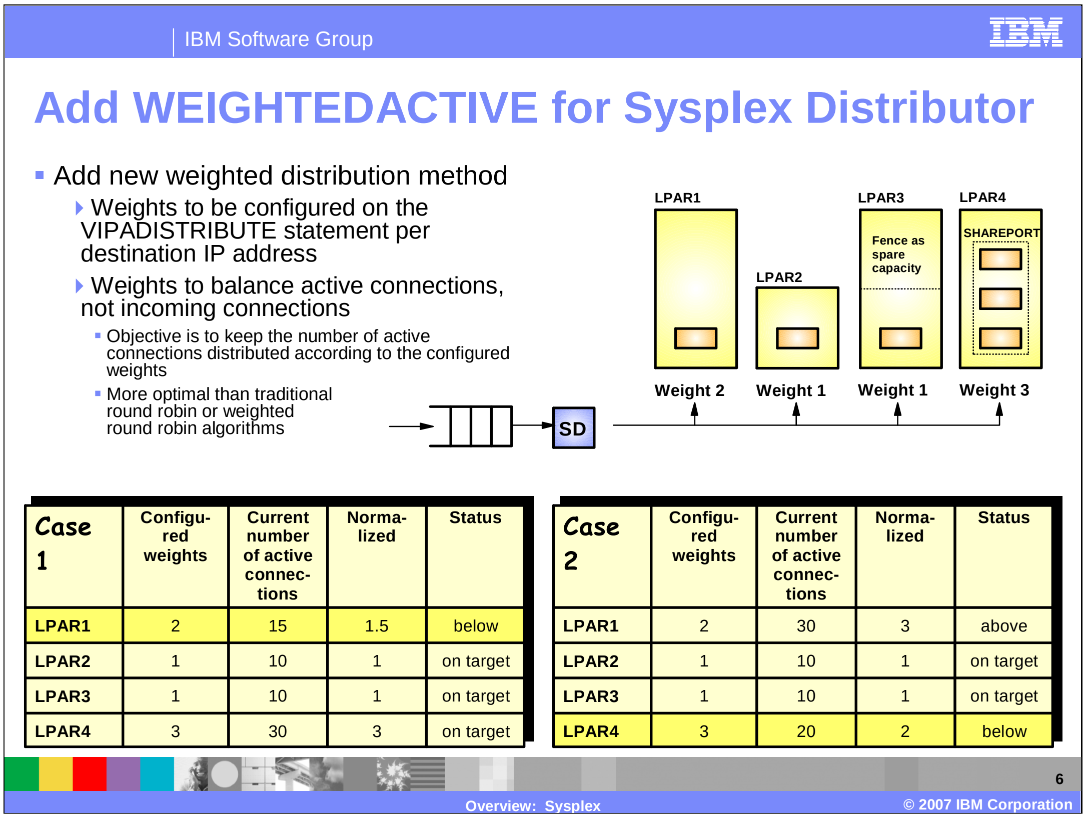
Overview: Sysplex

© 2007 IBM Corporation

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 partially addresses this by reducing the recommendation if the application’s Performance goals are not met.

SHAREPORT is deployed, yet not all systems have the same number of SHAREPORT server instances (one has three the other has only one). The current RR or WLM recommendations do not change distribution based on the number of server instances on each target. RR 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).



IBM Software Group

Add WEIGHTEDACTIVE for Sysplex Distributor

- Add new weighted distribution method
  - Weights to be configured on the VIPADISTRIBUTE statement per destination IP address
  - Weights to balance active connections, not 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

Case 1	Configured weights	Current number of active connections	Normalized	Status
LPAR1	2	15	1.5	below
LPAR2	1	10	1	on target
LPAR3	1	10	1	on target
LPAR4	3	30	3	on target

Case 2	Configured weights	Current number of active connections	Normalized	Status
LPAR1	2	30	3	above
LPAR2	1	10	1	on target
LPAR3	1	10	1	on target
LPAR4	3	20	2	below

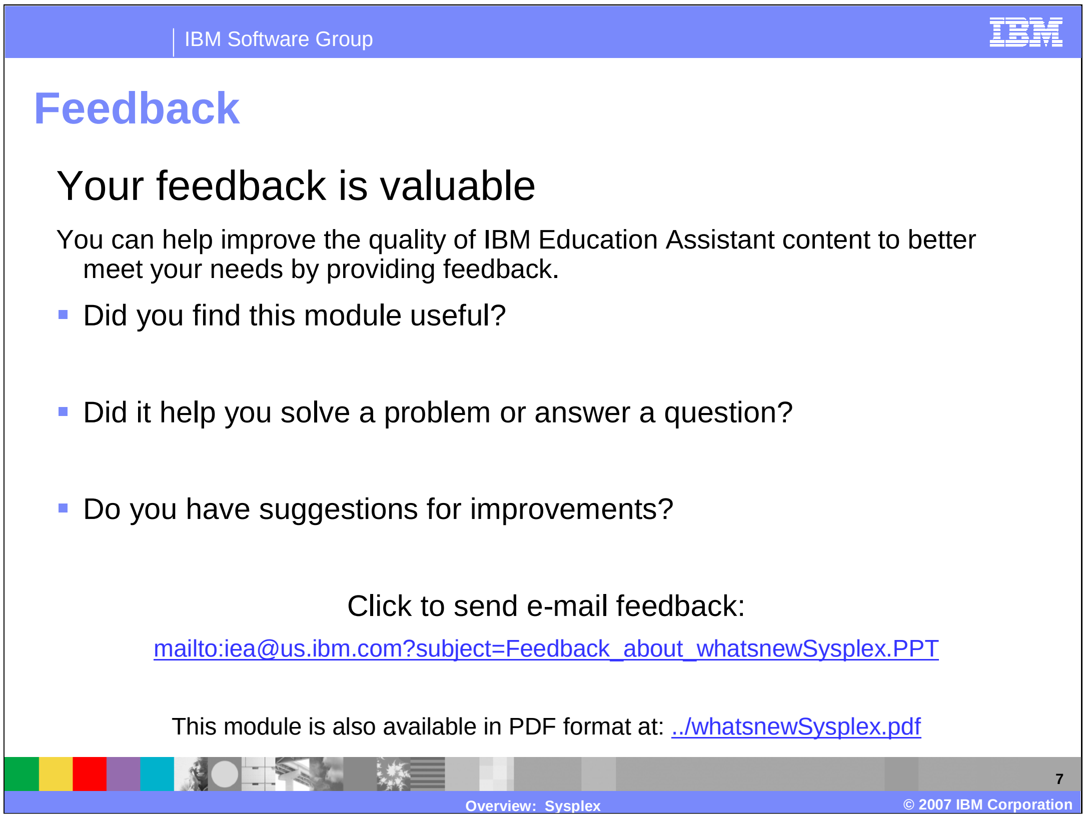
6

Overview: Sysplex

© 2007 IBM Corporation

A new distribution method, Weighted Active Connections is supported. This distribution method (WEIGHTEDACTIVE), provides granular control over workload distribution based on predetermined active connection count proportions for each target (fixed weights). If this distribution method is being used, then for each target TCP/IP stack, an active connection weight may be configured. The distributor balances incoming connection requests across the targets with a goal of having the number of active connections on each target proportionally equivalent to the configured active connection weight of each target. The connection weight defaults to one, so by not specifying any connection weights, the distributor's goal will be to have an equal number of active connections on each target.

The Target Server Responsiveness (TSR) fraction, abnormal completion rate fraction, and General Health fraction are applied against the configured weight to determine a modified weight. Connection goals are established based on the modified weight and the active connection count. Normalized weights are established by dividing the modified weight by 10.



IBM Software Group

Feedback

Your feedback is valuable

You can help improve the quality of IBM Education Assistant content to better meet your needs by providing feedback.

- Did you find this module useful?
- Did it help you solve a problem or answer a question?
- Do you have suggestions for improvements?

Click to send e-mail feedback:

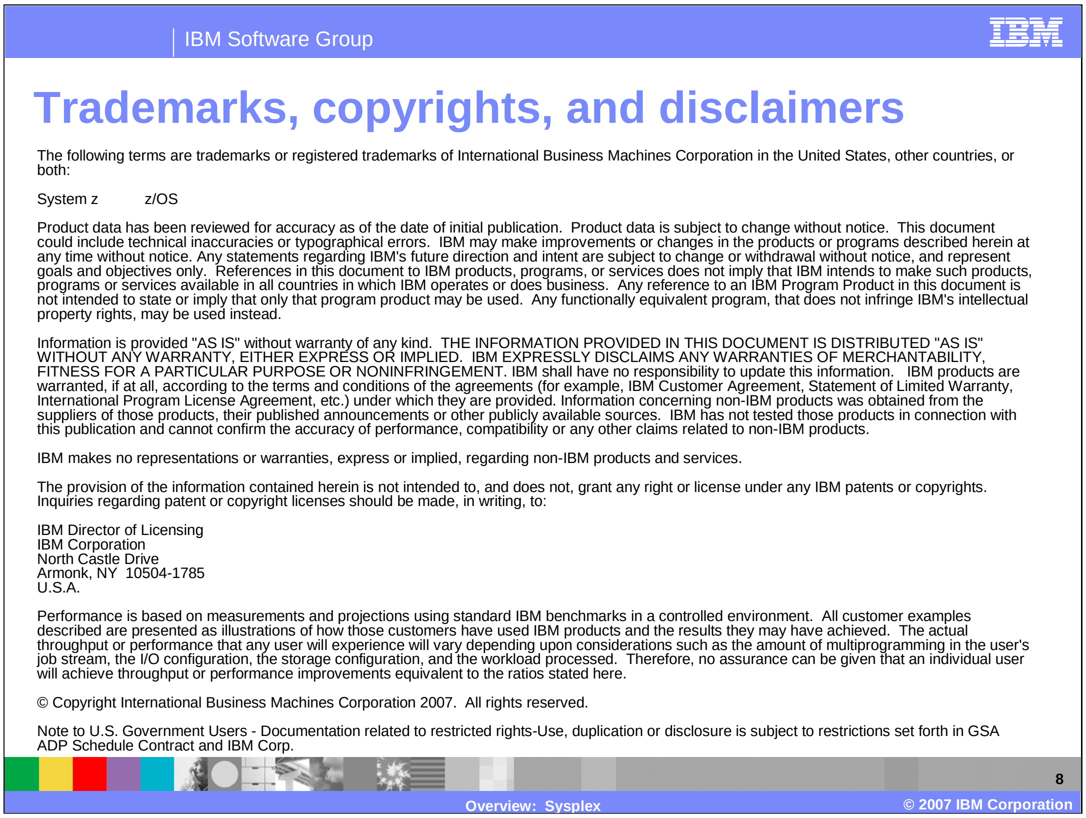
mailto:iea@us.ibm.com?subject=Feedback_about_whatsnewSysplex.PPT

This module is also available in PDF format at: ../whatsnewSysplex.pdf

7

Overview: Sysplex

You can help improve the quality of IBM Education Assistant content by providing feedback.



Trademarks, copyrights, and disclaimers

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

System z z/OS

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This document could include technical inaccuracies or typographical errors. IBM may make improvements or changes in the products or programs described herein at any time without notice. Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business. Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program, that does not infringe IBM's intellectual property rights, may be used instead.

Information is provided "AS IS" without warranty of any kind. THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IBM EXPRESSLY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT. IBM shall have no responsibility to update this information. IBM products are warranted, if at all, according to the terms and conditions of the agreements (for example, IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products.

IBM makes no representations or warranties, express or implied, regarding non-IBM products and services.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. The actual throughput or performance 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 or performance improvements equivalent to the ratios stated here.

Note to U.S. Government Users - Documentation related to restricted rights-Use, duplication or disclosure is subject to restrictions set forth in GSA ADP Schedule Contract and IBM Corp.