



IBM Software Group

z/OS® V1R9 Communications Server

Sysplex load balancing



@business on demand.

© 2008 IBM Corporation
Updated January 10, 2008

This presentation discusses the Sysplex load balancing enhancements in z/OS V1R9 Communications Server.

Agenda

- Support for WLM routing service enhancements for zAAP and zIIP
- Add WEIGHTEDACTIVE distribution method for Sysplex Distributor
- Support to configure the WLM Polling Interval



Sysplex Distributor and the Load Balancing Advisor use WLM information about the specialty processors, zAAP and zIIP, for workload balancing.

Support for a new distribution method WEIGHTEDACTIVE is added in this release.

As part of an APAR recently added in z/OS V1R6 Communications Server, configuration of a WLM polling interval is also supported.

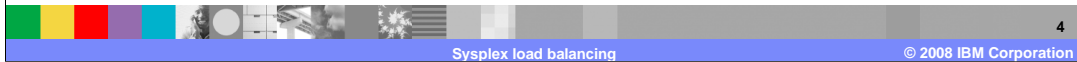
**Support for WLM routing service
enhancement for zAAP and zIIP**

This section describes the Sysplex Distributor and Load Balancing Advisor enhancements added to support the specialty processors.

Background - Distribution using WLM weights

- The Sysplex Distributor and Load Balancing Advisor support two types of distribution using WLM recommendations:
 - ▶ WLM System weights – based on a comparison of conventional CP capacity (BASEWLM)

 - ▶ WLM Server-Specific weights – based on a comparison of (SERVERWLM)
 - ✓ The CP capacity given the importance of the server's work
 - ✓ How well each server is meeting the goals of its service class



System weights (BaseWLM) and Server-specific weights (ServerWLM) are relative weights that range in value between 0 and 64.

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).

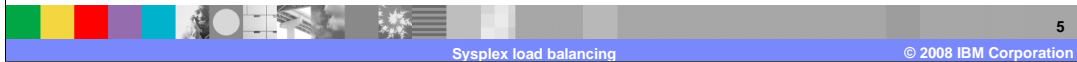
ServerWLM weights are based on a comparison of target servers within the same service class

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?

Problem - Support needed for specialty processors

- The zSeries® platform recently introduced “specialty” processors that are designed for specific z/OS workloads:
 - ▶ zAAP (zSeries Application Assist Processor)
 - ▶ zIIP (System z™ Integrated Information Processor)
- These new processors need to be considered when determining target weights



The zSeries platform has recently introduced the concept of 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).

- z/OS IPSEC workloads

Solution - Processors considered for target weights

- When ServerWLM is being used:
 - ▶ For each processor, WLM will return server-specific weights
 - ✓ Raw processor weights - zIIP, zAAP, and CP weights
 - ✓ Proportional weights - raw weights modified by actual usage by this server
 - ✓ Composite weight - based on the proportional weights
 - ▶ Sysplex distributor & LBA will display these weights
 - ▶ Sysplex distributor will make routing decisions using the composite weight
 - ▶ LBA will report the composite weights to external load balancers in place of the conventional CP weight
- When BaseWLM is used:
 - ▶ For each processor, WLM will return system weights
 - ✓ Raw processor weights - zIIP, zAAP, and CP
 - ▶ Sysplex distributor & LBA
 - ✓ Display the raw processor weights returned by WLM
 - ✓ Allow configuration of expected processor usage proportions
 - PROCTYPE CP x zAAP y zIIP z
 - ✓ Determine and display the proportional zIIP, zAAP, and CP weights.
 - ✓ Determine and display a composite weight from the proportional weights
 - ▶ Sysplex distributor will make routing decisions using the composite weight
 - ▶ LBA will report the composite weights to external load balancers in place of the conventional CP weight

6

This slide provides an overview of the changes in server-specific weights and system weights support for WLM, Sysplex distributor, and LBA.

For server-specific weights zIIP, zAAP, and CP weights are based on how each server is meeting the goals of its service class and a comparison of that processor's (available or displaceable) capacity on each target system given the importance of the server's work. For each processor, WLM will return a composite weight that Sysplex Distributor will use when making routing decisions and the Load Balancing Advisor will report to external load balancers. No additional configuration is required when ServerWLM is being used.

For system weights, zIIP, zAAP, and CP weights are based on the system-level displaceable capacity of each processor type. Because WLM is unaware of how applications are utilizing the various processors, some configuration may be required when BaseWLM is used. In these cases, it will be up to you to indicate the proportion of each type of processor those workloads will consume. The Communication Server workload distribution technologies retrieve the system WLM raw weights of each type of processor and apply the configured proportions to arrive at the composite weight to be used for workload distribution.

For the Sysplex Distributor, the proctype parameter on the VIPADISTRIBUTE statement can be configured to indicate the expected processor usage for each processor when the distribution method is BaseWLM. For the LBA, proctype can be defined on the WLM statement or the port_list statement to indicate the expected processor usage when baseWLM is being used. When the wlm parameter is not configured on the port_list, it defaults to the WLM statement configuration.

Values for each processor type, specified on proctype, can range between 0 and 99 so that the proportions can be expressed as percentages if required. The default for proctype is to only consider convention CP weights and not consider zIIP or zAAP when determining a weight. When proctype is coded, at least one processor type must be specified; any processor types that are not specified will be assigned a value of 0. Users should evaluate whether SERVERWLM distribution could be used as an alternative to BASEWLM distribution for their application. SERVERWLM has the added advantage that processor proportions will be automatically determined and dynamically updated by WLM based on the actual processor usage by the application. If BASEWLM is needed, to determine the processor proportions to configure, users need to study their workload usage of processors by analyzing SMF records, and performance monitor reports, such as RMF Workload Activity Reports to determine the expected utilization proportion for each processor type.

Configuration examples

```
VIPADISTRIBUTE BASEWLM PROCTYPE CP 20 ZAAP 80
                201.2.10.11 PORT 8000
                DESTIP ALL
VIPADISTRIBUTE SERVERWLM
                201.2.10.13 PORT 9000
                DESTIP ALL
```

```
Wlm basewlm
{
  Proctype
  {
    CP 30
    zAAP 70
  }
}

Port_list
{
  8000
  {
    WLM basewlm
    {
      Proctype
      {
        CP 20
        zAAP 80
      }
    }
  }
  9000
}
```

This slide shows an example of how to define processor type proportions, on the VIPADISTRIBUTE statement, when distributing to Port 8000 so that 20% of CP capacity, 80% of zAAP capacity, and 0% of zIIP capacity should be considered when determining the composite System weight. There are no configuration changes needed when the distribution method being used is ServerWLM.

This slide also shows how to define the processor type proportions for the Load Balancing Advisor (LBA). Values for each processor type can range between 0 and 99 so that the proportions can be expressed as percentages if required. When a processor type subparameter is not specified on the proctype parameter, it is set to 0. Proctype can be defined on the WLM statement or the port_list statement. When the wlm parameter is not configured for a port, the wlm settings default to the WLM statement configuration. Each statement and brace must be on a separate line, and each parameter within a bracket must also be on a separate line. In the example port 8000 has processor type proportions of CP 20% and zAAP 80%. Since no proctype is configured for port 9000, it defaults to the wlm statement settings to use basewlm weights with proportions of CP 30% and zAAP 70%.

Display command example - VIPADCFG

- **Netstat VIPADCFG/-F Detail Changes**

```
NETSTAT VIPADCFG DETAIL
VIPA Distribute:
  Dest:      201.2.10.11..8000
  DestXCF:   ALL
  SysPt:    No  TimAff: No  Flg: BaseWLM
  OptLoc:   No
  ProcType:
    CP: 20  zAAP: 80  zIIP: 00

  Dest:      201.2.10.13..9000
  DestXCF:   ALL
  SysPt:    No  TimAff: No  Flg: ServerWLM
  OptLoc:   No
```

The VIPADCFG Detail display is modified to display the configured processor proportions for each target when the distribution method is BaseWLM. The proportions will be used to modify the raw weights received from WLM. There are no changes to the display when ServerWLM is being used.

Display command example - VDPT detail (BaseWLM)

Netstat VDPT/-O DETAIL changes

The screenshot shows two entries in the Netstat VDPT detail report. The first entry is for destination 201.2.10.11..8000 (BaseWLM) and the second is for 201.2.10.13..9000 (ServerWLM). Callouts explain the weight calculations for the BaseWLM entry:

- Normalized weight:** $13 = 54/4$
- Composite weight:** $54 = CP: 6 + zAAP: 48$
- Raw weights:** CP: 30, zAAP: 60, zIIP: 60
- Proportional weights:** CP: 6, zAAP: 48, zIIP: 00. These are determined from ProcType: CP: 20, zAAP: 80, zIIP: 00. For example, $CP: 6 = 30 * 20\%$.

```

NETSTAT VDPT DETAIL
Dynamic VIPA Destination Port Table:
Dest: 201.2.10.11..8000
DestXCF: 201.3.10.15
TotalConn: 0000084011 Rdy: 001 WLM: 13 TSR: 100
Flg: BaseWLM
TCSR: 100 CER: 100 SEF: 100
Weight: 54
Raw CP: 30 zAAP: 60 zIIP: 60
Proportional CP: 6 zAAP: 48 zIIP: 00
ActConn: 0000000201
QosPlcAct: *DEFAULT*
W/Q: 00
Dest: 201.2.10.13..9000
DestXCF: 201.3.10.16
TotalConn: 0000020340 Rdy: 001 WLM: 10 TSR: 100
Flg: ServerWLM
TCSR: 100 CER: 100 SEF: 100
Weight: 40
Raw CP: 40 zAAP: 40 zIIP: 60
Proportional CP: 4 zAAP: 36 zIIP: 00
ActConn: 0000000058
QosPlcAct: *DEFAULT*
W/Q: 00
  
```

Use the Netstat VDPT/-O DETAIL report to display the raw weights, proportionally modified weights, raw composite weight and composite weight after normalization when BaseWLM or ServerWLM is being used.

With z/OS V1R9 Communications Server some of the detailed displays are simplified to only show values that pertain to a distribution method.

This example shows how the weights are determined for BaseWLM given the Proctype configuration on the previous slide of CP 20 zAAP 80.

Looking at the port 8000 BaseWLM target, the WLM weight of 13 is determined as follows. The raw weights, (CP 30, zAAP 60, and zIIP 60), were received from WLM. Each raw weight ranges from 0 through 64. The configured proportions are CP 20 zAAP 80. A processor's proportional fraction is determined by dividing the configured proportion by the sum of all configured processor proportions: CP proportional fraction 20% = $CP\ 20 / (CP\ 20 + zAAP\ 80 + zIIP\ 0)$ and zAAP proportional fraction 80% = $zAAP\ 80 / (CP\ 20 + zAAP\ 80 + zIIP\ 0)$. Each proportional weight is determined using the proportional fraction against the raw weight received from WLM: $CP\ 6 = (Raw\ CP\ 30) * (CP\ Proportional\ fraction\ 20\%)$ and $zAAP\ 48 = (Raw\ zAAP\ 60) * (zAAP\ Proportional\ fraction\ 80\%)$. The composite raw weight is the sum of the proportional weights (Weight 54 = CP 6 + zAAP 48). The TSR fraction is applied against the composite weight (no change since TSR fraction is 100%). The Normalized weight is determined by dividing the TSR modified weight by 4 (WLM 13 = $54/4$).

Looking at the port 9000 ServerWLM target, the weight of 10 is determined as follows. The following raw weights, (CP 40, zAAP 40, and zIIP 60), were received from WLM. The proportional weights, (CP 4, zAAP 36, and zIIP 0), received from WLM are based on Sysplex load balancing by the application. The composite weight is the sum of raw weights (Weight 40 = CP 40 + zAAP 40 + zIIP 0).

Display command example - LBA details (BaseWLM)

▪ MODIFY command—z/OS Load Balancing Advisor

```

F LBADV,DISP,LB,I=0
EZD1243I LOAD BALANCER DETAILS
LB INDEX      : 00          UUID       : 637FFF175C
GROUP NAME    : CICS_SYSTEM_FARM
GROUP FLAGS   : BASEWLM
ProcType      :
  CP: 20  zAAP: 80  zIIP: 00
IPADDR..PORT: 201.2.10.11.8000
SYSTEM NAME:  MVS209  PROTOCOL : TCP  AVAIL   : YES
WLM WEIGHT : 00054   CS WEIGHT : 100  NET WEIGHT: 00001
Raw          CP: 30  zAAP: 60  zIIP: 60
Proportional CP: 06  zAAP: 48  zIIP: 00
FLAGS       :
...
  
```

Composite weight
54 = CP: 6 + zAAP 48

Normalized weight

Raw weights

BaseWLM Proportional weights are determined from ProcType:
CP: 20 zAAP: 80 zIIP: 00
e.g. CP: 6 = 30 * 20%

10

Sysplex load balancing

© 2008 IBM Corporation

The load balancing advisor detail report shows the proctype proportions configured when BaseWLM is being used (CP: 20 zAAP: 80 zIIP:00), the Raw weights received from WLM (CP: 30 zAAP: 60 zIIP:60), and the proportional weights (CP: 06 zAAP: 48 zIIP: 00). When BaseWLM is being used, the proportional weights are modified by the advisor based on the configured proctype proportions.

LBA determines a normalized weight by dividing by the highest common denominator of all WLM weights received for a Port and Protocol. In this case with only one WLM weight (54), the highest common denominator is 54. $54/54 = 1$.

IBM Software Group IBM

Display command example - LBA details (ServerWLM)

- MODIFY command—z/OS Load Balancing Advisor

```

F LBADV,DISP,LB,I=0
EZD1243I LOAD BALANCER DETAILS
LB INDEX      : 00      UUID      : 637FFF175C
...
GROUP NAME    : CICS_SYSTEM_FARM
GROUP FLAGS   : SERVERWLM
IPADDR..PORT : 201.2.10.13..9000
SYSTEM NAME   : MVS209   PROTOCOL  : TCP  AVAIL    : YES
WLM WEIGHT    : 00040    CS WEIGHT  : 100  NET WEIGHT: 00004
  Raw         ← CP: 40  zAAP: 40  zIIP: 60
  Proportional ← CP: 04  zAAP: 36  zIIP: 00
  FLAGS      :
...

```

Composite weight
40 = CP: 4 + zAAP 36

Normalized weight

Raw weights

Proportional weights

What was the processor usage?
CP 10% = (CP prop. Wt. 4) / (CP Raw wt.: 40)
zAAP 90% = 36/40

11

Sysplex load balancing © 2008 IBM Corporation

The load balancing advisor detail report shows the Raw weights received from WLM (CP: 40 zAAP: 40 zIIP:60) and the proportional weights received from WLM (CP: 04 zAAP: 36 zIIP: 00).

When ServerWLM is being used, the proportional weights are received from WLM. WLM determines the proportional weights based on current processor usage for that application.

LBA normalizes weights by dividing by the highest common denominator of all WLM weights received for that Port and Protocol. In this case with only one WLM weight (40), the highest common denominator is 40. $40/40 = 1$.

Things to think about

- zAAP and zIIP capacity will only be returned by WLM if all systems in the sysplex are V1R9 or later.
- In a mixed release environment only conventional CP weights will be used to determine the WLM System or Server-specific weight
- DNS/WLM will not exploit the new zAAP and zIIP processors. WLM recommendations will continue to only consider general processor capacity.
- LBA will consider zAAP and zIIP weight recommendations for server members but not system members. WLM recommendations for system members will continue to only consider general processor capacity.

12


Sysplex load balancing

© 2008 IBM Corporation


This first two bullets describe concerns when using zAAP and zIIP capacity in a mixed release environment.

In the 4th bullet,

- server members are those that are identified by IP address, port, and protocol - a server member is considered available when there is a protocol “listener” for that IP address, port, and protocol
- system members are only identified by IP address – a system member is considered available when its address is active (in the Home list)

IBM Software Group 

Add WEIGHTEDACTIVE distribution method for Sysplex Distributor

 13
Sysplex load balancing © 2008 IBM Corporation

This section describe why support for this function was added and how it works.

Background information - Sysplex distribution

- Incoming connections are routed to multiple target stacks using one of 3 distribution methods:
 - ▶ RoundRobin – Even distribution to all targets
 - ▶ BaseWLM – Uses WLM system weights
 - ▶ ServerWLM – Uses WLM server-specific weights
 - ▶ WLM weights (BaseWLM and ServerWLM) are normalized by dividing by 4. Weighted round robin distribution to targets uses the normalized weight.
- The distributor can reduce the Server or Base WLM Weights by using
 - ▶ Target Server Responsiveness fractions (TSR)
 - ✓ Connectivity between the distributing stack and the target stack - are new connection requests reaching the target? Target Connectivity Success Rate (TCSR)
 - ✓ Network connectivity between Server and client - are new connections being established? Connection Establishment Rate (CER)
 - ✓ Is the server accepting new work? Server accept Efficiency Fraction (SEF)

Currently, the Sysplex Distributor supports three distribution methods RoundRobin which evenly distributes incoming connections to all targets, BASEWLM which uses WLM system weights, and ServerWLM which uses WLM server-specific weights. When WLM weights are being used, they are normalized or reduced by dividing by 4. Incoming connection requests are distributed based on the normalized weights.

System weights (BaseWLM) and Server-specific weights (ServerWLM) are relative weights that range in value between 0 and 64. BaseWLM weights are based on a comparison of target Systems in the sysplex. How much processor 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). The distributor polls WLM for system weights each minute. 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? The target systems poll WLM for their server weights each minute and forward the weights to the distributor.

The received weights can be optionally modified by a QoS Service level fraction. A Service Level fraction measures the performance of the established connections that map to a DVIPA/Port on a target server. This includes the target to client performance, the ratio of retransmits and timeouts to number of packets sent, overall throughput and throughput/connection against required values, and the ratio of current connections against maximum connection limits. After the fraction is applied the weights are normalized (reduced) by dividing by 4. If all of the received WLM weights for a DVIPA/Port are less than or equal to 16, normalization is not done. After the fraction is applied against the raw weight, the weights are left unchanged.

The Target Server Responsiveness (TSR) fraction consists of 3 components, Target Connectivity Success Rate (TCSR) which is a measure of connectivity between the distributing stack and the target stack, Connection Establishment Rate (CER) which is a measure of network connectivity between Server and client (is the 3-way connection set up exchange completing?), and Server accept Efficiency Fraction (SEF) which is a measure of the Target Server's health. The weights are modified by the TSR fraction, and optionally the QoS fraction, before normalizing.

Background information - Sysplex distribution

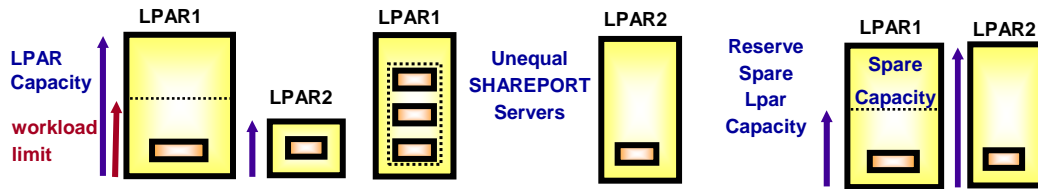
- WLM provides an interface which allows a server to pass additional information about its overall health:
 - ▶ Abnormal transaction completion Rate
 - ✓ Applications that use WLM monitoring of transactions (for example, CICS® Transaction Server for z/OS) can report an abnormal transaction completion rate to WLM
 - ✓ The value is between 0 and 1000 with 0 meaning no abnormal completions per 1000 transactions.
 - ▶ General health of the application
 - ✓ Applications can report their general health to WLM.
 - ✓ The value is between 0 and 100 with 100 meaning that a server has no general health problems (100% healthy).
- WLM will reduce the reported weight based on Abnormal Completion Rate and the General Health.
 - ▶ The Health Metrics are passed from WLM to Target System to Distributor for display purposes

WLM provides an interface which allows a server to pass additional information about its overall health. The following information may be used to reduce the weight passed to the stack.

- Abnormal transaction completion Rate - Applications such as the CICS Transaction Server for z/OS act as Subsystem Work Managers. They establish WLM Service Class goals, using WLM to monitor transactions against these goals; as part of this monitoring process, they can report an abnormal transaction completion rate to WLM (abnormal completions per 1000 transactions). The value is between 0 and 1000 with 0 meaning no abnormal completions.
- General health of the application - Applications can report their general health to WLM. The value is between 0 and 100 with 100 meaning that a server has no general health problems (100% healthy).

WLM will reduce the reported weight based on Abnormal Completion Rate and the General Health. The Health Metrics are passed from WLM to Target System to Distributor for display purposes.

Problem - Another distribution method needed



Application Scaling

- ▶ WLM recommendations may favor larger systems significantly.
- ▶ However, an application may not scale well to larger systems
- ▶ Application becomes overloaded when processor capacity is available, but the workload limit has been reached
- ▶ ServerWLM partially addresses this by reducing the recommendation if the application's Performance goals are not met

Shareport with unequal Servers per stack

- ▶ RoundRobin distributes 1 connection per target stack regardless of number of shareport servers
- ▶ WLM Server-specific weights from a target stack reflect the average weight of all shareport servers

Reserve Spare Capacity for timer driven workloads

- ▶ Batch workloads are injected into a system during specific times with specific completion requirements
- ▶ But WLM evenly consumes available capacity on all systems
- ▶ If the system is also a target for long running DDVIPA connections, these batch jobs may not complete on time if they are unable to displace the connection workload, OR the connection work is displaced and connection performance is affected

16

Sysplex load balancing

© 2008 IBM Corporation

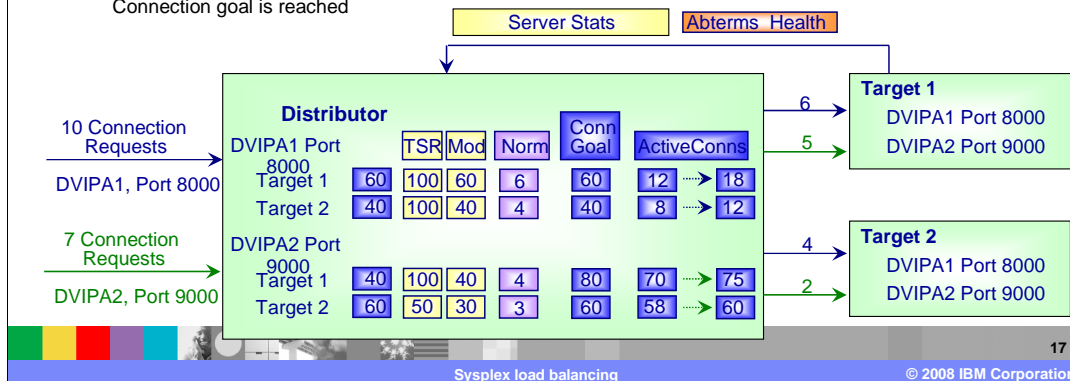
Application Scaling - Target systems can vary significantly in terms of capacity (small systems along with larger systems). WLM recommendations may favor the larger systems significantly. However, a target application may not scale well to larger systems; due to its design it may not be able to take full advantage of the additional processor capacity on the larger systems. As a result this type of server can get inflated WLM recommendations when running on larger systems causing it to be overloaded with work.

Unequal numbers of SHAREPORT Servers – If SHAREPORT is used, but not all systems have the same number of SHAREPORT server instances (one has 2 the other has 3). 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.

No reservation ability for time driven workloads – Users prefer to reserve some capacity on certain systems for types of batch workloads that run during specific time periods with specific completion requirements. If that system is also a target for long running DDVIPA connections, WLM recommendations will allow that available capacity to be consumed. This could potentially impact the completion times of the batch jobs if they are not able to displace the existing non-batch workloads or vice versa (the connections on that system may suffer from a performance perspective if the batch jobs displace those workloads).

Solution - WeightedActive distribution

- WeightedActive Distribution provides a more granular control over workload distribution
 - ▶ The comparative workload required on each system must be understood so that appropriate connection weights can be configured (fixed weights)
 - ✓ Configure these new parameters on the VIPADISTRIBUTE statement:
 - DISTMethod WEIGHTEDActive
 - Configure a Weight for each target destination
 - ▶ TSR values and Health Metrics (Abterms, Health) are applied to create a modified weight
 - ▶ Active connection count goals are determined based on the modified connection weights and the active connections on each target (multiple of modified weight > active connections)
 - ▶ Modified Weights are normalized by dividing by 10
 - ▶ Distribution is still weighted RoundRobin based on the normalized weight, but a target is skipped if a Connection goal is reached



17

Sysplex load balancing

© 2008 IBM Corporation

WeightedActive Distribution provides more granular control over workload distribution. The comparative workload required on each system must be understood so that appropriate connection weights can be configured (fixed weights). A new distribution method value of WEIGHTEDActive is added to the DISTMethod parameter. A weight can be configured for each DESTIP destination. Each weight can range in value from 1 to 99 so that the weights can be expressed as percentages. This example was configured using this method; the configured weights added up to 100, so that each weight could be shown as a percentage. Ideally each weight should be greater than 10 so that granularity is preserved when Autonomic fractions need to be applied to determine a modified weight. It defaults to 10, so if DESTIP ALL is configured, then the default weight of 10 is assumed which results in a connection distribution goal 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.

In the example, the Port 9000 Server distribution is determined as follows. Based on configuration, it is required that Target 1 will have 40% of the connection load and Target 2 will have 60% of the connection load. Since the TSR, abnormal terminations, and health are normal for Target 1, but 50 % for Target 2, the modified weight for Target 1 is 40 and the modified weight for Target 2 is 30 (60 * 50%). The Active Connection Goal is a value for each target such that if achieved would exactly match the required distribution proportions (it is always a multiple of the modified weight). The total number of active connections for both targets is 128. The total modified weight is 70. The multiplier used to determine the connection weight goal is 2 (128/70 + 1). Active connection goals are determined using the modified weight of each target and the multiplier. Target 1's goal is 80 (40 * 2) and Target 2's goal is 60 (30 * 2). The Normalized weight is the modified weight divided by 10. Therefore Target 1's normalized weight is 4 (40/10) and Target 2's normalized weight is 3 (30/10). As 7 connection requests are received: After the first 4 requests are evenly distributed between Target 1 & Target 2, Target 1 will have 72 active connections (Unused Normalized weight is 2) and Target 2 will have 60 active connections (Unused Normalized weight is 1). The next 3 requests will go to Target 1; although the normalized weight for Target 2 is not used up, the connection goal of 60 has been reached while Target 1's connection goal of 80 has not been reached. Assuming that the active connection counts do not change, the next 5 connection requests will go to Target 1. At this point both Target 1 and 2 will have reached their connection goals so the next connection request will cause a calculation of new target goals.

The existing MIB object, `ibmMvsDVIPADistConfDistMethod`, will indicate if WeightedActive Distribution is configured. A new MIB object, `ibmMvsDVIPADistConfTargetWeight`, will display the configured weight for each target.

Display command example VIPADCFG DETAIL

- Use the Netstat VIPADCFG/F Detail report to display the configured distribution method and the weights (if DISTMethod is WEIGHTEDActive)

```
NETSTAT VIPADCFG DETAIL
VIPA Distribute:
  Dest:      201.2.10.11..8000
  DestXCF:   201.3.10.15
  SysPt:    No  TimAff: No  Flg: WeightedActive
  OptLoc:   No  Weight: 80
  Dest:      201.2.10.11..8000
  DestXCF:   201.3.10.16
  SysPt:    No  TimAff: No  Flg: WeightedActive
  OptLoc:   No  Weight: 20
```

VIPADCFG is modified to display the new configured distribution method of WEIGHTEDActive along with the configured weights for each target.

Display command example - VDPT DETAIL

- Use the Netstat VDPT/O DETAIL report to display the active distribution method, the modified weight, and the active connection counts for each target

```

NETSTAT VDPT DETAIL
Dynamic VIPA Destination Port Table:
Dest:      201.2.10.11..8000
DestXCF:   201.3.10.15
TotalConn: 0000084011 Rdy: 001 WLM: 20 TSR: 50
Flg: WeightedActive
TCSR: 100 CER: 100 SEF: 50
Abnorm: 0000 Health: 50
ActConn: 0000000240

Dest: 201.2.10.11..8000
DestXCF: 201.3.10.16
TotalConn: 0000020340 Rdy: 001 WLM: 20 TSR: 100
Flg: WeightedActive
TCSR: 100 CER: 100 SEF: 100
Abnorm: 0000 Health: 100
ActConn: 0000000058

```

20 = configured weight (80) *
TSR(50%) * Health (50%) * Abterms (100%)



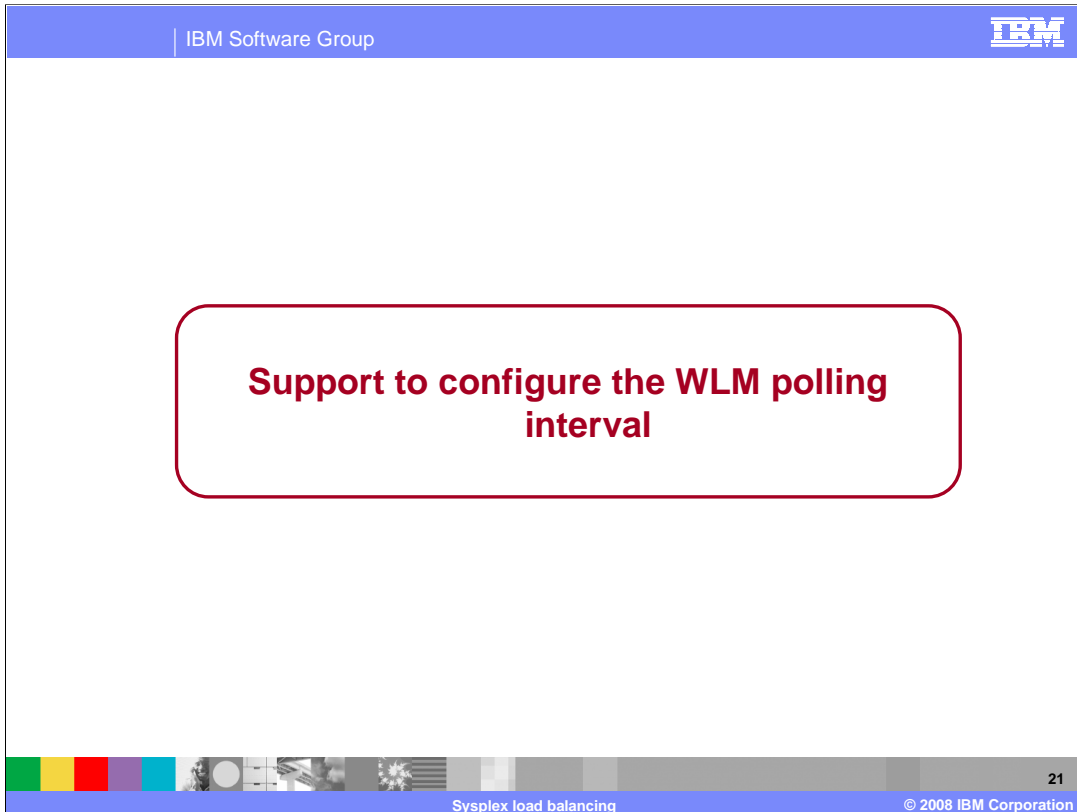
With this release some of the detailed displays are simplified to only show values that pertain to a distribution method.

The detailed version of the report when the distribution method is WEIGHTEDActive shows the TSR metric components (TCSR, CER, and SEF) and the Health metrics (abnormal terminations and health) that are used to determine the modified weight along with the active connection counts.

Things to think about

- If the distributor is V1R9, WeightedActive distribution can be used regardless of the Target stack release level. However:
 - ▶ A target stack needs to be at least V1R7 so that TSR metrics are reported to the distributor
 - ✓ TSR is considered to be 100% when a target is pre-V1R7
 - ▶ A target stack needs to be at least V1R8 so that health metrics (abnormal terminations and health) are reported to the distributor
 - ✓ Health and normal termination rate are considered to be 100% when a target is pre-V1R8
- Each backup stack needs to be V1R9 or later to allow WeightedActive distribution to be inherited during a takeover, otherwise BASEWLM will be used

This slide describes concerns when systems are in a mixed release environment.



The next group of slides describe a new function APAR which allows the sysplex WLM polling interval to be configured.

Background information - Wlm polling interval

- Interaction between the stack and WLM
 - ▶ The TCP/IP stacks poll WLM every 60 seconds for weights
 - ✓ BASEWLM - the distributor polls WLM for system weights
 - ✓ SERVERWLM - each target polls WLM for server-specific weights
 - ▶ WLM calculates new weights
 - ✓ Based on a comparison of the last 10 seconds of processor utilization on registered sysplex systems or servers
 - ✓ It keeps a 3 minute rolling average of these calculations
 - ✓ This average is returned by WLM when it is polled
 - ▶ The TCP/IP polling interval assumes weights will not change significantly from minute to minute.

Interaction between the Sysplex stacks and WLM depends on the distribution method. When BASEWLM is being used, the Sysplex Distributor polls WLM every 60 seconds for weights from all systems. When SERVERWLM is being used, each target polls WLM for server-specific weights which are then sent to the distributor. The distributor uses the received WLM weights to determine how to distribute connections to the target systems; a weighted round-robin distribution is used based on the WLM weights.

WLM calculates new weights based on a comparison of the last 10 seconds of processor utilization on registered sysplex systems. It keeps a three minute rolling average of these calculations. When WLM receives a poll request, the three minute rolling average is returned for each system in the sysplex. The 3 minute average is used to smooth the weight changes.

The distributor's 1 minute polling interval was determined based on the assumption that WLM weights do not change significantly from minute to minute.

Problem statement - See-saw distribution

- A user environment had two distribution targets
 - ▶ BASEWLM
 - ▶ Systems running at 100% capacity
 - ▶ High volumes of short lived connections
- The 3 minute rolling average *changed significantly* between the 1 minute polling intervals
- Distributor reacted to WLM changes too slowly:
 - ▶ When noticed, one server was overloaded and the second server was underutilized.
 - ▶ During the next polling interval, the distributor overloaded the second server and the first server became underutilized.
 - ▶ This "see-saw" distribution continued during each successive polling interval
 - ✓ The connection load shifting back and forth between the two servers
 - ✓ Never reaching a steady WLM weight and connection load for each server.

In this environment, the load on target systems was close to 100% capacity with a workload consisting of high volumes of short lived connections. In this type of environment the 3 rolling minute average *changed significantly* between the 1 minute polling intervals. So the original design point was no longer valid for this environment.

The distributor was reacting too slowly to changes in the WLM recommendations between the target servers. At the time the problem was noticed, the first target server was overloaded with connections and the second server was underutilized. From this point on, there was a "see-saw" distribution. During the next minute interval, the new WLM weights caused the distributor to direct most of the connections to the second server causing it to be overloaded and the first server to be underutilized. This continued from interval to interval with the distributor continually shifting most of the connection load back and forth between the two servers, never reaching a steady WLM weight and connection load for each server.

Solution – Make polling interval configurable

- Allow the polling interval to be configured
 - ▶ GLOBALCONFig SYSPLEXWLMPoll
- The default will continue to be 60 seconds since this continues to work well in most environments
- The polling interval will be applied to both SERVERWLM and BASEWLM polling
- This support is available on prior releases in APAR PK24752
- Mixed sysplex environment considerations
 - ▶ BASEWLM
 - ✓ Since the distributor polls WLM for the system weights of all target stacks, only the distributor needs to be V1R9 to change the polling interval
 - ✓ If the backup stack is not V1R9, then APAR PK24752 must be applied
 - ▶ SERVERWLM
 - ✓ Since target stacks poll WLM for server-specific weights, all target stacks, backup stack, and the distributor should be V1R9 or APAR PK24752 must be applied to change the polling interval
 - ✓ The polling interval needs to be consistent on all target stacks and the distributor to be effective

A new parameter, SYSPLEXWLMPoll, is added to the GLOBALCONFig statement to allow a user to control the polling interval. It can range between 1 and 60 seconds. The default polling interval will remain 60 seconds. As a guideline the polling interval should not be lower than the WLM weight calculation interval (currently 10 seconds). The polling interval will apply to both SERVERWLM and BASEWLM polling.

In a mixed sysplex environment where some of the systems are not at the z/OS V1R9 Communications Server level, then APAR PK24752 may need to be applied to the pre-V1R9 system to activate this support.

Display command example - NETSTAT CONFIG

- Use the Netstat CONFIG/-f report to display the polling interval

```
D TCP/IP,TCPCS1,NETSTAT,CONFIG
...
GLOBAL CONFIGURATION INFORMATION:
TCPSTATS: NO   ECSALIMIT: 0000000K  POOLLIMIT: 0000000K
MLSCHKTERM: NO  XCFGRPID:      IQDVLANID:  0
SEGOFFLOAD: YES  SYSPLEXWLPOLL: 060
EXPLICITBINDPORTRANGE: 10000-11023
SYSPLEX MONITOR:
    TIMERSECS: 0060  RECOVERY: YES  DELAYJOIN: YES  AUTOREJOIN: YES
    MONINTF:  NO    DYNROUTE: NO
ZIIP:
    IPSECURITY:NO
```

The sysplex WLM polling interval will be displayed as part of the Global Configuration Information.

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_Syplex_LoadBal.ppt

This module is also available in PDF format at: [../Syplex_LoadBal.pdf](..../Syplex_LoadBal.pdf)



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:

CICS System z z/OS zSeries

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.

© Copyright International Business Machines Corporation 2008. All rights reserved.

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.

