#SHAREorg

SHARE
Technology · Connections · Results

Route Me, WLM Workload Manager Functions for Dynamic Workload Routing

Horst Sinram
IBM Germany Research & Development
z/OS Workload Management

11 Mar 2014
Session 15217

SHARE in Anaheim

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

AIX*	DS8000*	Language Environment*	SystemPac*	z10
BladeCenter*	FICON*	Parallel Sysplex*	System Storage	z10 BC
DataPower*	HiperSockets	POWER7*	System z	z10 EC
DB2*	Hyperwap	PrintWay	System z9	z/OS*
DFSMS	IBM*	ProductPac*	System z10	zEnterprise
DFSMSdss	IBM eServer	RACF*	System z10 Business Class	zSeries*
DFSMShsm	IBM logo*	REXX	WebSphere*	
DFSMSrmm	ibm.com	RMF	z9*	
DFSORT	Infiniband*	ServerPac*		
DS6000*	InfoPrint			

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

InfiniBand is a registered trademark of the InfiniBand Trade Association (IBTA).
Intel is a trademark of the Intel Corporation in the United States and other countries.
Linux is a trademark of Linux Torvalds in the United States, other countries, or both.
Java and all Java-related trademarks and logos are trademarks or registered trademarks of Sun Microsystems, Inc., in the United States and other countries.
Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation.
UNIX is a registered trademark of The Open Group in the United States and other countries.
All other products may be trademarks or registered trademarks of their respective companies.
The Open Group is a registered trademark of The Open Group in the US and other countries.

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.

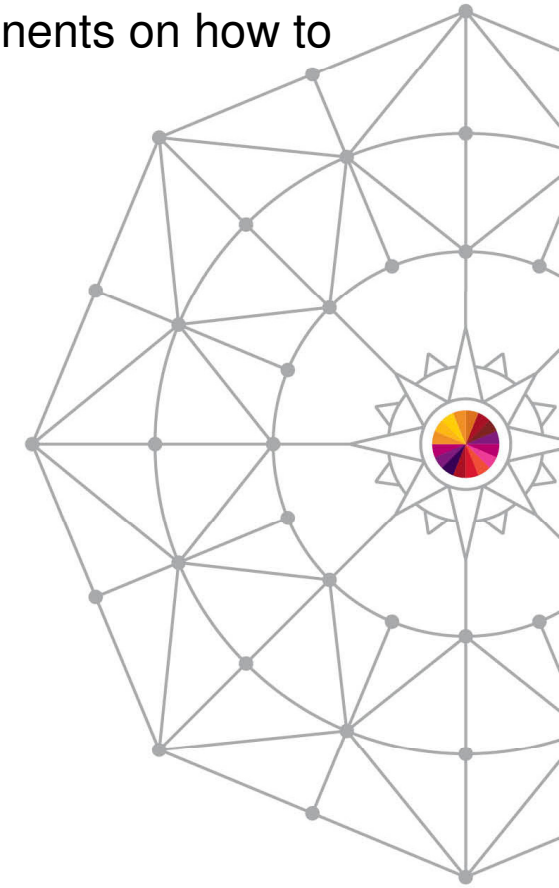
This presentation and the claims outlined in it were reviewed for compliance with US law. Adaptations of these claims for use in other geographies must be reviewed by the local country counsel for compliance with local laws.

Agenda

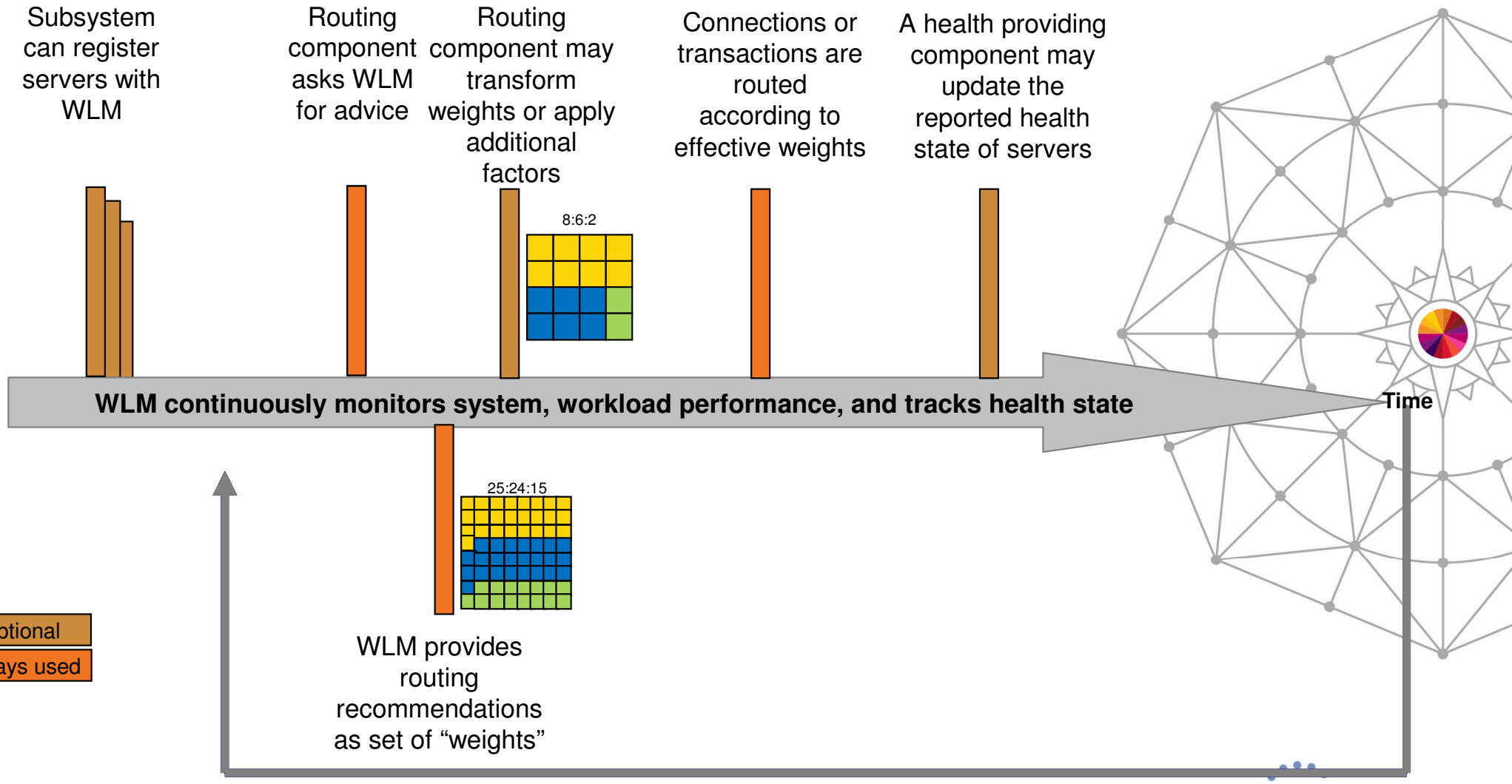
- Concepts
  - Importance levels
  - Displaceable capacity
  - Free capacity
- WLM Sysplex Routing Services
  - IWMWSYSQ
  - IWMSRSRS
  - IWM4SRSC
  - Basic capacity-based weights and additional influencers
- Observations, best practices and optimization approaches

WLM Dynamic Workload Routing Services

- WLM Sysplex routing services provide guidance to routing components on how to distribute
  - Transactions
  - Connections
- Multiple sets of routing APIs are offered by WLM
  - Same underlying capacity view but different algorithms and influencing parameters
- Scope
  - Multiple systems of one Sysplex, one or more servers per system
- Primary objectives for balancing:
  - Capacity – Route work according to capacity available
  - Performance – WLM goal attainment
  - Integrity – Avoid shortages
  - Reliability – Avoid not healthy work consumers

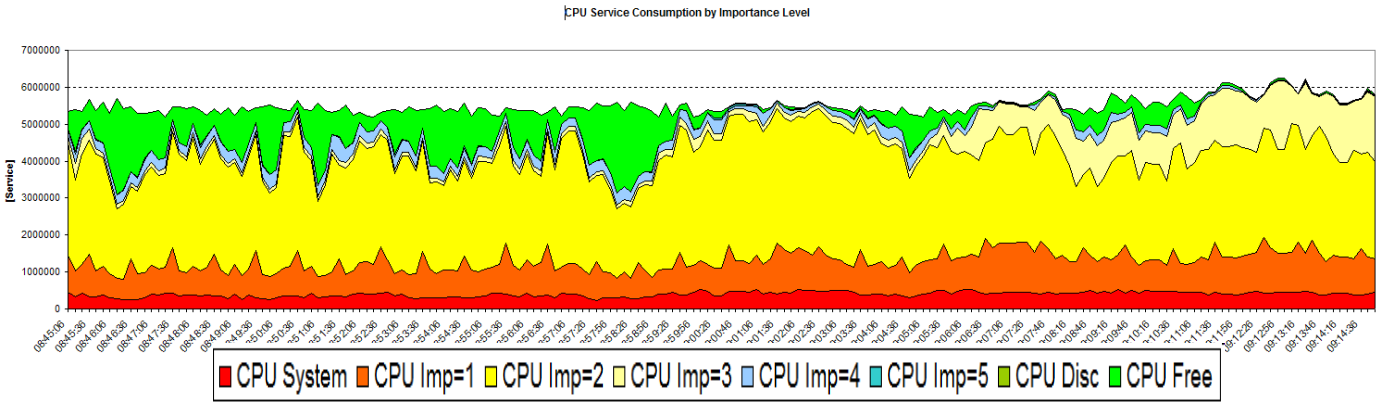


The life cycle of workload routing recommendations



Concepts: Service consumption by importance level

- WLM/SRM tracks the consumption of *CPU service* by importance level
- WLM management will sacrifice less important work to allow more important work to achieve the goals. Less important work may be *displaced* entirely.



- Level 0:
  SYSTEM and SYSSTC
- Level 1-5:
  Importance 1 through 5
- Level 6:
  Discretionary
- Level 7:
  Free (unused) capacity

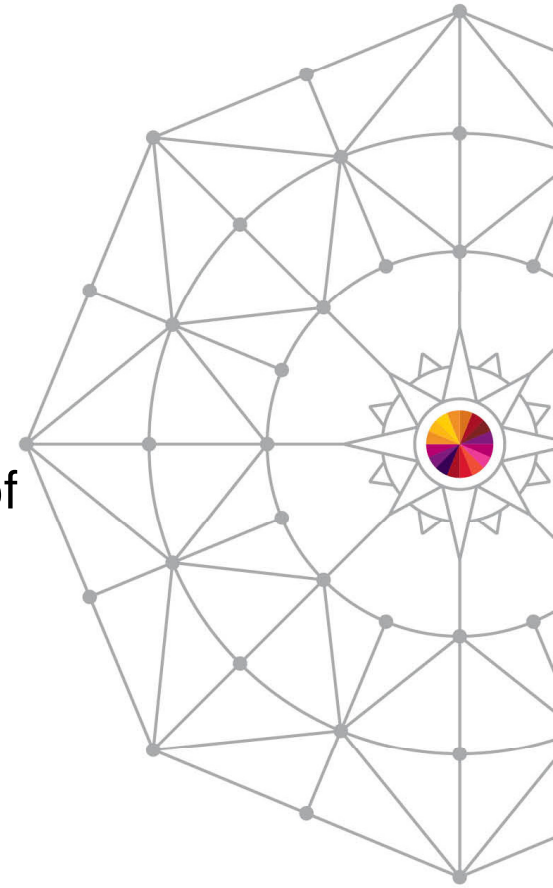
Concepts: WLM determination of displaceable capacity

- An important metric for routing decisions is the *displaceable capacity* at a given importance level (i):

$$DisplaceableCapacity_i = FreeCapacity + \sum_{j=i+1}^{6} CapacityConsumed_j$$

or

$$DisplaceableCapacity_i = \sum_{j=i+1}^{7} CapacityConsumed_j$$

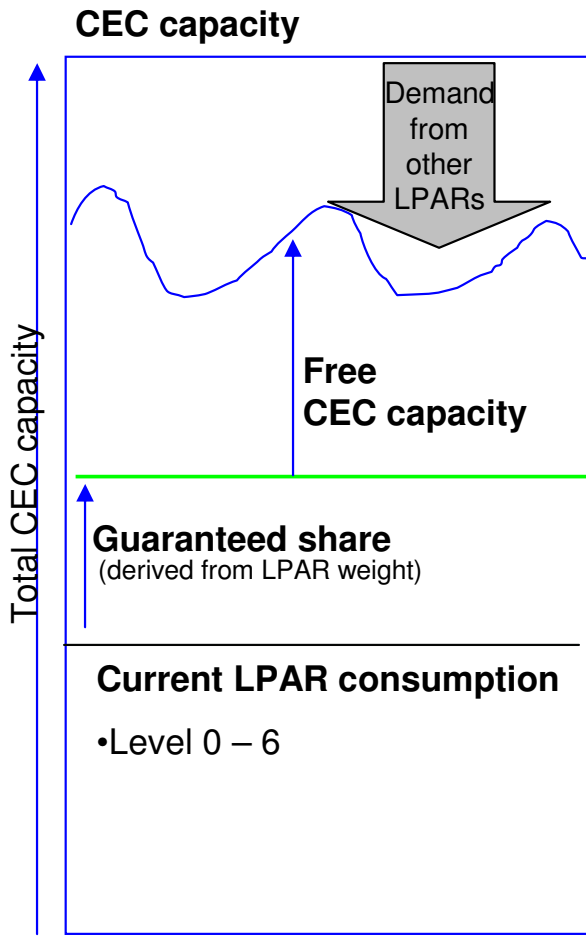


- For the purpose of routing the **3 min rolling averages** of consumption and free capacity are considered
- The *consumed capacity* is usually well understood
  - *Free capacity* may be harder to understand
  - Needs to reflect many different constraints that could limit the capacity that can be consumed by an LPAR.
- All processor types to be assessed independently

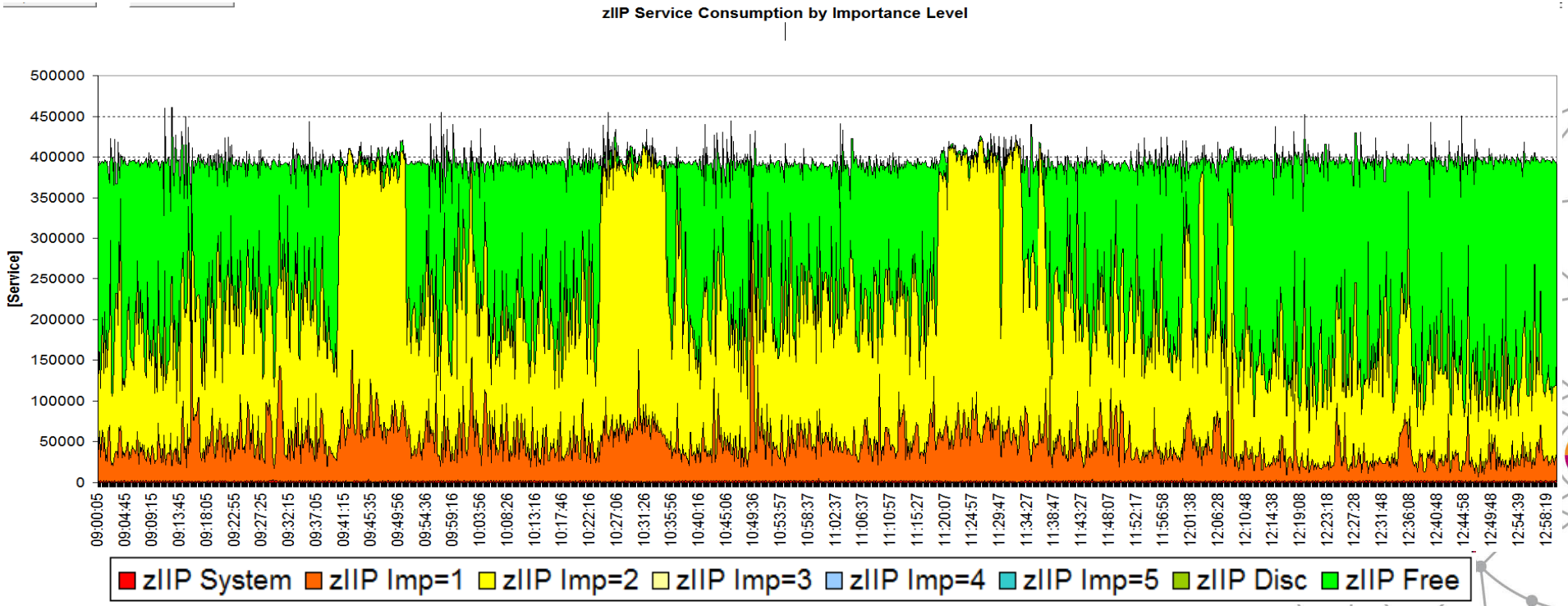
Concepts: LPAR Capacity
What limits an LPAR's capacity?

- Logical capacity (number of logical processors)
- LPAR initial cap (aka hard cap), LPAR absolute cap (zEC12 GA2)
- Defined capacity (aka soft cap)
  - LPAR level defined capacity
  - Group capacity

  Defined capacity is only considered while capping is in effect
- LPAR weight
  - Guaranteed capacity unless configuration parameters prohibit the guaranteed capacity to be consumed
  - IRD weight management may change weights dynamically hence guaranteed capacity changes
- Available CEC capacity – unused CEC capacity can be consumed beyond weight
- In addition, consider
  - MVS Busy (MVS wait time)
  - LPAR configuration: shared vs. dedicated vs. „Wait completion =Yes"

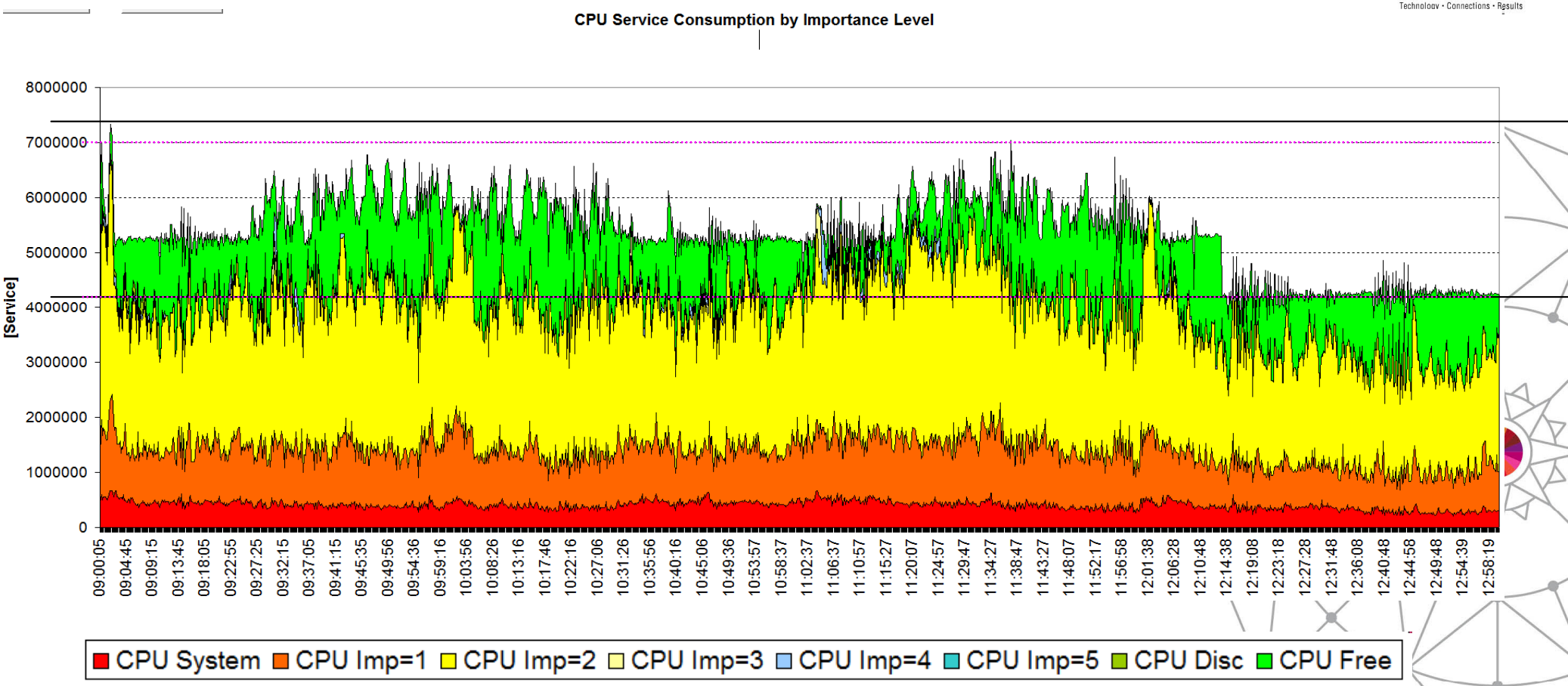


Free LPAR Capacity - Example 1



- While an LPAR is running below its weight entitlement and no capping is in effect the total consumed plus free capacity is usually pretty constant.

Free LPAR Capacity - Example 2



- Capping, group capping, and influences by other LPARs can **heavily** and **frequently** change the total capacity available to an LPAR

Free LPAR Capacity – some considerations

- *A single capacity value can hardly represent all the different preferences that installations may have*.

  Examples:

  - Preferentially displace the lowest importance work
  - Minimize/control crossover of zIIP/zAAP work to CPs
  - Equal distribution of used capacity
  - Preferential use of guaranteed capacity vs. free CEC capacity
  - Leave whitespace for expected workloads, e.g. batch
  - Anticipation of capping before capping becomes active
  - Availability/anticipation of not activated temporary capacity (On/Off Capacity on Demand)
  - Avoid usage of activated temporary capacity
  - …

Blue: Controls are available

Agenda

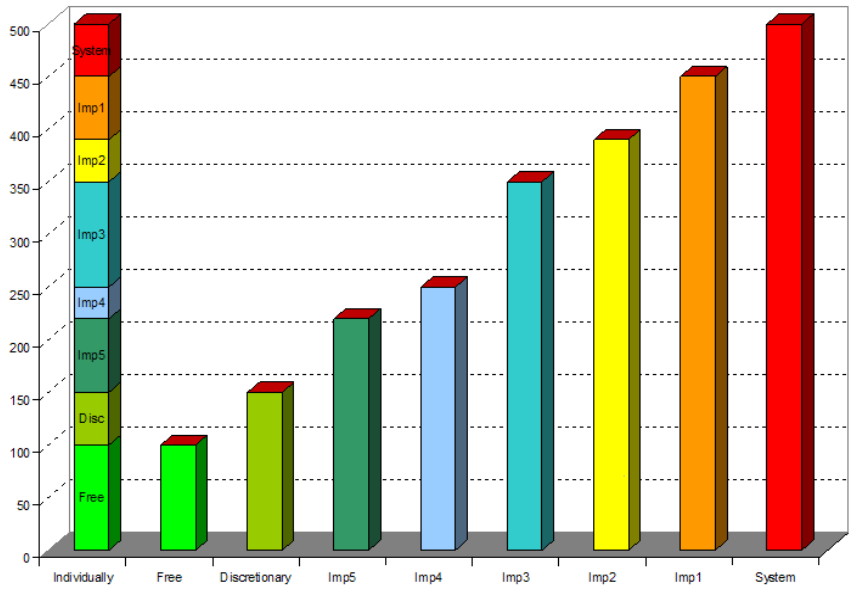
- Concepts
  - Importance levels
  - Displaceable capacity
  - Free capacity
- WLM Sysplex Routing Services
  - IWMWSYSQ
  - IWMSRSRS
  - IWM4SRSC
  - Basic capacity-based weights and additional influencers
- Observations, best practices and optimization approaches

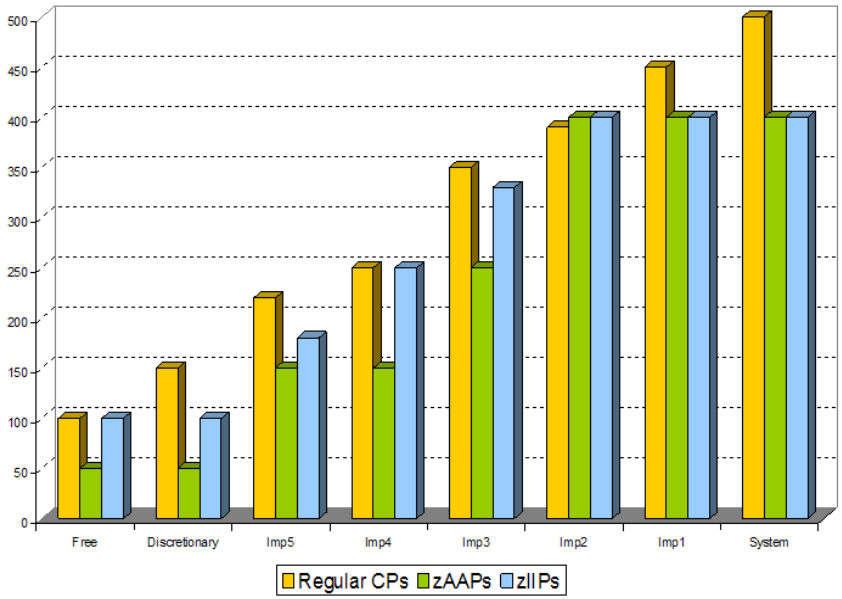
WLM Sysplex Routing Services Overview

Interface	Purpose	Typical Use (not exhaustive)
IWMWSYSQ	Obtain free & displaceable capacity of systems in Sysplex (1, 3, and 10 min rolling averages).	Customer applications and subsystems that want to consider free and displaceable capacity.
IWMSRSRS FUNCTION=SELECT (IWMSRSRG,DRS)	Obtain best suited registered servers to route work to. Only **capacity** considered.	Sysplex Distributor BASEWLM,
IWMSRSRS FUNCTION=SPECIFIC (IWMSRSRG,DRS)	Obtain list of registered eligible servers and recommended weights. Besides **capacity** goal achievement (**PI**), **queue time** for enclaves, **health** indicator is considered.	DDF
IWM4SRSC	For a specific server address space obtain recommendation how suitable a server is. No registration required. Besides **server-specific capacity** goal achievement (**PI**), **abnormal termination** rate, **health** indicator is considered.	Sysplex Distributor SERVERWLM
IWM4HLTH	Provide health status for an address space. Value is considered by IWM4SRSC and IWMSRSRS FUNCTION=SPECIFIC	CICS Transaction Gateway, DDF, LDAP.

Routing Services: IWMWSYSQ

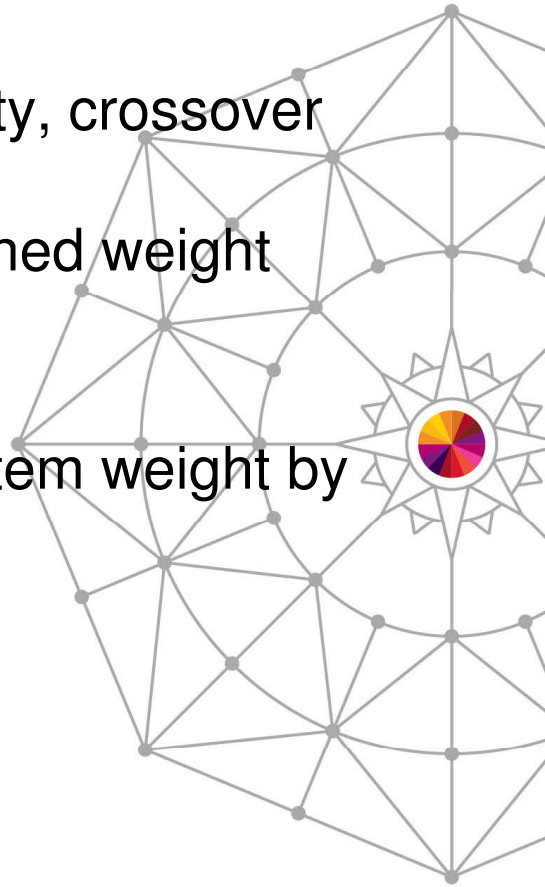
- Provides displaceable capacity at each importance level
  - The system level contains the total system capacity, including SYSTEM work
  - Rolling average over 60, 180, and 600 sec.
- Data are returned for all processor types
- In addition: System shortages information, uniprocessor speed of a single processor, zAAP and zIIP normalization factors– required for subcapacity models
  - EXTENDED_DATA=NO returns the output area in the pre-z/OS V1.9 format





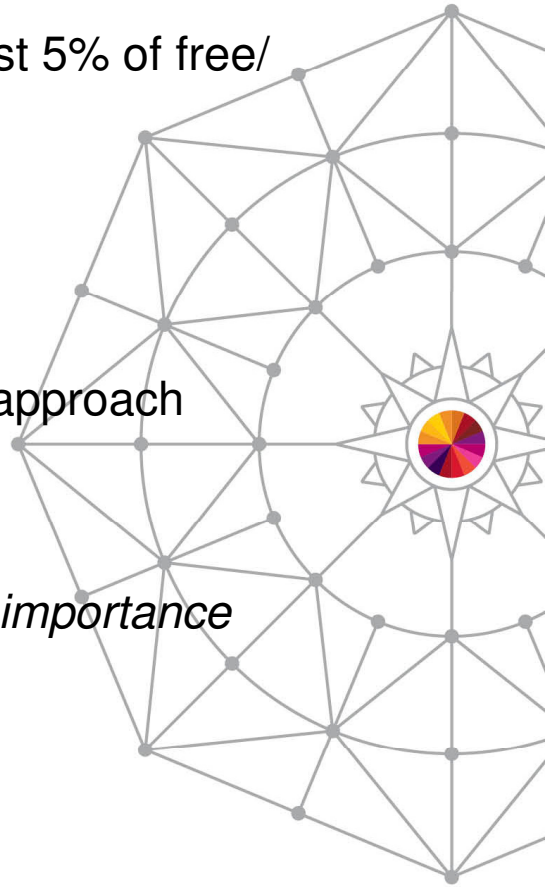
WLM Routing Weights Computation Overview: *Steps Involved*

- Compute capacity-based weights for systems
  - Includes adjustment for specialty processor capacity, crossover cost, and importance level weighting
  - Return weights for each processor type and combined weight
  - Frequently scaled to 64
- When multiple servers run on a system divide the system weight by #servers to derive a server's weight
- Only for IWMSRSRS SPECIFIC and IWM4SRSC modify weights based on
  - Performance index
  - Queue time ratio
  - Health indicator



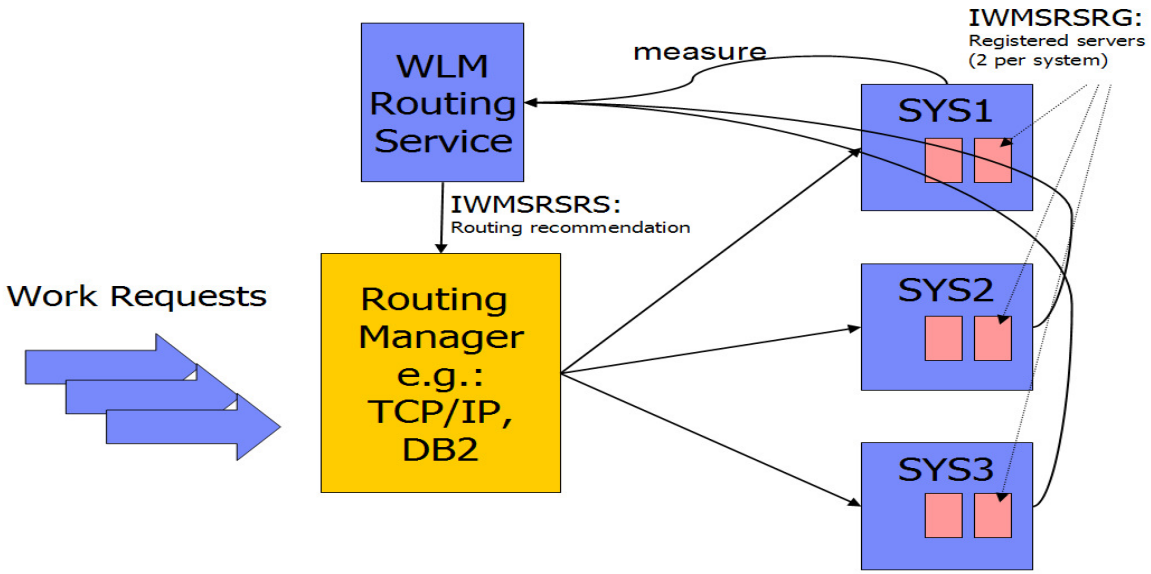
IWMSRSRS vs. IWM4SRSC Capacity calculations

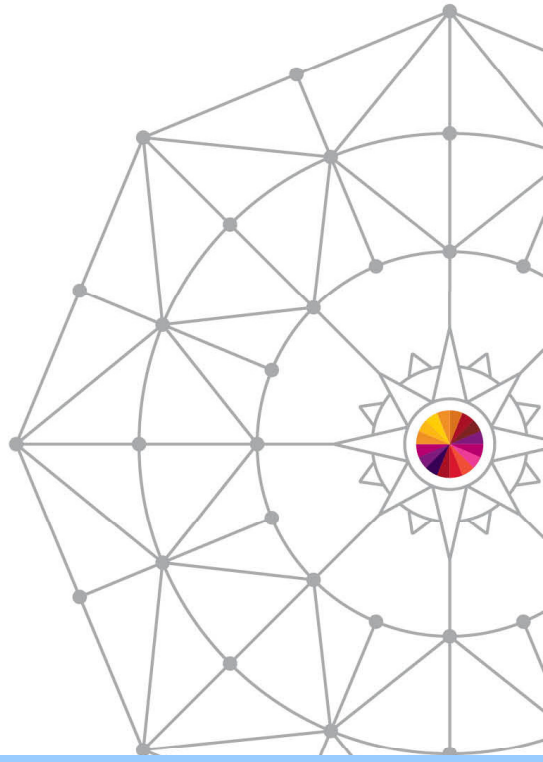
- **BASEWLM (WLM service IWMSRSRS)**
  - Locate the importance level –searching bottom-up- where at least 5% of free/ displaceable capacity is available on one system
  - Disadvantage
    - Does not consider the importance of the work
    - May result in oscillations
  - Advantage
    - Considers the low important work because it is a bottom up approach



- **SERVERWLM (WLM service IWM4SRSC)**
  - Calculates the weight based on the *displaceable capacity at the importance level that the work will run* on the systems.
    Advantage
    - Considers the importance of the work
    - Avoids the oscillation of routing recommendations
  - Possible Disadvantage
    - Lower important work isn't distinguished from free capacity
      - *Solution: Importance Level Weighting*

Sysplex Routing with IWMSRSRS: Bottom-Up Weight Calculation





Algorithm

1. Select the importance level that provides at least **5% of cumulative capacity on at least one** system
2. Calculate system weight on each system
3. Calculate server weight:

$$\text{System Weight} = \frac{\text{SUs}_{\text{at selected level}}[\text{this system}]}{\sum_{\text{I for all systems}} \text{SUs}_{\text{at selected level}}[i]} \bullet 64$$

$$\text{Server Weight} = \frac{\text{System Weight}}{\#\ \text{of servers on system}}$$

Sysplex Routing with IWMSRSRS: Bottom-Up Weight Calculation

Example:

Two Servers per System

Selected Importance Level: 5

Sum:

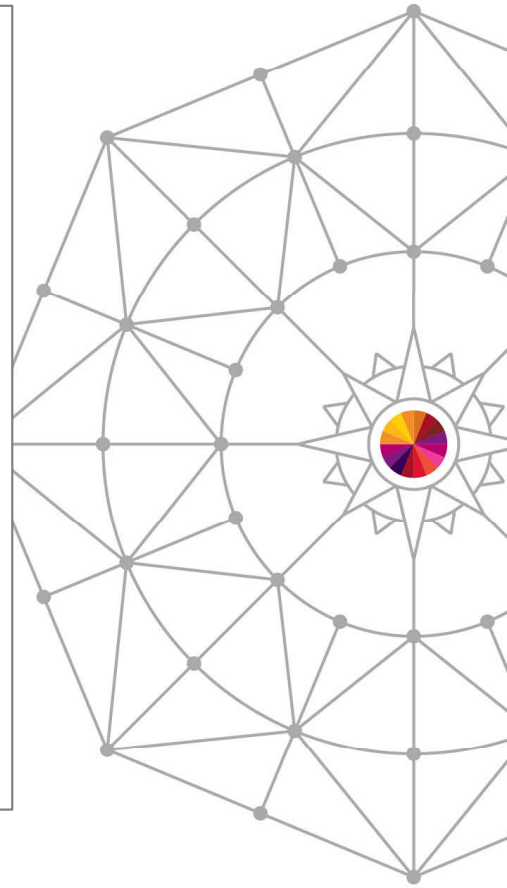
200+400+300 = 900

	System 1		System 2		System 3	
Level	SUs	%	SUs	%	SUs	%
0	2000	100	2000	100	2000	100
1	1800	90	1900	95	1840	92
2	1600	80	1500	75	1600	80
3	1100	55	1500	75	800	40
4	400	20	1200	60	800	40
5	200	10	400	20	300	15
6	80	4	20	1	0	0
7	0	0	0	0	0	0

System 1 weight = 200 *64 / 900 = 14 **Server weight** = 14 / 2 = **7**

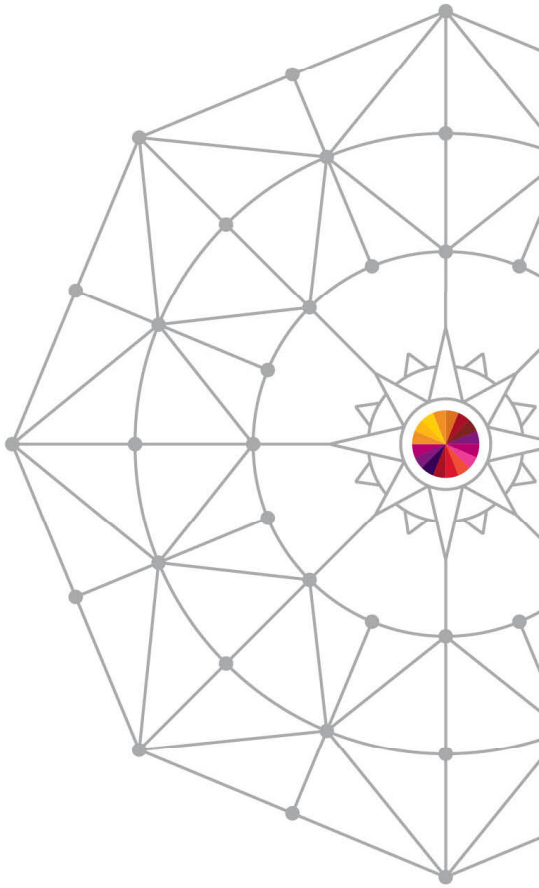
System 2 weight = 400 *64 / 900 = 28 **Server weight** = 28 / 2 = **14**

System 3 weight = 300 *64 / 900 = 21 **Server weight** = 21 / 2 = **10**

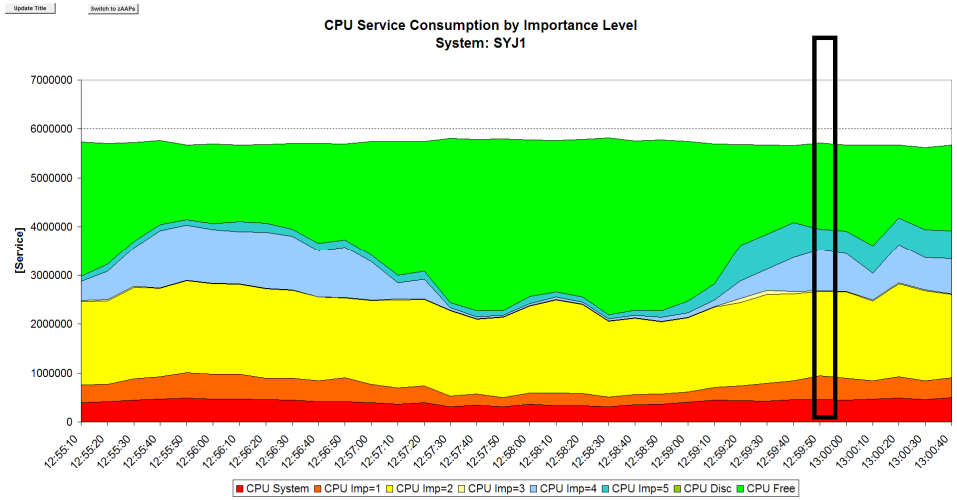


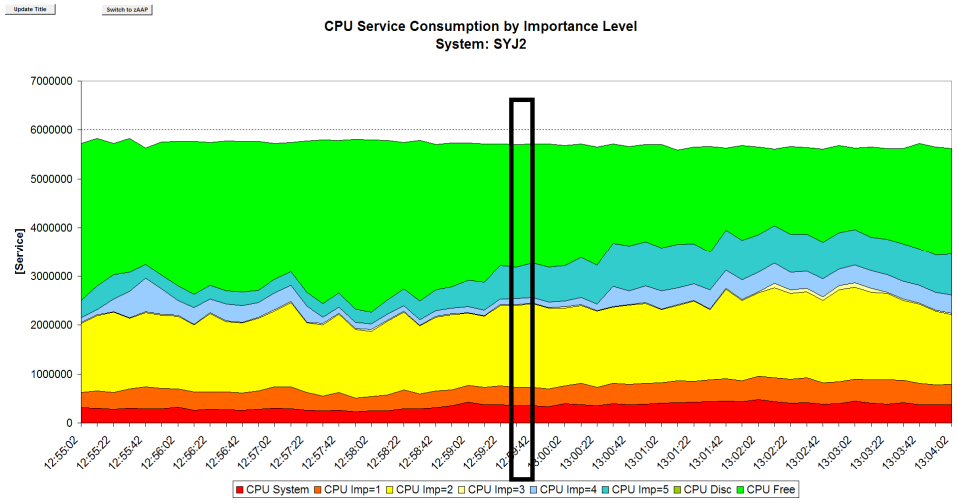
Example: How WLM computes weights
Base Data for following Examples

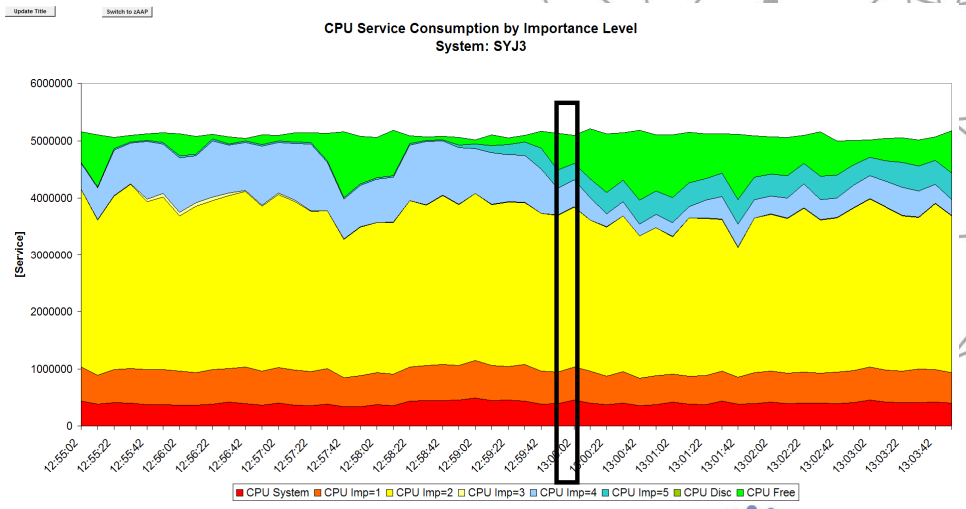
- The 3 charts on the next page show the service consumption by importance level for three systems of a Sysplex environment
- Observations:
  - The consumption at importance level 1 is similar on all systems
  - The consumption at importance level 2 is much higher on system SYJ3
- On the following charts we use data from one time interval to show the calculations for BASEWLM and SERVERWLM
  - This is calculation is simplified but comparable to the actual WLM algorithm



Example: How WLM computes weights
Base Data for following Examples







Example: System weights for IWMSRSRS vs. IWM4SRSC

DATA

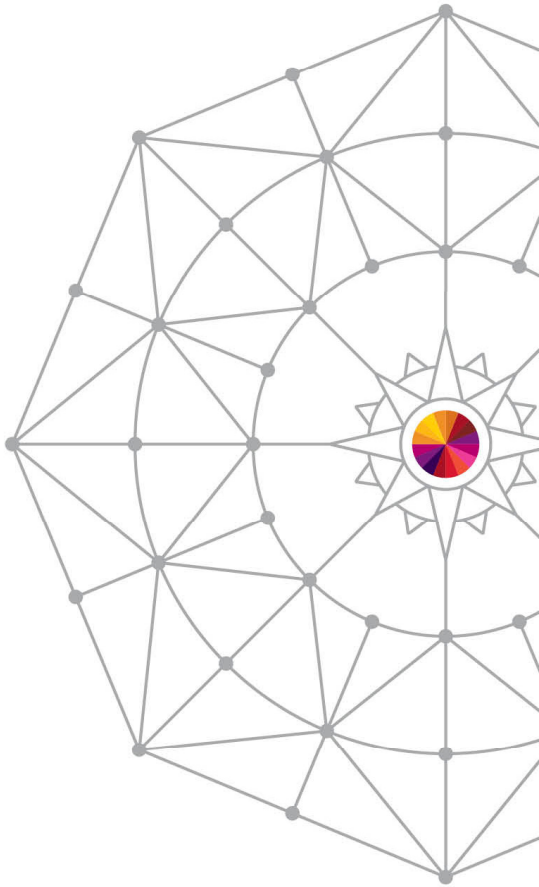
Time/System	13:00:00/J1			13:00:02/J2			13:00:02/J3		
	SU/IL	Sum	%	SU/IL	Sum	%	SU/IL	Sum	%
System	450307	5668854	100%	392443	5683002	100%	450802	5097430	100%
IL 1	438985	5218547	92%	366208	5290559	93%	578145	4646628	91%
IL 2	1762932	4779562	84%	1584407	4924351	87%	2813465	4068483	80%
IL 3	14875	3016630	53%	31141	3339944	59%	6443	1255018	25%
IL 4	773676	3001755	53%	116617	3308803	58%	473513	1248575	24%
IL 5	473036	2228079	39%	723396	3192186	56%	284036	775062	15%
Disc	0	1755043	31%	0	2468790	43%	0	491026	10%
Free	1755043	1755043	31%	2468790	2468790	43%	491026	491026	10%

BASEWLM

Selected Importance Level=		7
Capacity at the Importance Level=	1755043+2468790+491026	4714859
Weight for System J1=	1755043/4714859*64	24
Weight for System J2=	2468790/4714859*64	34
Weight for System J3=	491026/4714859*64	7

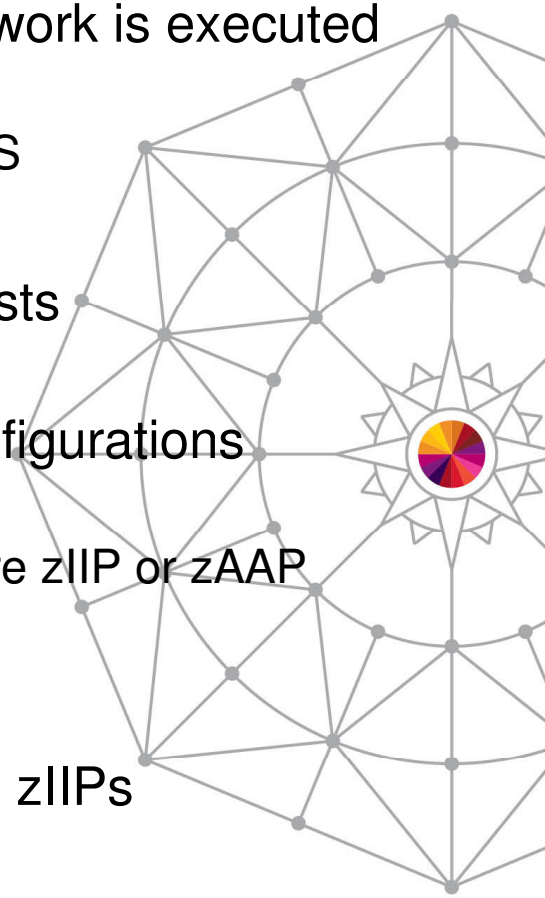
SERVERWLM

Selected Importance Level=		1
Max Capacity at the Importance Level=	System J2	5290559
Weight for System J1=	5218547/5290559*64	63
Weight for System J2=	5290559/5290559*64	64
Weight for System J3=	4646628/5290559*64	56

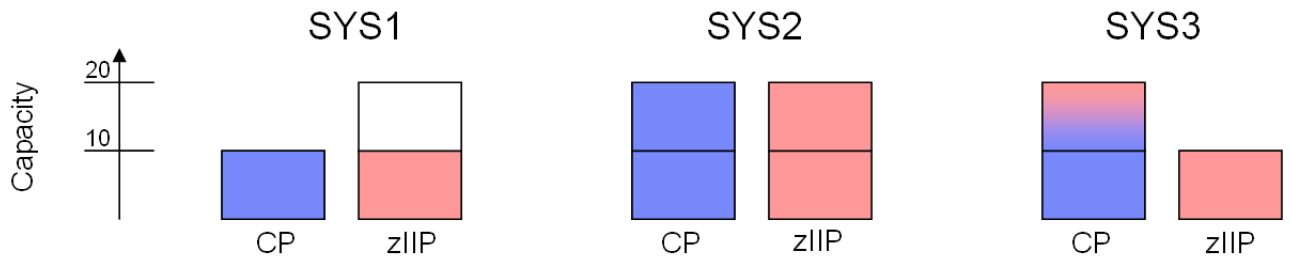


WLM Routing: Crossover to CP cost

- By default:
  - Assumption is that there is no penalty when zAAP or zIIP work is executed on regular CPs
    - For IFAHONORPRIORITY=YES or IIPHONORPRIORITY=YES
- In Reality:
  - Executing zAAP or zIIP work on regular CPs may incur costs
    - Work is priced by its consumption on regular CPs
  - Systems to which work is routed to may have different configurations
    - Number of zIIPs or zAAPs may differ between systems
    - It might be of advantage to route the work to systems with more zIIP or zAAP capacity
- Solution
  - SERVERWLM allows to specify ProcXCost for zAAPs and zIIPs
    - Begin with small cost values
- As a result WLM will use a different method to calculate the routing weights which reflects capacities of different processor types much better



WLM Routing: Proportional vs. Equivalent CPU



Assumption: Work uses 50% CP and 50% zIIP, HONORPRIORITY=YES

Algorithm	SYS1		SYS2		SYS3	
	CP	zIIP	CP	zIIP	CP	zIIP
Proportional	32	64	64	64	64	32
Proportional Combined	48		64		48	
Equivalent CPU	10		20		15	
Equivalent CPU Combined and Scaled	32		64		48	

- METHOD=EQUIVALENT
  - Is required when different cost factors should be applied to regular CPs and zIIPs or zAAPs
  - Advantageous if significant zIIP and/or zAAP capacity is installed
  - Is not required and does n'o provide any different results when only regular CP capacity is installed or only little zIIP or zAAP capacity is installed.

WLM Routing: Importance level weighting

Time	System	System	IL 1	IL 2	IL 3	IL 4	IL 5	Disc	Free
13:00:00	J1	450307	438985	1762932	14875	773676	473036	0	1755043
13:00:02	J2	392443	366208	1584407	31141	116617	723396	0	2468790
13:00:02	J3	450802	578145	2813465	6443	473513	284036	0	491026

Weighted Sus			
Constant	Square Root	Linear	Quadratic
5218547	10206412	21754659	117826411
5290559	11043437	24993423	147905587
4646628	7449427	12975818	50627374

IWM4SRSC weights			
Constant	Square Root	Linear	Quadratic
63	59	56	51
64	64	64	64
56	43	33	22

Square Root	IL_Weighting=1	1.00	1.41	1.73	2.00	2.24	2.45	2.65
Linear	IL_Weighting=2	1	2	3	4	5	6	7
Quadratic	IL_Weighting=3	1	4	9	16	25	36	49

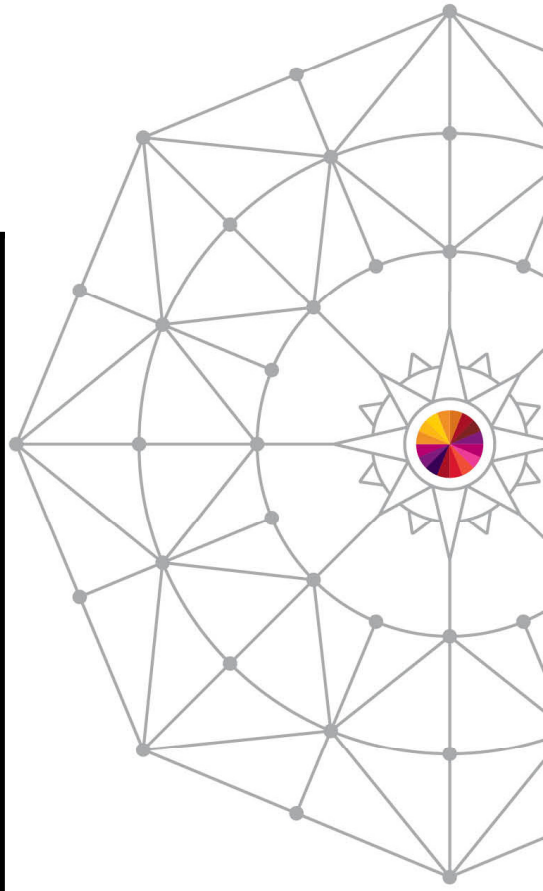
- Importance Level Weighting is available with service IWM4SRSC (SD SERVERWLM)
- Default routing algorithm uses “Constant” - no weighting of the lower importance levels
  - As a result the weights returned by service IWM4SRSC to Sysplex Distributor are nearly identical for all three systems (see column “Constant”)
    With importance level weighting it is possible to factor in the work running at lower importance levels.
  - Three weighting levels exist: Square Root (mildly), Linear and Quadratic (heavy) weighting
  - You can observe that the biggest effect is for system J3 on which much more work runs at importance level 2
  - Concern: A too high weighting can cause oscillation effects

Performance Index (PI) effect on routing weight

If PI>1 the weight will be divided by the performance index

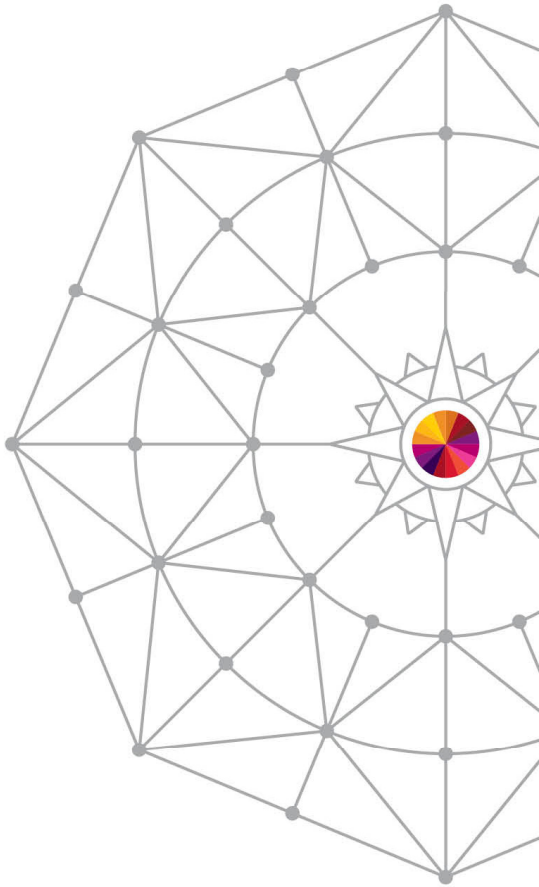
- with default IEAOPT RTPIFACTOR = 100

SYS	Avail Cap	Orig. Server weight	PI	**WLM weight**
SYS1	110	18	1.3	**14**
SYS2	100	16	0.8	**16**
SYS3	95	15	1.0	**15**
SYS4	95	15	2.0	**8**
Total		64		53



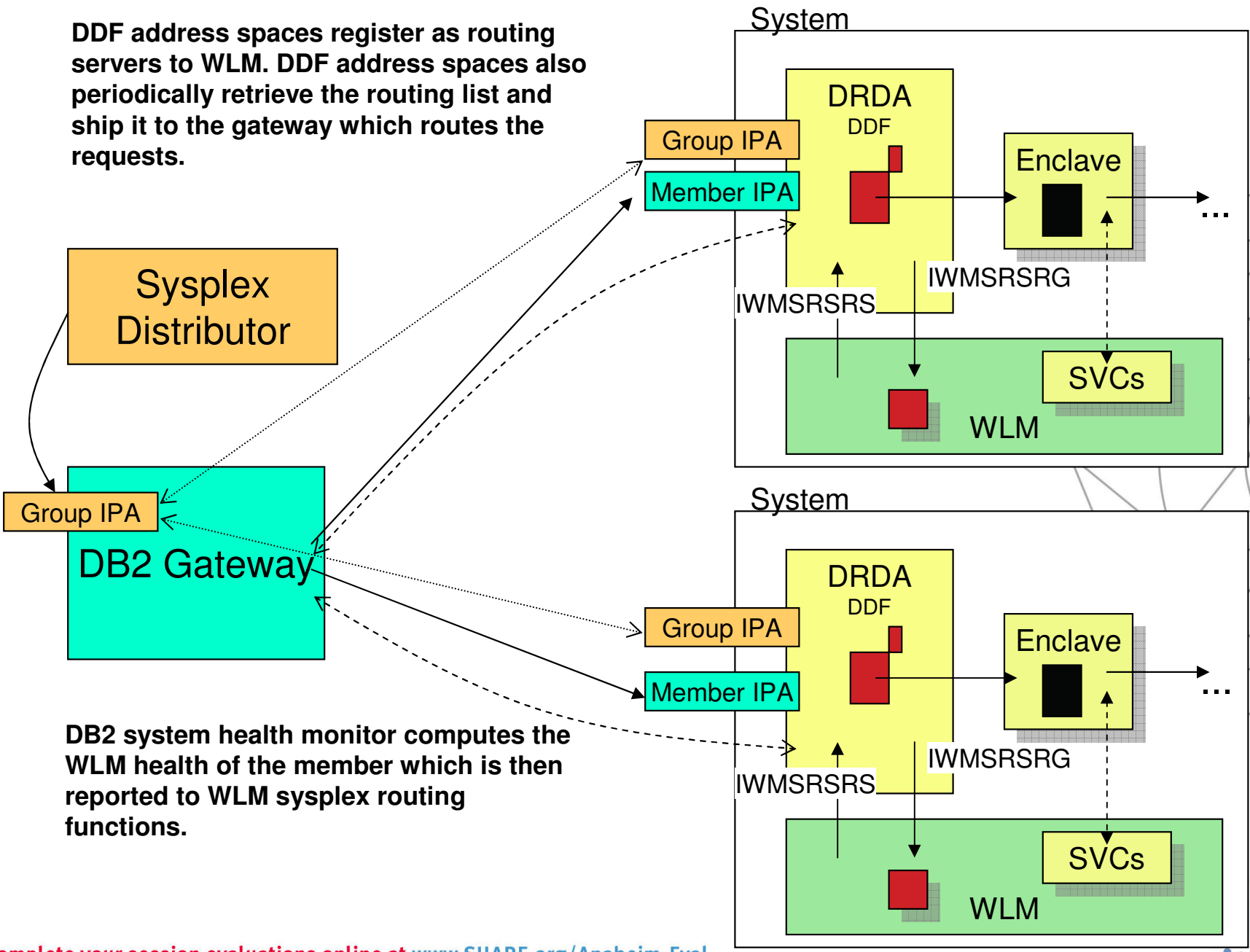
Health indicator effect on routing weight

- A health indicator can be set per server address space
- Health=100 is default and remains in effect until a different value is set via IWM4HLTH
  - Each IWM4HLTH invocation replaces previous health indicator values
- If the health indicator of a server is <100 its capability is reduced
- The server weight will be reduced by applying a factor of health/100



Background: Routing Services: DB2

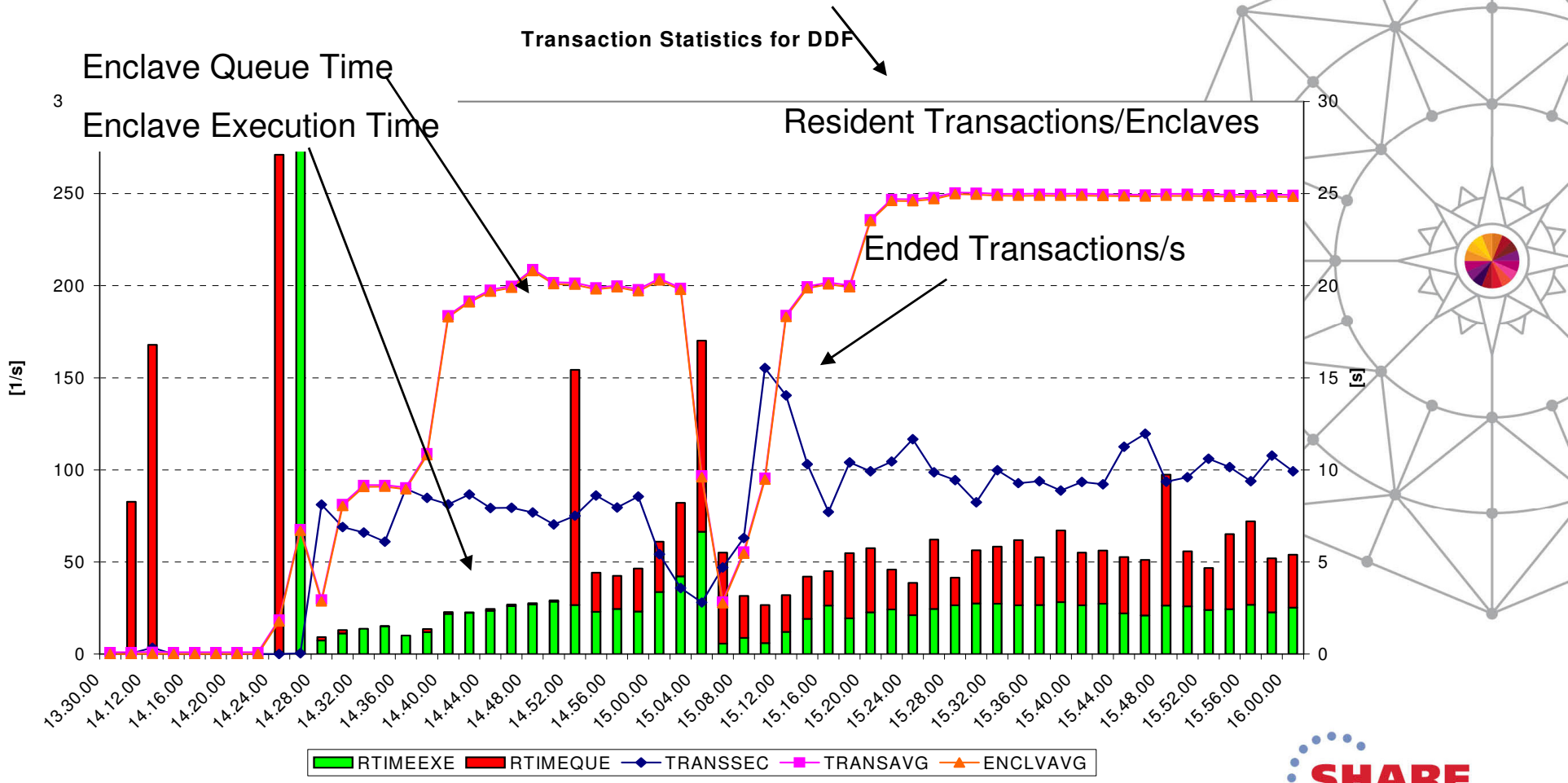
DDF address spaces register as routing servers to WLM. DDF address spaces also periodically retrieve the routing list and ship it to the gateway which routes the requests.



DB2 system health monitor computes the WLM health of the member which is then reported to WLM sysplex routing functions.

Sysplex Routing for DB2: Example Queue Time Ratio

- Servers with a better enclave queue time : execution time ratio will be favored
  - Server weight reduced by factor execution time / (execution time+ queue time)
- Only effective if DB2 is configured with "DDF Threads" INACTIVE

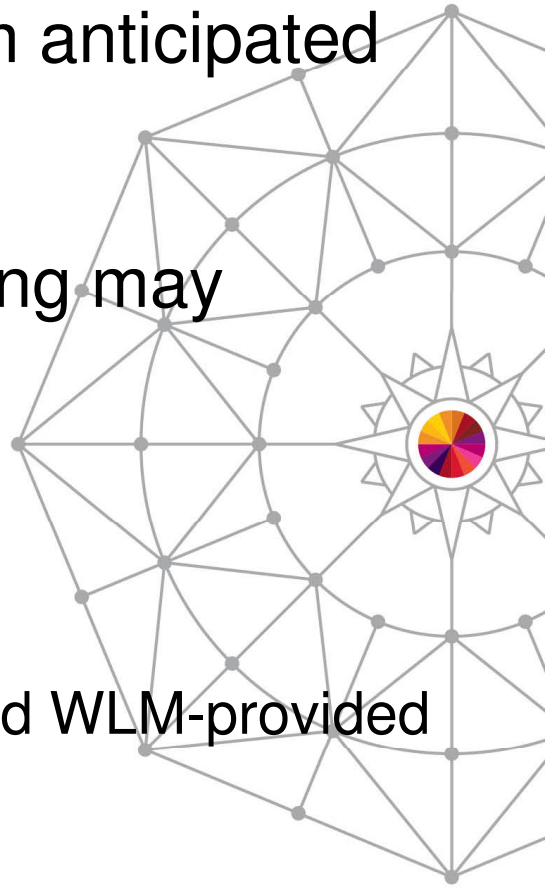


Agenda

- Concepts
  - Importance levels
  - Displaceable capacity
  - Free capacity
- WLM Sysplex Routing Services
  - IWMWSYSQ
  - IWMSRSRS
  - IWM4SRSC
  - Basic capacity-based weights and additional influencers
- Observations, best practices and optimization approaches

Understanding routing behavior

- Actual workload distribution may deviate from anticipated or warranted distribution
- Understanding and optimizing workload routing may require skills from multiple domains:
  - Applications
  - Subsystems involved
  - Routing product & configuration
    - Routing provide usually commands to understand WLM-provided weights and overrides
    - First step to understand raw WLM weights
    - Most routing services parameters are specified here
  - LPAR configuration & WLM



Drill-down into balancing issues

- Use routing component commands to understand WLM recommendations vs. routed work
- What routing mechanism is being used?
- Understand impact due to
  - Capacity
  - Performance Index
  - Health

  Use CPU activity report and Workload activity reports to understand LPAR/CEC configuration, load and performance index
  - RMF Mon III data can provide better granularity

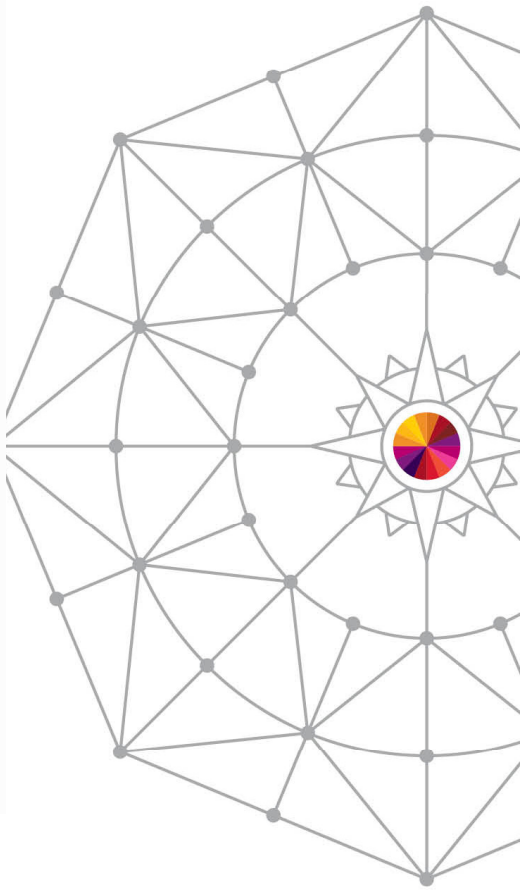
NETSTAT –O

```
$ netstat -O -
   P15150


MVS TCP/IP NETSTAT CS V1R12       TCPIP
   Name: TCPIP           10:31:18
Dynamic VIPA Destination Port Table for
   TCP/IP stacks:
Dest:        …..15150

  DestXCF:    …

  TotalConn: 0000059767  Rdy: 001
   WLM: 12  TSR: 100

  DistMethod: ServerWLM
```

- *The WLM weight in this summary display is derived by the weight value returned by IWM4SRSC (ServerWLM)*
  - *However, it has been post processed by Sysplex Distributor*
    - *Potentially reduced based on a number of health factors and*
    - *Normalized (divided by 4 to yield a value between 0-16 vs 0-64).*
  - *This value is what SD will use for load balancing and can be compared to the values of the other targets*

NETSTAT VIPADCFG DETAIL

```
VIPA Distribute:
 IP Address        Port   XCF Address       SysPt  TimAff  Flg
 ----------        ----   -----------       -----  ------  ----
 201.2.10.11       n/a    ALL               Yes    200     R
   DistMethod: Roundrobin
   OptLoc: No
 201.2.10.13       243    ALL               No     No      0
   DistMethod: BaseWLM
   OptLoc: 1
   ProcType:
     CP: 60  zAAP: 00  zIIP: 40
 201.2.10.14       243    ALL               No     No      1
   DistMethod: ServerWLM
   OptLoc: No
   ProcXCost:
     zAAP: 003  zIIP: 001
   ILWeighting: 1
```



DDF DISPLAY Command

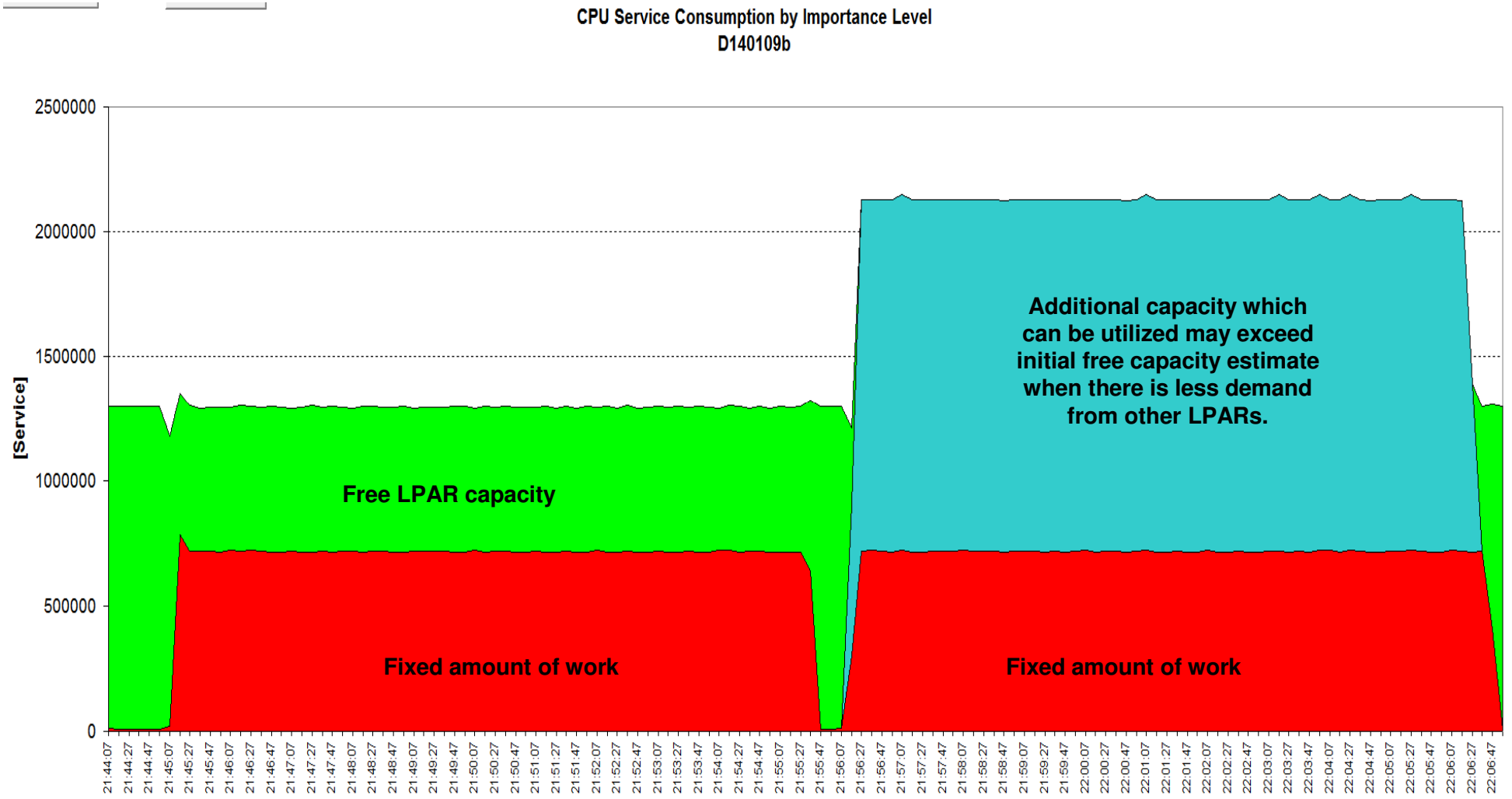
- -DIS DDF [DETAIL] returns WLM weight information
  - The following server list entry information is displayed for each DDF location that registered to WLM as part of the data sharing group:
  - DSNL100I LOCATION SERVER LIST: DSNL101I WT IPADDR IPADDR DSNL102I *weight ipv4-address ipv6-address*

```
-DISPLAY DDF DETAIL
```

With the DETAIL option, the following additional information is included in the output:

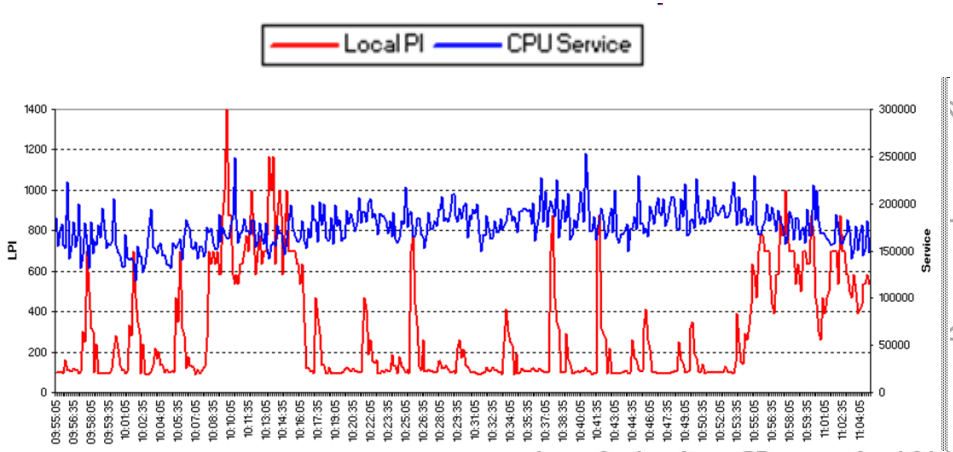
```
DSNL090I DT=      A CONDBAT=  64 MDBAT=   64
DSNL092I ADBAT=   1 QUEDBAT=   0 INADBAT=  0 CONQUED=      0
DSNL093I DSCDBAT= 0 INACONN=   0
DSNL100I     LOCATION SERVER LIST:
DSNL101I     WT IPADDR             IPADDR
DSNL102I     64 ::9.110.115.111    2002:91E:610:1::111
DSNL102I        ::9.110.115.112    2002:91E:610:1::112
DSNL099I DSNLTDDF DISPLAY DDF REPORT COMPLETE
```

Example: Initial Free LPAR capacity may be under-estimated



Performance Index (PI) effect on routing weight

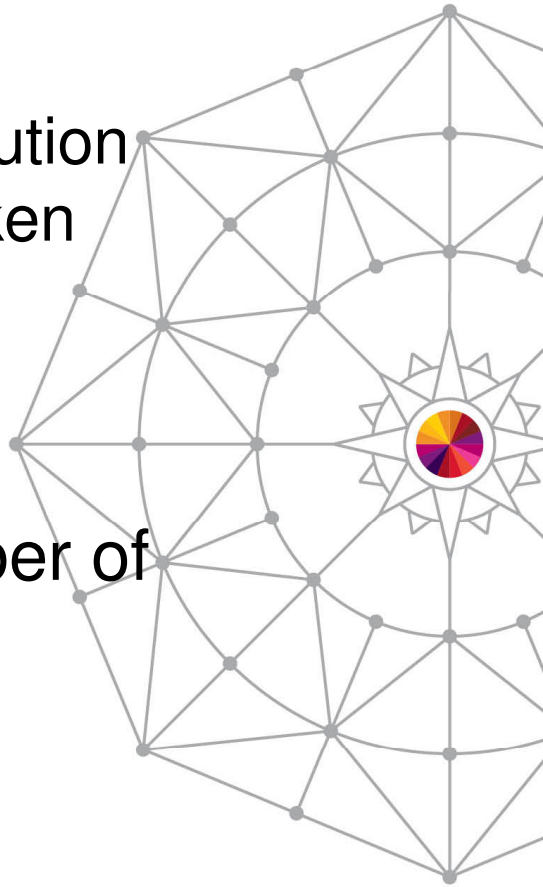
- Heavily fluctuating PI values can distort routing recommendations.
- In such cases it can be beneficial to scale back the impact of the PI via the IEAOPT RTPIFACTOR control.



- When RTPIFACTOR=0, the server weight is independent from the server PI
- When RTPIFACTOR=100 and server PI >1, the server weight is divided by the server PI.
- When 0<RTPIFACTOR<100 it results in a proportional influence of the server PI on the server weight.

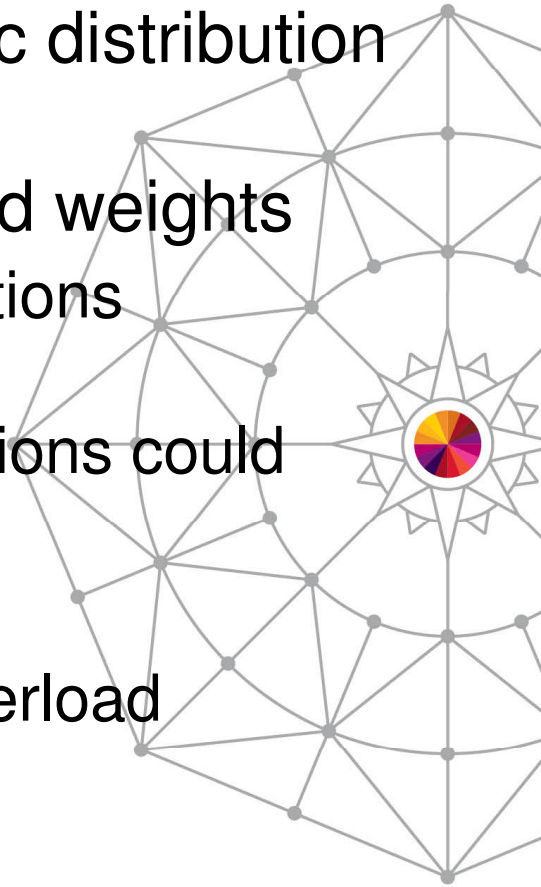
Observation: Connections vs. transaction routing

- Long living connections are… long living
  - May be established due to a given load distribution but not redistributed until connections are broken up and re-established
- The number of transactions routed to some systems may be *not proportional* to the number of connections that were established



Observation: Asymmetric configurations

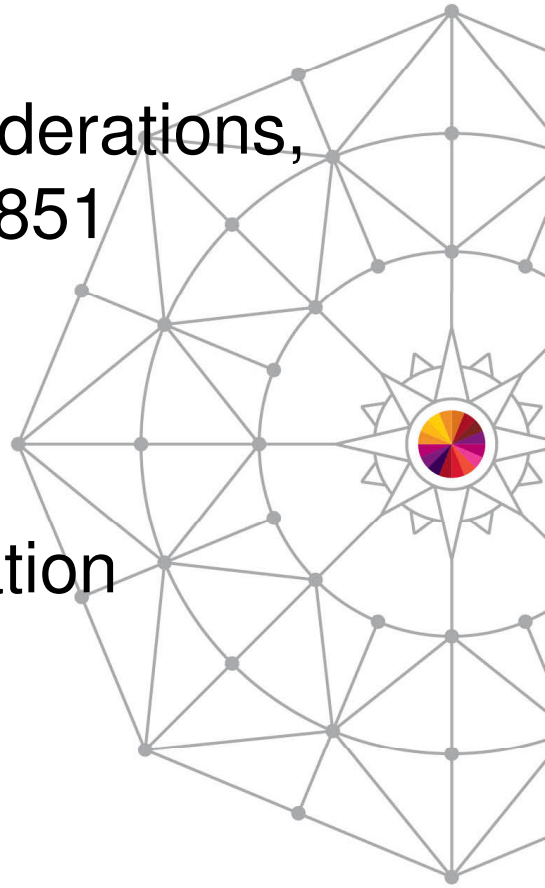
- Usually not a problem at all - unless a specific distribution is warranted
- Asymmetric configuration may result in biased weights
  - E.g. different weights, different CEC configurations
    - Consider zIIP, zAAP pools, too, when relevant
  - Depending on subsystems the routed transactions could deviate more
- Consider
  - SERVERWLM - if PI is a good indicator for overload
  - IL Weighting
    - IL weighting=1 is usually a good starting point
  - Round-robin or another, non-WLM based distribution method



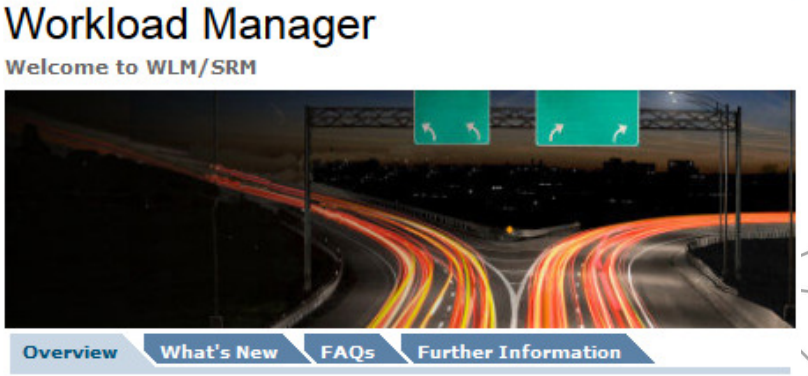
Sysplex Distributor and DB2 DDF - More Information -

- Gus Kassimis:
  Sysplex Networking Technologies and Considerations,
  SHARE in San Francisco, 2013, Session: 12851

- Jim Pickel:
  DB2 9 for z/OS Data Sharing: Distributed
  Load Balancing and Fault Tolerant Configuration
  http://www.redbooks.ibm.com/abstracts/redp4449.html



z/OS Workload Management - More Information -



- z/OS WLM Homepage:

  http://www.ibm.com/systems/z/os/zos/features/wlm/

- z/OS MVS documentation
  - z/OS MVS Planning: Workload Management: http://publibz.boulder.ibm.com/epubs/pdf/iea2w1c0.pdf
  - z/OS MVS Programming: Workload Management Services: http://publibz.boulder.ibm.com/epubs/pdf/iea2w2c0.pdf
- *IBM Redbooks publication*s:
  - System Programmer's Guide to: Workload Manager: http://publib-b.boulder.ibm.com/abstracts/sg246472.html?Open
  - ABCs of z/OS System Programming Volume 12 http://publib-b.boulder.ibm.com/abstracts/sg247621.html?Open

What is a DDF Transactions?

- ACTIVE MODE threads are treated as a single enclave from the time they are created until the time they are terminated. This means that the entire life of the database access thread is reported regardless of whether SQL work is actually being processed.
- INACTIVE MODE threads are treated differently. If the thread is always active, the duration of the thread is the duration of the enclave. When the thread is pooled, such as during think time, it is not using an enclave. In this case, inactive periods are not reported.