Capping Technologies and 4HRA Optimization

- Comparison of Hard and Soft-capping Controls
- Capping Implementation
- Defined Capacity and Group Capacity Management using z/OS Capacity Provisioning

Horst Sinram, STSM, z/OS Workload and Capacity Management, , sinram@de.ibm.com
IBM Germany Research & Development
16 Apr 2016

IBM

Trademarks

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

AIX*	DFSMSdfp	DS8000*	IBM*	Language Environment*	REXX	System z9*	z/Architecture*
BladeCenter*	DFSMSdss	Easy Tier	IBM eServer	MQSeries*	RMF	System z10	zEnterprise*
CICS*	DFSMShsm	FICON*	IBM logo*	MVS	SYSREXX	Tivoli*	z/OS*
DataPower*	DFSMSrmm	FlashCopy*	IMS	Parallel Sysplex*	System Storage	WebSphere*	
DB2*	DFSORT	HiperSockets	InfinBand*	PR/SM	System x*	z10 BC	
DFSMS	DS6000*	HyperSwap*	Infoprint*	RACF*	System z*	z10 EC	

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.

ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.

Java and all Java based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

OpenStack is a trademark of OpenStack LLC. The OpenStack trademark policy is available on the OpenStack website.

TEALEAF is a registered trademark of Tealeaf, an IBM Company.

Windows Server and the Windows logo are trademarks of the Microsoft group of countries.

Worklight is a trademark or registered trademark of Worklight, an IBM Company.

UNIX is a registered trademark of The Open Group in the United States and other countries.

* Other product and service names might be trademarks of IBM or other companies.

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 information provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g, zIIPs, zAAPs, and IFLs) ("SEs"). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the “Authorized Use Table for IBM Machines” provided at www.ibm.com/systems/support/machine_warranties/machine_code/aut.html (“AUT”). No other workload processing is authorized for execution on an SE. IBM offers SE at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.

Notice Regarding Specialty Engines (e.g., zIIPs, zAAPs and IFLs):

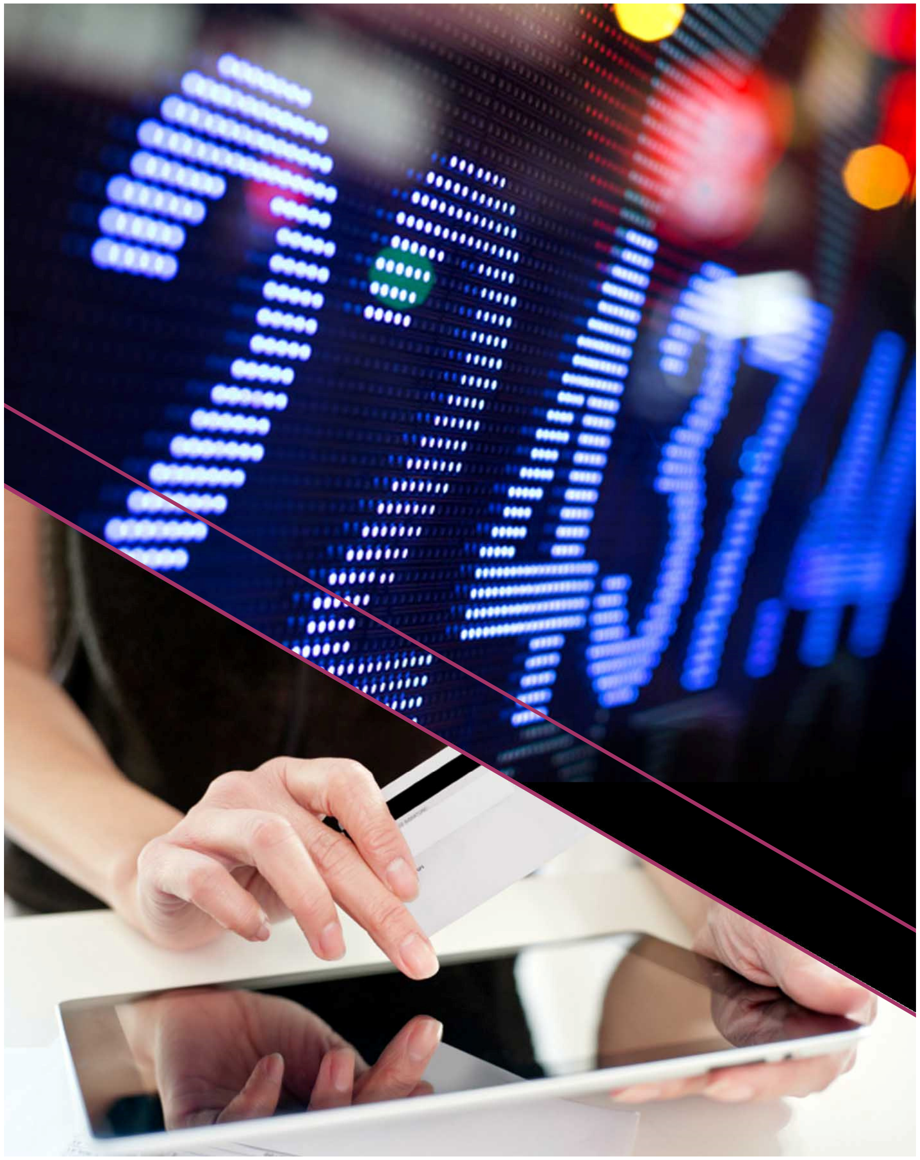
Any information contained in this document regarding Specialty Engines ("SEs") and SE eligible workloads provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g., zIIPs, zAAPs, and IFLs). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the "Authorized Use Table for IBM Machines" provided at www.ibm.com/systems/support/machine_warranties/machine_code/aut.html ("AUT").

No other workload processing is authorized for execution on an SE.

IBM offers SEs at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.

Agenda

- **Overview of capping types**
- Initial capping
- Absolute capping
- Defined capacity & group capacity
- Resource group capping
- 4HRA management
- Additional Material



Reasons you would consider capping techniques…

Technical motivation

- Protect/isolate LPARs against other LPARs, e.g. multi-tenancy
- Influence capacity-based workload routing
- Guarantee unused CPC processor capacity
- Protect workloads (sets of service classes) against other workloads

Financial motivation

- Limit software cost
  - Guaranteed capacity limit for one or more LPARs
  - Four hour rolling average (4HRA) consumption

- Possible impact of capping needs to be monitored and accepted
- Cap limits should be adjusted as appropriate
  - Watch your SLAs

Comparison of capping types

➡ New capping types (spring 2016)

Type of capping	Scope	Specification unit	Proc types	Stability of limit under configuration changes	Suitable to isolate LPARs or LPAR groups	Control point
Initial (hard) capping	LPAR	LPAR share of CPC capacity	Any	–	+	SE/HMC
LPAR Absolute capping (zEC12 GA2 and later)	LPAR	Fractional #processors	Any	O	+	SE/HMC
➡ LPAR Group Absolute Capping (z13 GA2 and later)	Group of LPARs	Fractional #processors	Any	O	+	SE/HMC
Defined capacity (DC, soft capping)	LPAR	MSU (4HRA)	CP	+	–	SE/HMC
LPAR group capacity (GC, soft capping)	Group of LPARs	MSU (4HRA)	CP	+	–	SE/HMC
➡ Absolute MSU Capping	LPAR or Group	MSU	CP	+	+ (CP only)	SE/HMC + IEAOPT
Resource group capping	Groups of service classes in Sysplex or per LPAR	Unweighted CPU SU/sec, fraction of LPAR share, or fractional #CPs	CP*	+	N/A	WLM Policy
Logical configuration	LPAR	Integer #processors	Any	O	+ but coarse grain	HMC+OS

PR/SM controlled — WLM controlled, PR/SM enforced — WLM controlled

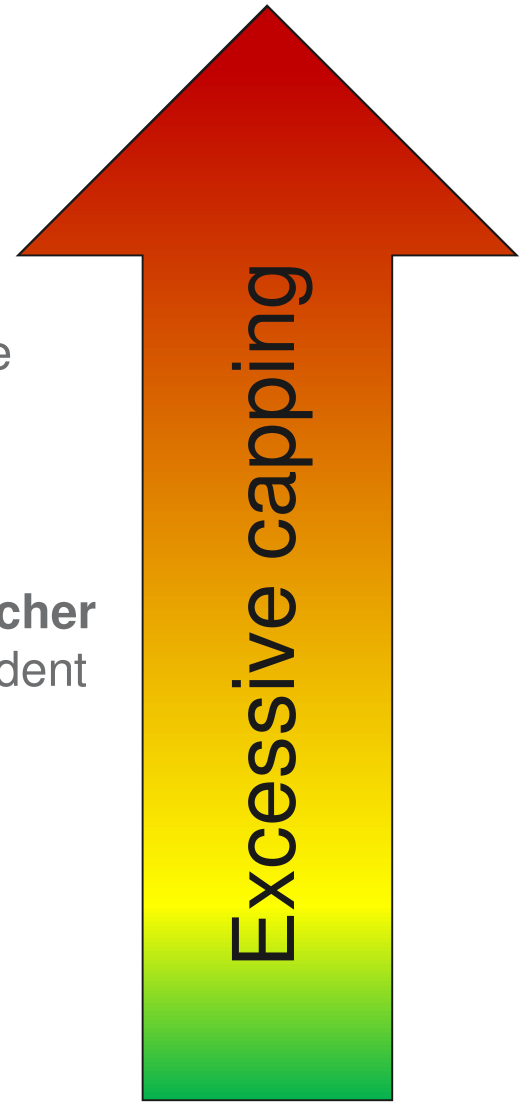
Which capping techniques may be combined?

Type of capping➔	Initial (hard capping)	LPAR Absolute capping	LPAR Absolute group capping	Defined capacity [1]	LPAR group capacity [1]	Resource group capping
Initial (hard capping)		+	+	–	–	+
LPAR Absolute capping	+		+	+	+	+
LPAR Group Absolute capping	+	+		+	+	+
Defined capacity [1]	–	+	+		+	+
LPAR group capacity [1]	–	+	+	+		+
Resource group capping	+	+	+	+	+	

(1) Includes ABSMSUCAPPING=NO and ABSMSUCAPPING=YES

Possible impacts of (excessive) capping

- **Sysplex / multi system outage**
  - E.g. for LOCKs or RESERVEs not being freed timely
- **System outage**
  - E.g. for resources not being freed timely
  - Storage shortages
  - Work (e.g. SRBs) backed up, common storage shortage
- **Important work displaced**
- **Service levels missed**
- **Contention and increased promotion by SRM or dispatcher**
  - Can be unproblematic if displaced work is truly independent from important work – no shared resources
- **Less important work displaced**
- **Goals missed**
- **Increased response times**
- **Increased CPU delays**



Agenda

- Overview of capping types
- **Initial capping**
- Absolute capping
  - H/W absolute capping
  - WLM Absolute capping
- Defined capacity & group capacity
- Resource group capping
- 4HRA management
- Additional Material

Initial capping (aka “hard capping”)

- Defined to PR/SM per processor type. Managed by PR/SM through limiting the processor time available to the LP’s logical processors
- The LPAR capacity is capped to LPAR share of CPC shared capacity

$$\text{LPAR}_i \text{ share} = \frac{\text{Weight}_i}{\sum\limits_{\text{All activated LPARs}} \text{Weight}_j}$$

- LPAR weight is distributed across online CPs of the given type
- With HiperDispatch=NO an LP’s share is divided by the number of online logical CPs
  - Capping is done on a logical CP basis.
    May result in over capping if not all LCPs can be utilized
  - Consider following example:
    zEC12-732, 10 CPs online, Share=5.6%, low CPC utilization
    Workload: 2 TCBs

Initial Capping with HiperDispatch=Yes vs. No

CPU Activity Reports

- CPU 2827 CPC CAPACITY 3665 SEQUENCE CODE 00000000000
- MODEL 732 CHANGE REASON=NONE **HIPERDISPATCH=YES**
- H/W MODEL H43

---CPU---		---------------- TIME % ----------------				LOG PROC	
NUM	TYPE	ONLINE	LPAR BUSY	MVS BUSY	PARKED	SHARE %	
0	CP	100.00	89.12	97.67	0.00	100.0	HIGH
1	CP	100.00	87.50	97.83	0.00	80.4	MED
2	CP	100.00	2.51	82.33	96.54	0.0	LOW
3	CP	100.00	1.87	63.68	96.54	0.0	LOW
4	CP	100.00	0.01	-----	100.00	0.0	LOW
5	CP	100.00	0.01	-----	100.00	0.0	LOW
6	CP	100.00	0.01	-----	100.00	0.0	LOW
7	CP	100.00	0.01	-----	100.00	0.0	LOW
A	CP	100.00	0.01	-----	100.00	0.0	LOW
B	CP	100.00	0.01	-----	100.00	0.0	LOW
TOTAL/AVERAGE			18.10	96.92		180.4	

- MODEL 732 CHANGE REASON=NONE **HIPERDISPATCH=NO**
- H/W MODEL H43

---CPU---		---------------- TIME % ----------------				LOG PROC
NUM	TYPE	ONLINE	LPAR BUSY	MVS BUSY	PARKED	SHARE %
0	CP	100.00	14.61	54.28	------	18.0
1	CP	100.00	13.00	46.80	------	18.0
2	CP	100.00	10.71	31.82	------	18.0
3	CP	100.00	6.77	18.55	------	18.0
4	CP	100.00	4.22	6.44	------	18.0
5	CP	100.00	4.87	13.16	------	18.0
6	CP	100.00	1.75	2.72	------	18.0
7	CP	100.00	4.54	13.05	------	18.0
A	CP	100.00	4.02	10.40	------	18.0
B	CP	100.00	3.08	6.88	------	18.0
TOTAL/AVERAGE			6.76	20.41		180.0

With HiperDispatch=Yes the high/medium processors receive a higher processor share.

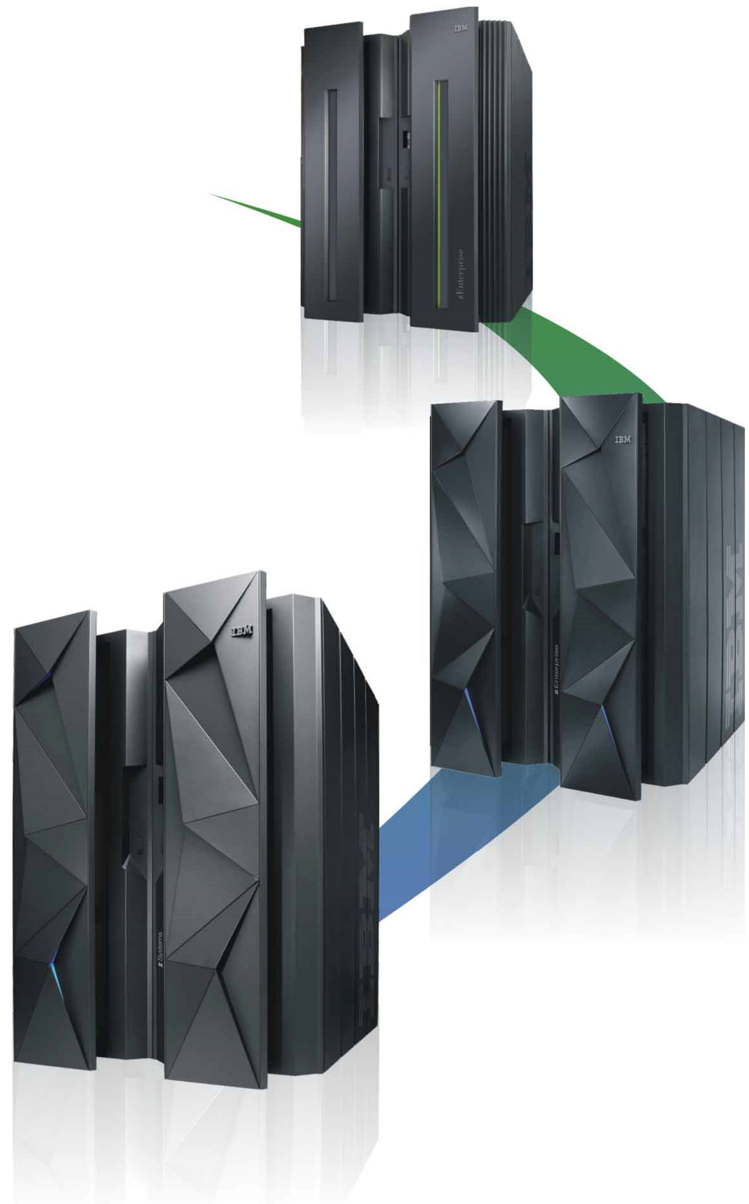
Stability of initial cap limits

- The **MSU equivalent** for an initial cap limit changes when…
  - The initial weight of the capped LPAR is changed
  - LPARs are de/activated or the total weight changes due to initial weight changes
  - Temporary capacity is de/activated
    - CBU, On/Off CoD…
- May require manual intervention when
  - A particular MSU/MIPS number is guaranteed for an LPAR
  - A particular MSU number must not be exceeded for licensing reasons

Agenda

- Overview of capping types
- Initial capping
- Absolute capping
  - **H/W absolute capping**
  - WLM Absolute capping
- Defined capacity & group capacity
- Resource group capping
- 4HRA management

- Additional Material



LPAR Absolute Capping and Group Capping

- We use “Absolute” to refer to capping independently of the four hour rolling average consumption
- Defined to PR/SM per processor type. Managed by PR/SM through limiting the number of PR/SM time slices available to the LPAR’s logical processors
- Specification in terms of (fractional) number of processors per processor type
  - E.g., 3.75 CPs
- LPAR absolute capping introduced with zEC12 GA2
  LPAR absolute group capping introduced with z13 GA2
- Primarily intended for non z/OS images
  - Not capping to a MSU figure, not aware of 4h rolling average consumption
- Can be specified independently from the LPAR weight
  - But recommended to specify absolute cap above weight
  - WLM algorithms consider weight

Absolute Capping Limit

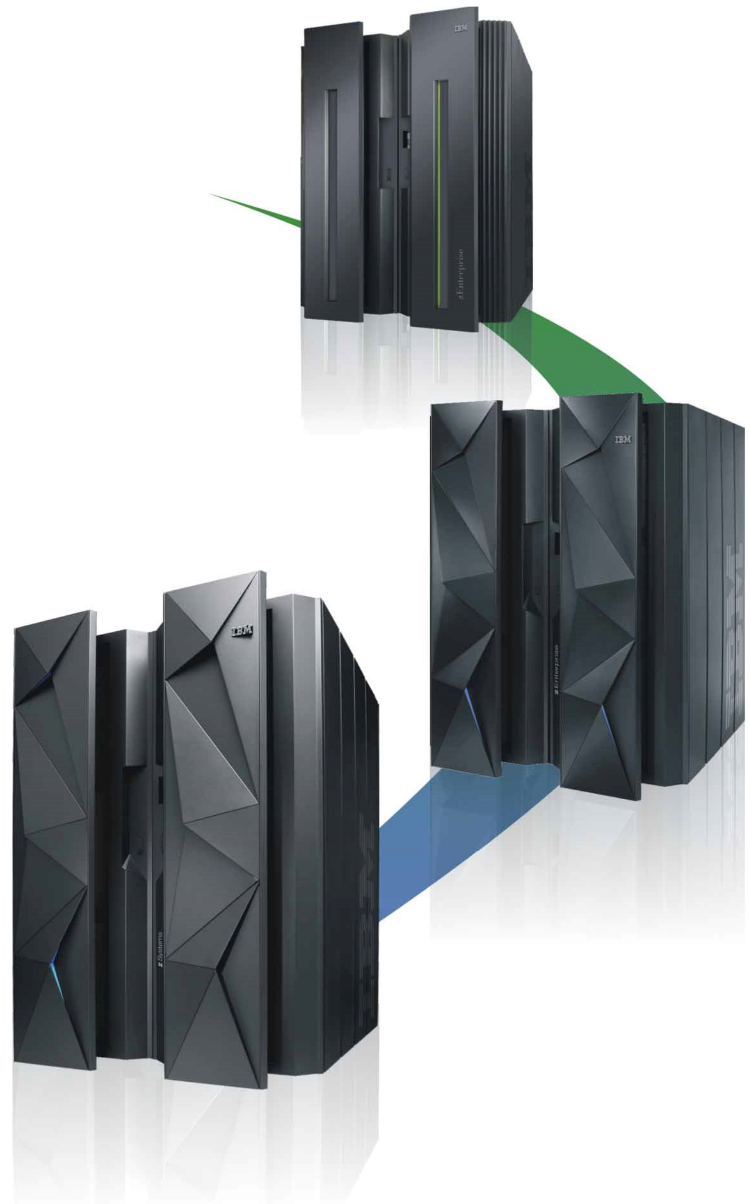
- Absolute capping may be used *concurrently* with defined capacity and group capacity management
  - The minimum of all limits becomes effective.
  - WLM/SRM is aware of the absolute cap, e.g. for routing decisions.
  - Partition capacity RCTIMGWU =
    MIN (absolute cap, absolute group cap,
    defined capacity, group cap,…)
    when all capping types are in effect
    - RMF provides RCTIMGWU in SMF70WLA
    - In addition, SMF70HW_Cap_Limit value in hundredths of CPUs

Stability of H/W absolute cap limits

- The effective limit for an absolute cap can change significantly when
  - the absolute cap value of the capped LPAR is changed, or
  - temporary capacity is de/activated AND the capacity level (processor speed) changes
    - I.e., general purpose processor CBU, On/Off CoD to/from subcapacity models
- But: the effective MSU rating for an absolute cap changes <u>also</u> when just the number of physical processors changes
  - I.e. CBU, On/Off CoD to/from within same capacity level, such as 7xx
  - If this effect is not desired, WLM absolute capping ca be considered

Agenda

- Overview of capping types
- Initial capping
- Absolute capping
  - H/W absolute capping
  - **WLM Absolute capping**
- Defined capacity & group capacity
- Resource group capping
- 4HRA management
- Additional Material



WLM Absolute capping (Permanent capping)

z/OS release / *Function*	*V2.2*	*V2.1*
z13 GA2 LPAR Absolute group capping	***OA47752***	***OA47752***
Absolute MSU capping	***OA49201***	***OA49201***

- Function of WLM provided by APAR OA49201
- Technically, based on WLM defined capacity capacity or group capacity
  - BUT: LPAR **will always be capped**, independent of 4 hour rolling average consumption.
    - Only general purpose processor
  - Same underlying mechanisms
- Specified in IEAOPTxx.

  Limit is the LPAR defined capacity or group capacity specified on the HMC **in MSU**.
- z/OS Capacity Provisioning will not manage partitions capped through WLM absolute capping
  - Provisioning Manager commands can be used when limit changes desired

Using absolute MSU capping

IEAOPTxx ABSMSUCAPPING=	
NO	Defined capacity limits and group capacity limits should be enforced only while the long term four hour rolling average consumption exceeds the respective limit (existing and usually desired behavior).
YES	Defined capacity limits and group capacity limit should be enforced **permanently, independently of the long term four hour rolling average consumption**.

- AbsMSUcapping=Yes limits LPAR consumption to a certain MSU number at all times.
  - I.e., the system loses the flexibility of consuming above the defined capacity limit while the four hour rolling average is below the limit.
- Limit remains stable even when CEC configuration changes, e.g. through On/Off CoD or CBU activations or deactivations.
  - Absolute MSU capping is an effective means to permanently limit the consumption of an LPAR to a specific MSU figure at all times
    - including times when the *four-hour rolling average* does not exceed the defined limit.
- The Capacity Provisioning Manager (CPM) will not change limits of AbsMSUCapping=YES systems.

Using absolute MSU capping with group capacity

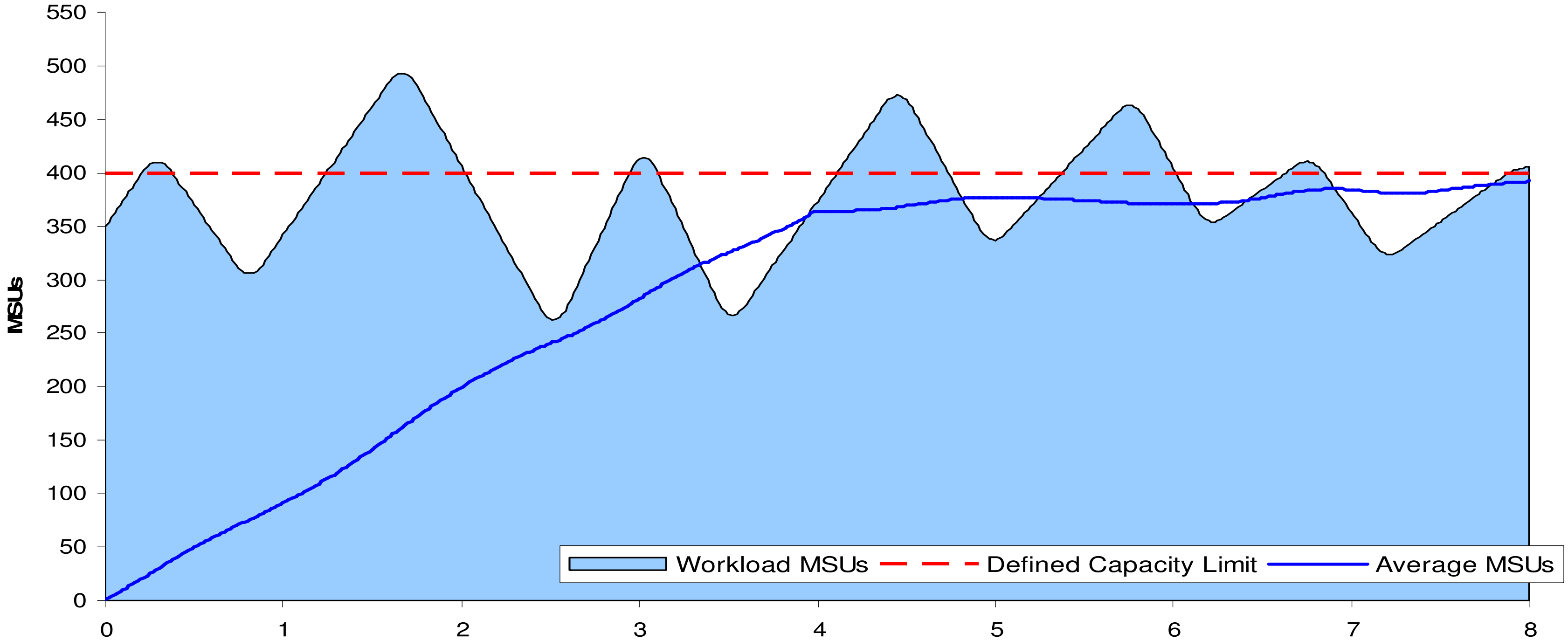
- When used with an LPAR capacity group:
  - Limit on behalf of the group entitlement will always be enforced
    - Regardless of the *four-hour rolling group average* consumption.
  - As with AbsMSUcapping=NO, an LPAR is allowed to take benefit of the unused group capacity
    - Unless the LPAR is also capped via other LPAR limits.
  - All members of a capacity group that use AbsMSUcapping=YES will permanently enforce the limit on behalf of the capacity group.
  - All members of a capacity group that do *not* use AbsMSUcapping=YES will be capped while the group *four-hour rolling group average* consumption is greater or equal to the group limit

Agenda

- Overview of capping types
- Initial capping
- Absolute capping
- **Defined capacity & group capacity**
- Resource group capping
- 4HRA management

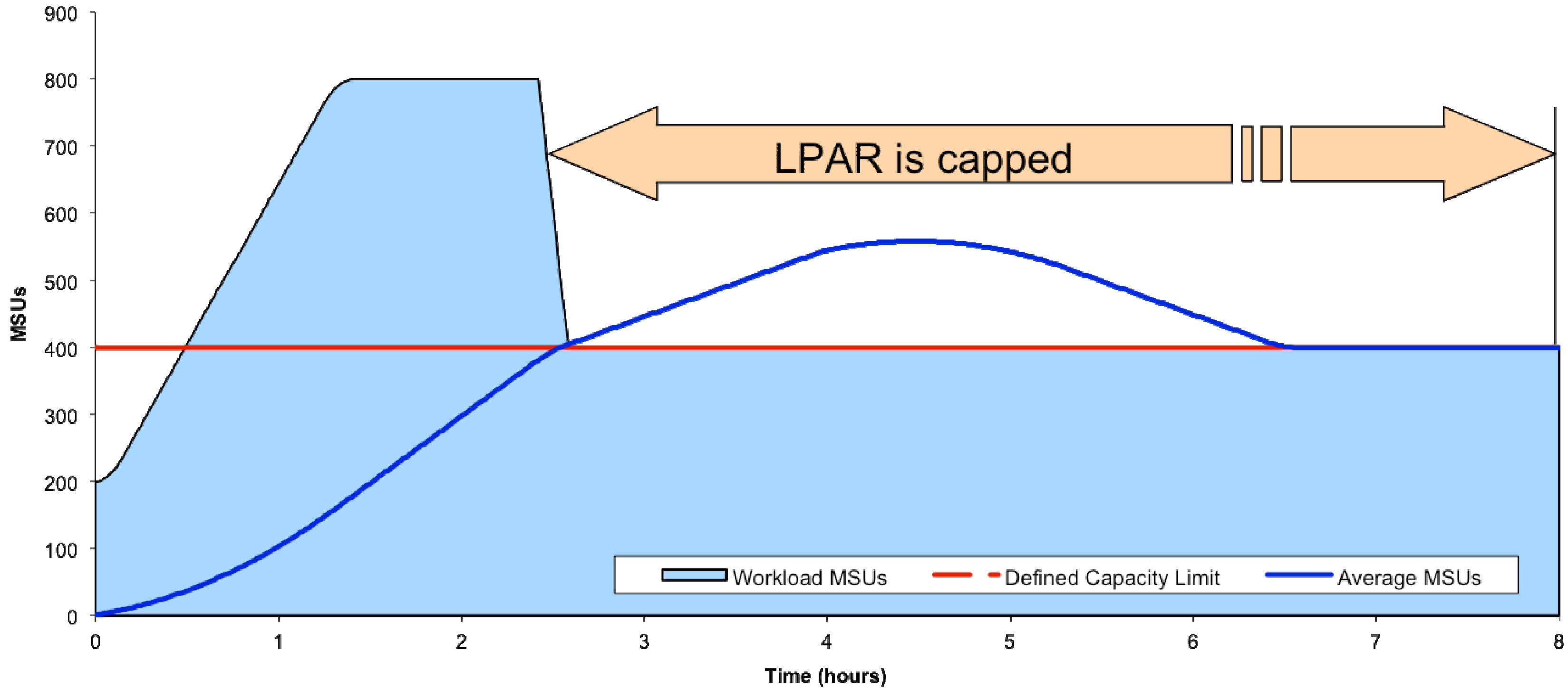
- Additional Material

4 Hour Rolling Average ("4HRA")



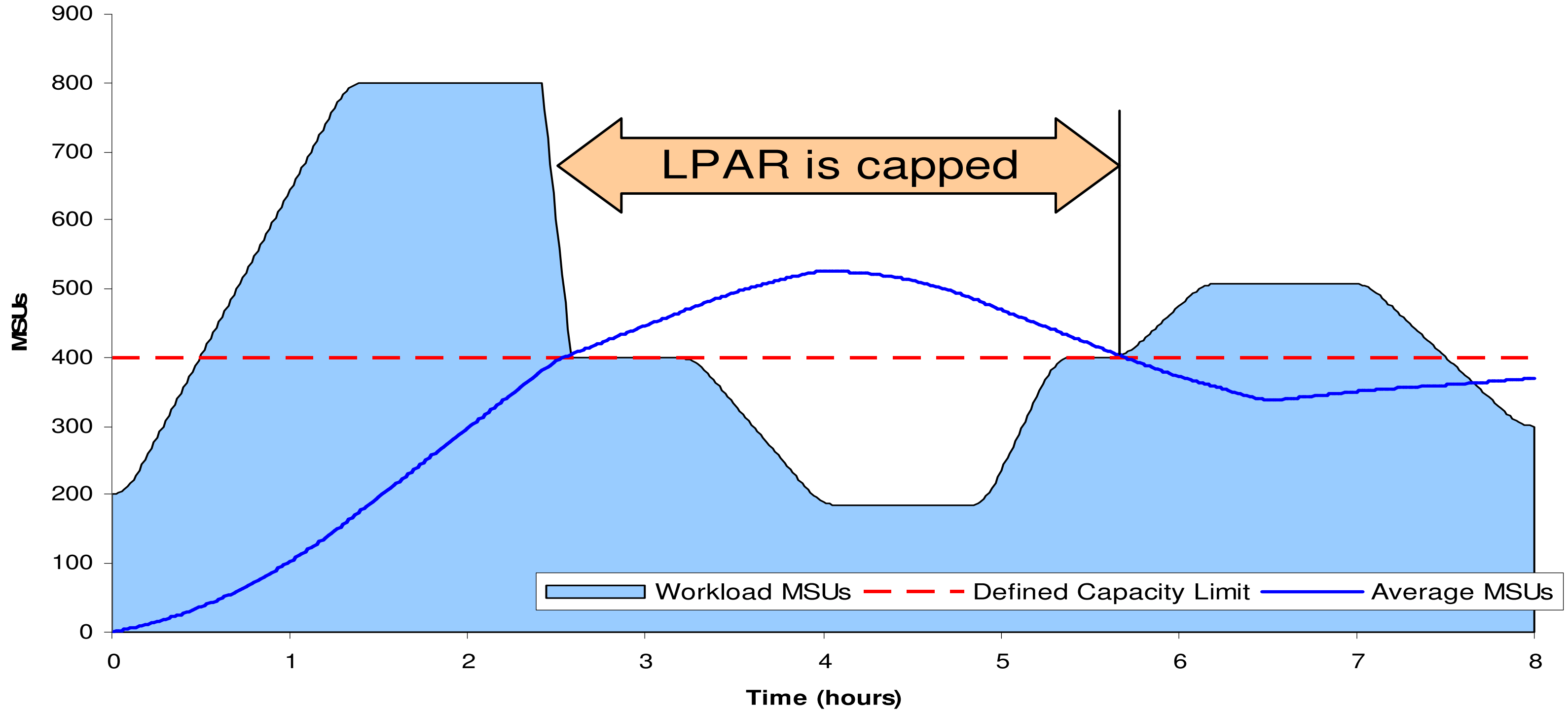
- Average consumption in LPAR over last 4h (rolling)
- ;easure in MSU ≡ "Million Service Units per hour"
  - ≠ Service Units • 3600 / 1000000
- Tracked as array of 48 intervals of 5 min = 4h

LPAR Capping



- An LPAR is -soft- capped when the 4HRA exceeds the defined capacity limit
- It remains capped until the 4HRA is below the defined limit
- While capped, the consumption is limited to the defined limit
- WLM advises PR/SM how to cap the LPAR

End of capping phase



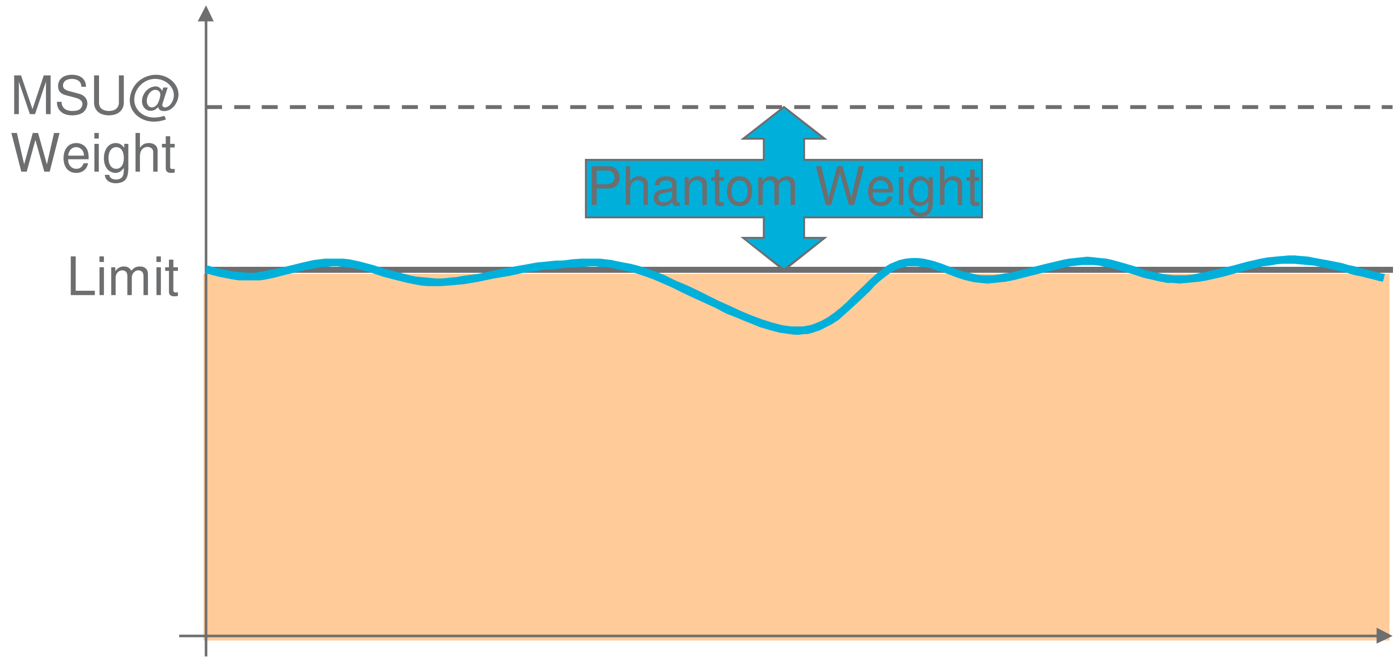
• Capping ends when the 4 hour average is below the softcap

Underlying soft capping techniques

- Historically, PR/SM algorithms were designed to cap a partition at its weight (hard capping)
- Therefore, WLM and PR/SM use particular interfaces to cap a partition to an arbitrary MSU figure

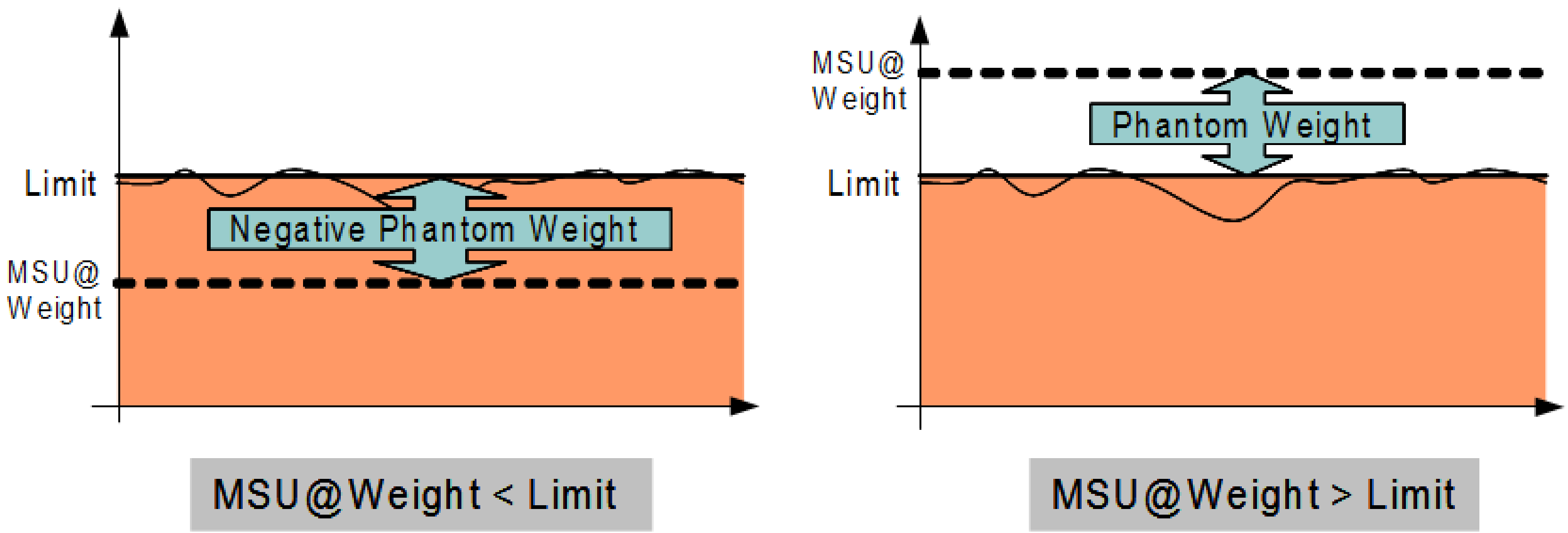
Weight vs. defined capacity limit	Hardware/Software level	Selected capping technique
MSU@weight > MSU imit	Any	**Phantom weight**
MSU@weight ≤ MSU limit	zEC12 GA2 and z/OS V2.1 or later	**Negative phantom weight**
	Other	Pattern capping

Phantom weight



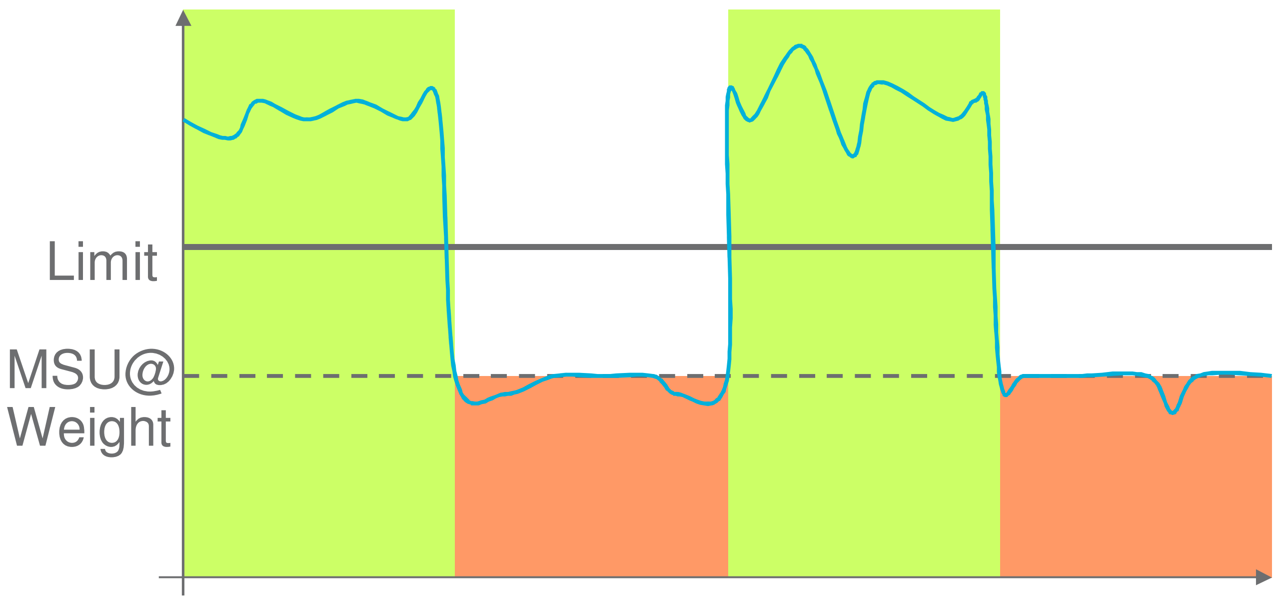
- Phantom weight is used to modify the PR/SM share of an LPAR
- WLM does not change a phantom weight as long as the limit and configuration do not change
  ➔smooth capping

Capping with phantom weight



- zEC12 with z/OS V2.1 and above support not only positive but also negative phantom weights.
  - Note: While a positive phantom weight changes the PR/SM entitlement of a partition, a negative phantom does not elevate the PR/SM dispatching priority.
    ➔ Only the capacity defined by the weight is guaranteed.

Cap pattern (only used with pre-EC12 GA2 H/W or older software levels)



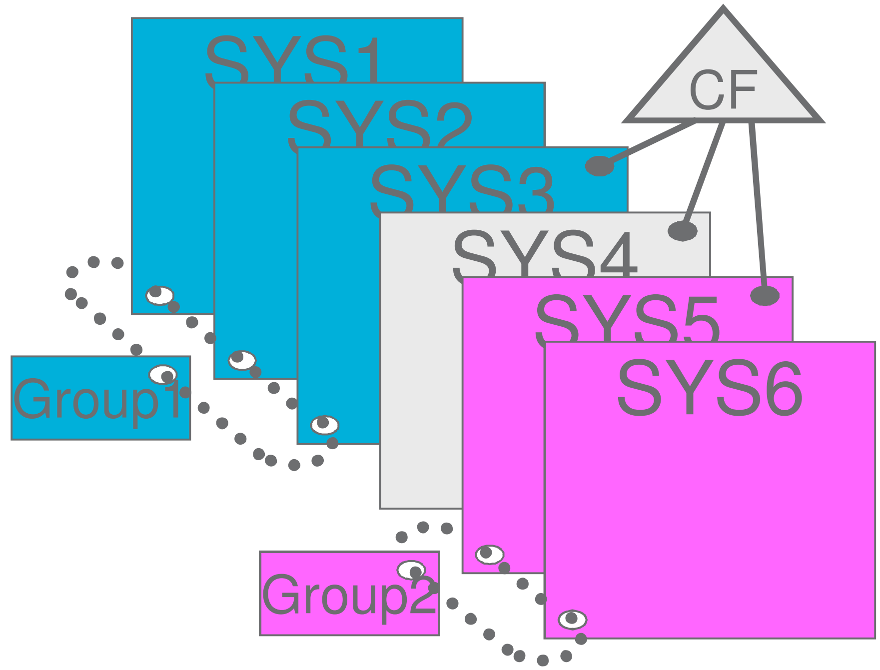
Prior to negative phantom weights WLM set up a cap pattern:
Alternating periods of

- LP capped to MSU@Weight, and
- LP uncapped

On average the MSU limit is enforced
…but interactive workloads can experience “pulsing”

Group Capping

An LPAR capacity group can be used to enforce a MSU limit for a set of one or more LPARs.



- A capacity group is limited to a **single CPC** but independent from the Sysplex
- A system can be joined to one group at most
- A system will not join or will leave the capacity group when requirements not met
  - Namely, initial capping must not be active

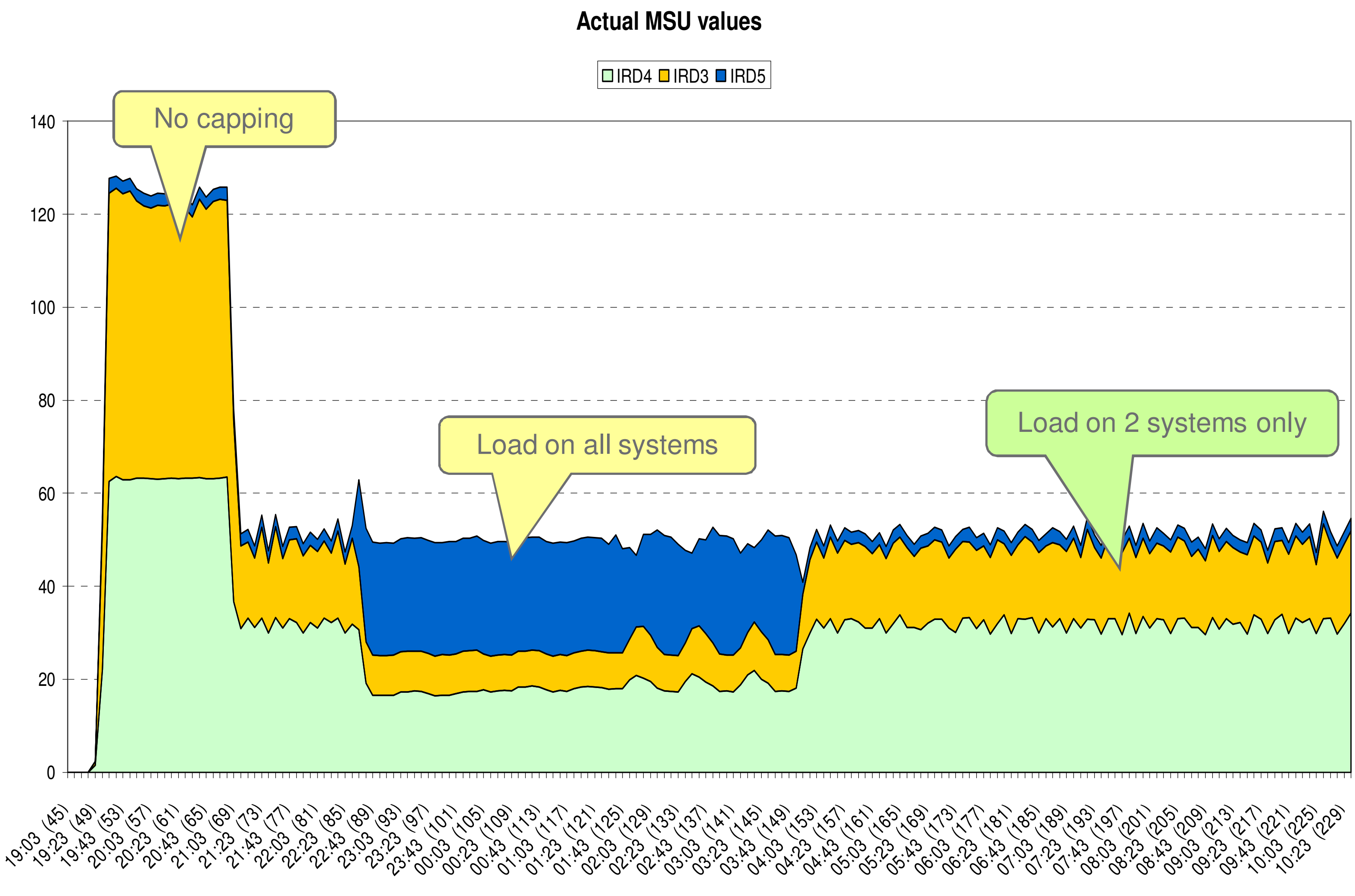
Group capping example

System	Weight	DC (MSU)	GC (MSU)	Initial GC Share (MSU)	Donation at full demand (MSU)	GC Entitle ment (MSU)
SYS1	600	-	400	200	–	240
SYS2	300	-		100	–	120
SYS3	300	40		100	60	40

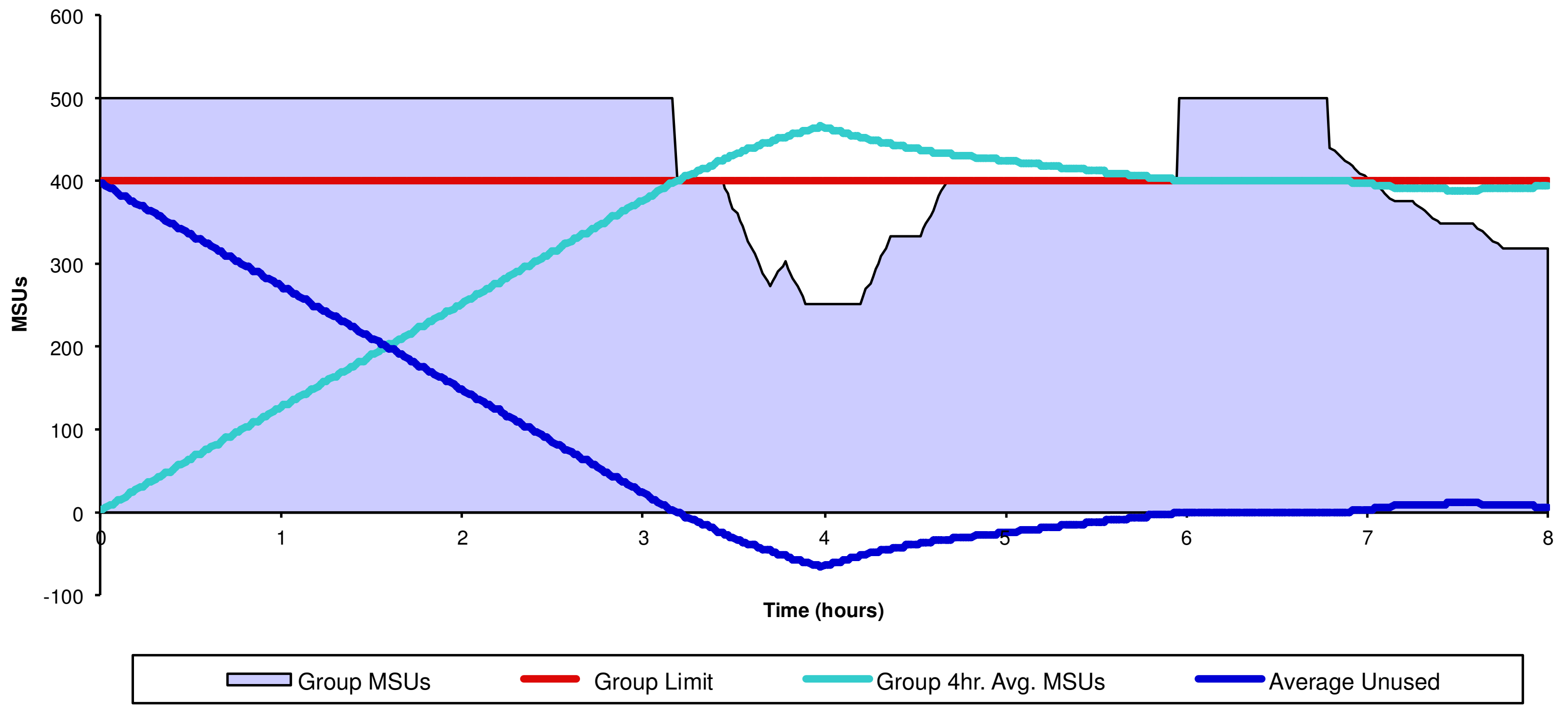
- The share of a group member is based on its *weight*
  - With IRD with zEC12 GA2 & z/OS V2.1: initial weight
  - With IRD in prior environments: current weight

- Unused capacity is donated to other group members
  - …and re-distributed based on weight
- The minimum of DC and GC entitlement is used for capping an LPAR

Group Capping behavior



Unused vector (group capping)



- Group capacity is tracked via an **un**used group capacity array of 48 intervals of 5 min
- Group capping is active when average unused group capacity negative
- Each system tracks unused capacity while joined to a capacity group
  - Not synchronized upon group changes: systems may have a different view for up to 4h

RMF: Partition Data Report

```
                              P A R T I T I O N   D A T A   R E P O R T
                                                                                            PAGE    2
        z/OS V1R12             SYSTEM ID SYS1           DATE 10/13/10           INTERVAL  14.59.678
                               RPT VERSION V1R12 RMF    TIME 09.30.00           CYCLE 1.000 SECONDS

MVS PARTITION NAME                     SYS1     NUMBER OF PHYSICAL PROCESSORS       9           GROUP NAME     N/A
IMAGE CAPACITY                          100               CP                        7                LIMIT     N/A
NUMBER OF CONFIGURED PARTITIONS           9               ICF                       2            AVAILABLE     N/A
WAIT COMPLETION                          NO
DISPATCH INTERVAL                   DYNAMIC

--------- PARTITION DATA -----------------  -- LOGICAL PARTITION PROCESSOR DATA --   -- AVERAGE PROCESSOR UTILIZATION PERCENTAGES -

--------- PARTITION DATA -----------------  PROCESSOR-    ----DISPATCH TIME DATA----   LOGICAL PROCESSORS   --- PHYSICAL PROCESSORS ---
                                            NUM   TYPE    EFFECTIVE        TOTAL      EFFECTIVE    TOTAL   LPAR MGMT  EFFECTIVE  TOTAL
                   ----MSU----  -CAPPING--
NAME        S   WGT  DEF    ACT  DEF   WLM%  1.2   CP     00.04.27.302  00.04.27.519    24.86      24.92     0.01       4.24      4.25
                                              4    CP     00.00.21.680  00.00.22.083     0.60       0.61     0.01       0.34      0.35
SYS1        A   20                                                                                         0.02       3.41      3.43
SYS2        A   1                                                                                          0.01      68.68     68.69
SYS3        A   10                                                                                         0.01      23.02     23.03
SYS4        A   300                                                                                        0.05                 0.05
SYS5        A   200                                                                                       ------     ------    ------
                                                                                                           0.11      99.69     99.80
CFC1        A   DED                                                                                        0.01      99.95     99.96
CFC2        A   DED                                                                                        0.00       0.00      0.00
*PHYSICAL*                                                                                                 0.03                 0.03
                                                                                                          ------     ------    ------
  TOTAL                                                                                                    0.04      99.95     99.99
```

--------- PARTITION DATA -----------------

NAME	S	WGT	----MSU---- DEF (1)	----MSU---- ACT (2)	-CAPPING-- DEF (3)	-CAPPING-- WLM% (4)
SYS1	A	20	100	10	NO	62.2
SYS2	A	1	0	1	YES	0.0
SYS3	A	10	5	8	NO	3.3
SYS4	A	300	95	155	NO	0.0
SYS5	A	200	50	52	NO	0.0

RMF: Partition Data Report

1. **MSU DEF** DC limit for this partition in MSU as specified on HMC
2. **MSU ACT** Actual avg. MSU consumption of this LPAR
3. **CAPPING DEF** Indicates whether this partition uses initial capping
4. **CAPPING WLM%** Portion of time the LPAR was capped during the RMF interval
   - Does not necessarily imply that the cap constrained the LPAR's consumption.

RMF: CPC Capacity

1. **CPC Capacity**
   Total capacity of the CPC in MSU/h

2. **Image Capacity**
   Maximum capacity available to this partition

3. **Weight % of Max**
   Average weighting factor relative to the maximum defined weight for this partition.

4. **WLM Capping %**
   Percentage of time that WLM had advised PR/SM to cap the LPAR

5. **4h Avg**
   Average consumed MSU/h during the last 4 hours
6. **4h Max**
   Maximum consumed MSUs during the last 4 hours

RMF: Group Capacity report

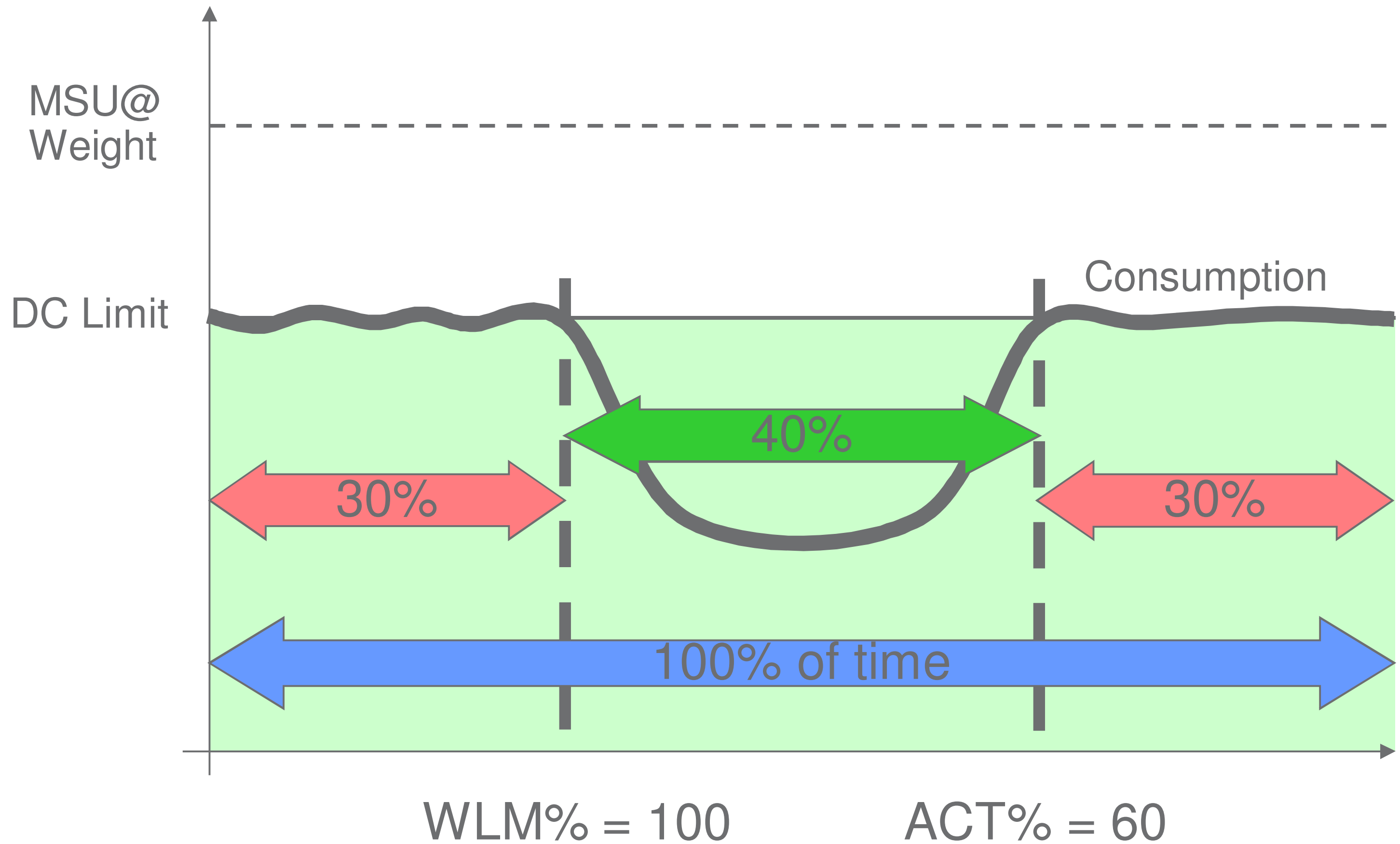
```
                          G R O U P   C A P A C I T Y   R E P O R T

      z/OS V1R12          SYSTEM ID SYS1              DATE 10/13/2010             INTERVAL 14.59.968
                          RPT VERSION V1R12 RMF       TIME 15.15.00               CYCLE 1.000 SECONDS

----GROUP-CAPACITY----  PARTITION   SYSTEM         -- MSU --    WGT   ---- CAPPING ----    - ENTITLEMENT -
NAME      LIMIT   AVAIL                            DEF    ACT         DEF   WLM%    ACT%   MINIMUM MAXIMUM
  1         2       3                               4      5           6     7       8         9      10
GROUP1    1500     -22  SYS1        SYS1           80      3    600   NO     25      23        80      80
                        SYS2        SYS2           80      3    500   NO    100      46        80      80
-----------------------------------------          ---------------------------------------------------------
                                   TOTAL                   6   1100
```

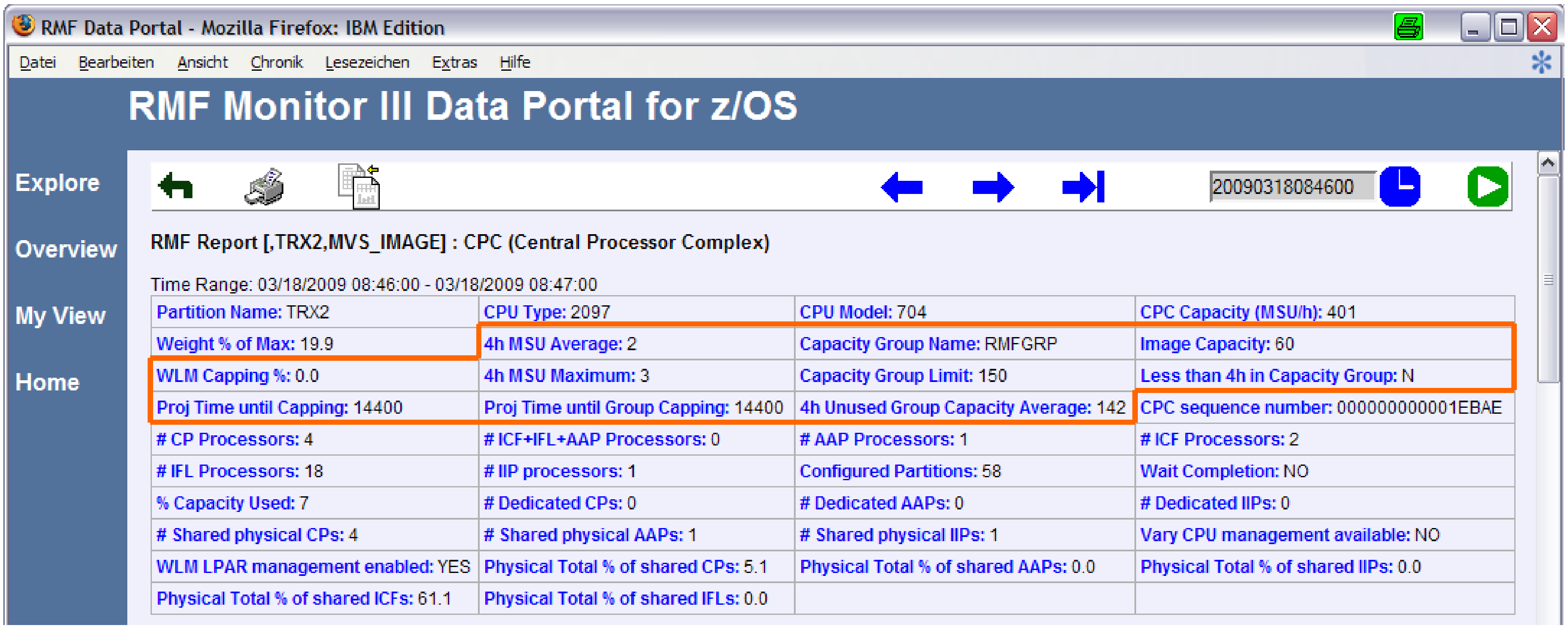
1. NAME	Name of the WLM capacity group
2. LIMIT	Group limit
3. AVAIL	Average unused capacity in MSUs (avg. unused vector)
4. MSU DEF	Defined capacity limit
5. MSU ACT	Average used capacity
6. CAPPING DEF	YES indicates that initial capping is active
7. CAPPING WLM%	Percentage of time that WLM had set up a cap for the partition
8. CAPPING ACT%	Percentage of time found capping actually limited the usage of processor resources for the partition
9. MINIMUM ENT.	Minimum of the GC member share and the DC limit
10.MAXIMUM ENT.	Minimum of the GC limit and the DC limit

Phantom weight: WLM% vs. ACT% in RMF



- RMF: WLM% capping is always 100 in case of phantom weight

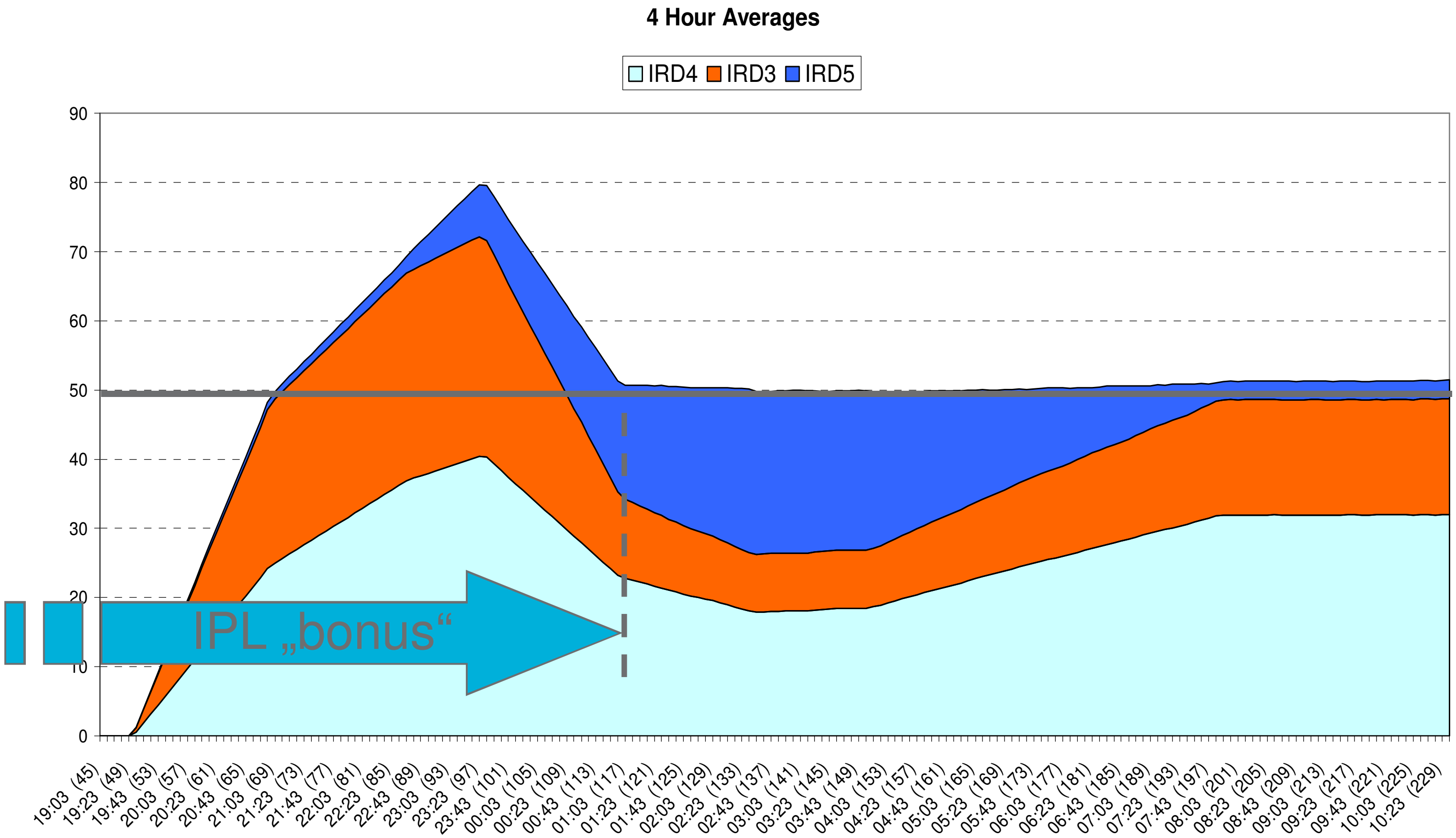
RMF Data Portal



Partition Name: TRX2	CPU Type: 2097	CPU Model: 704	CPC Capacity (MSU/h): 401
Weight % of Max: 19.9	4h MSU Average: 2	Capacity Group Name: RMFGRP	Image Capacity: 60
WLM Capping %: 0.0	4h MSU Maximum: 3	Capacity Group Limit: 150	Less than 4h in Capacity Group: N
Proj Time until Capping: 14400	Proj Time until Group Capping: 14400	4h Unused Group Capacity Average: 142	CPC sequence number: 00000000001EBAE
# CP Processors: 4	# ICF+IFL+AAP Processors: 0	# AAP Processors: 1	# ICF Processors: 2
# IFL Processors: 18	# IIP processors: 1	Configured Partitions: 58	Wait Completion: NO
% Capacity Used: 7	# Dedicated CPs: 0	# Dedicated AAPs: 0	# Dedicated IIPs: 0
# Shared physical CPs: 4	# Shared physical AAPs: 1	# Shared physical IIPs: 1	Vary CPU management available: NO
WLM LPAR management enabled: YES	Physical Total % of shared CPs: 5.1	Physical Total % of shared AAPs: 0.0	Physical Total % of shared IIPs: 0.0
Physical Total % of shared ICFs: 61.1	Physical Total % of shared IFLs: 0.0		

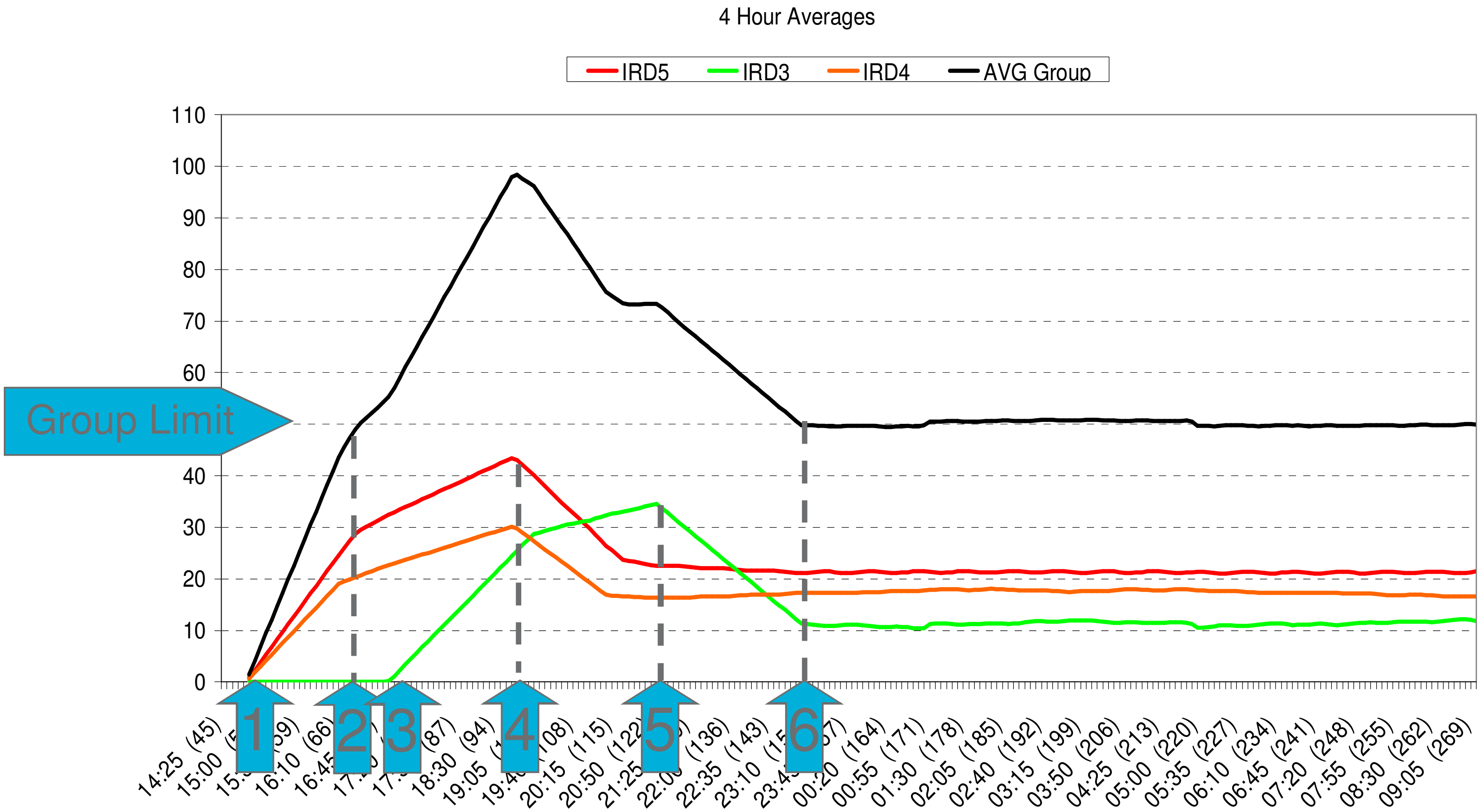
Many capping related fields are available in RMF Monitor III Data Portal

4 hour rolling average at IPL



Average is always for 4 hours even when the IPL was less than 4 hours ago

A member joins the capacity group



1. Workloads begin on IRD4 & 5
2. Group limit reached
3. System IRD3 joins group
4. IRD4 & 5: Four hours since (1.)
5. IRD3: Four hours since (3.). All systems have same GC view.
6. Group Avg. = Group limit

Capping and HiperDispatch

- WLM capping can influence the HiperDispatch configuration of an LPAR:
  - *limit<MSU@weight :* Capping through positive phantom weight reduces the PR/SM priority of an LPAR. Therefore, the number of Vertical High or Medium (VH, VM) processors may be reduced.
  - *Limit>MSU@weight:* Capping through negative phantom weight does not increase the PR/SM priority of a partition and the HiperDispatch configuration of the capped LPAR remains unchanged
- z/OS V2.2 and V2.1 with APAR OA43622 provide some HiperDispatch enhancements that become effective when running capped, or when capped LPARs are present on the CPC.

	z/OS release / **Function**	V2.2	V2.1	V1.13
Hiper-Dispatch z13 & zEC12	***Unpark while capped*** ***Unused capacity refinement*** *Prime cycle elimination*	**+**	***OA43622***	

HiperDispatch “Unpark while capped”

- Previously, HiperDispatch
  - Parked all Vertical Low (VL) processors when a system capped via positive phantom weight
    - VLs are used for discretionary capacity and not required to absorb the LPAR weight
    - However, it was seen that, for some workloads, the reduced number of logical processors made it difficult to fully utilize the cap target capacity.
  - Unparked all VL processors when a system was capped by negative phantom weight, or some cases of PR/SM absolute capping

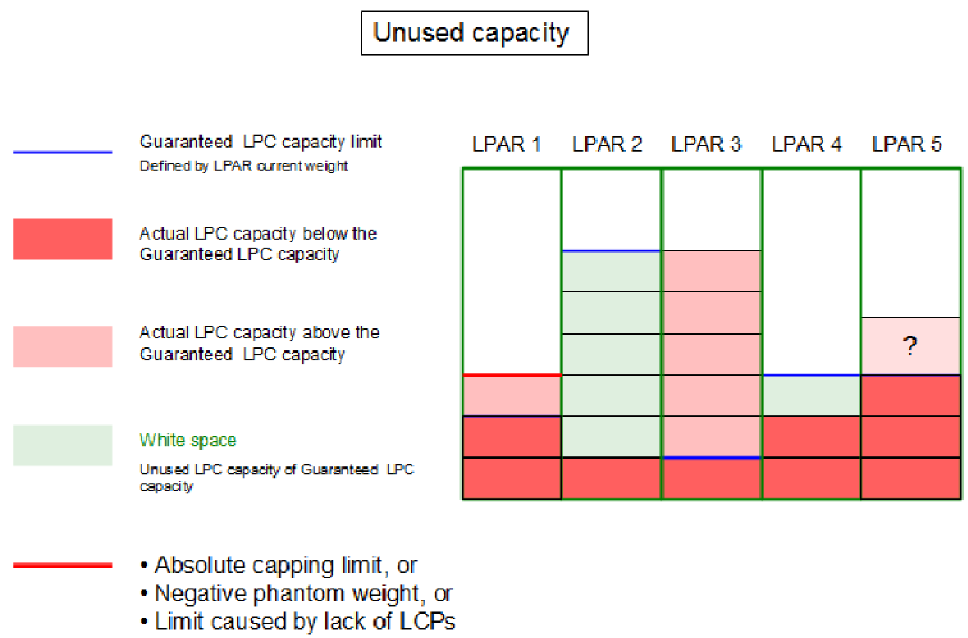
- Now, HiperDispatch can unpark VL processors <u>if</u> the processors can be used efficiently.

HiperDispatch refinement of “unused capacity” use

- HiperDispatch decisions are based on the CPC-wide unused capacity situation
- The ‘unused capacity share’ calculation was enhanced to also consider the LPAR configuration values
  - absolute capping value
  - negative phantom weight
  - number of logical processors
  - effective defined capacity and group capacity limit

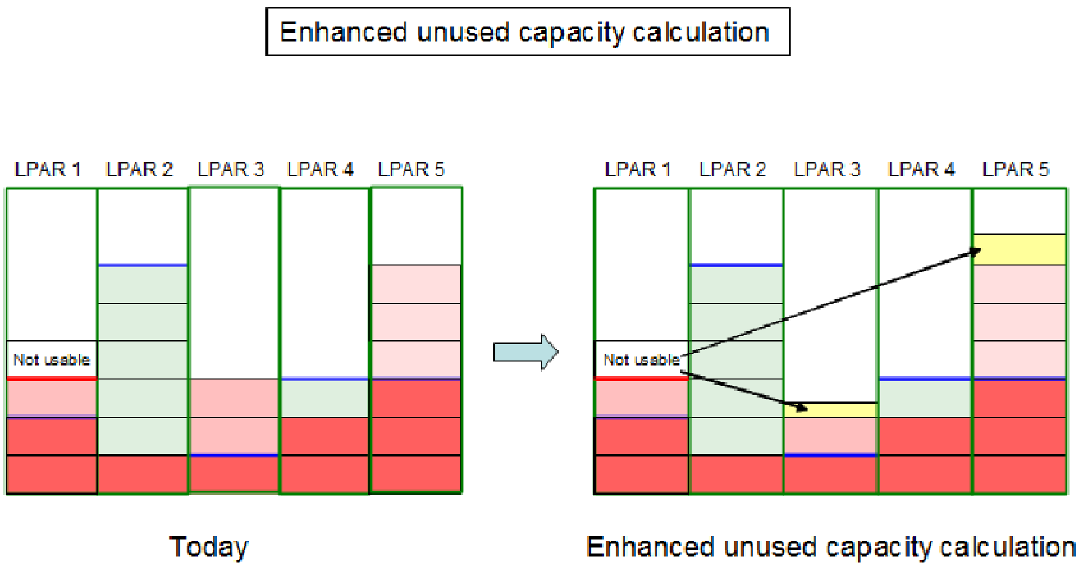
  of possible ‘unused capacity’ receivers

CPC with 5 LPARs. LPAR1 has an absolute capping limit, which is indicated with the red line. LPAR2, and LPAR4 are unused capacity donors, while LPAR1 / 3 / 5 are unused capacity receivers.



* Statements regarding IBM future direction and intent are subject to change or withdrawal, and represent goals and objectives only.

HiperDispatch refinement of “unused capacity” use



- Figure on the left shows today’s unused capacity calculation, which does not consider LPAR capping limits.
- Unused capacity calculation is only based on the receiver’s weight share.
- Figure on the right shows an example of enhanced unused capacity calculation. It considers the capping limits of the receivers.
- Because LPAR1 is not able to use its total unused capacity share its ‘not usable’ unused capacity share portion increases the unused capacity share of LPAR5.

Agenda

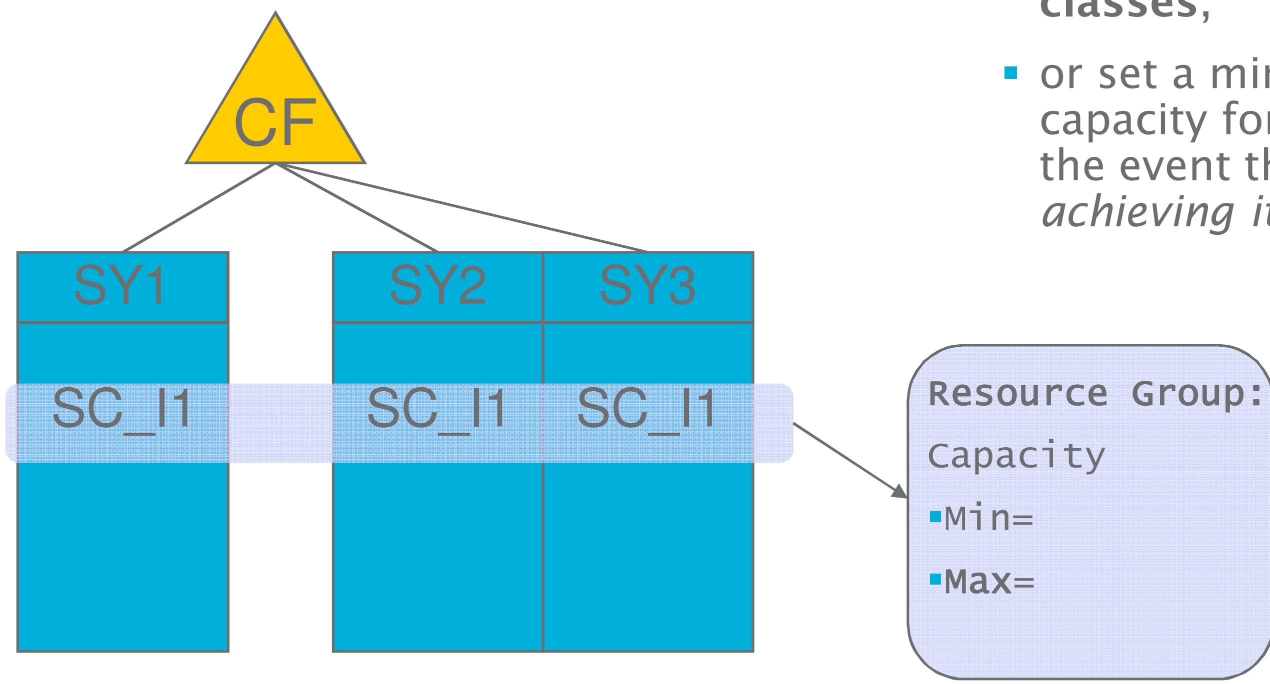
- Overview of capping types
- Initial capping
- Absolute capping
- Defined capacity & group capacity
- **Resource group capping**
- 4HRA management

- Additional Material

 50

What is a Resource Group?

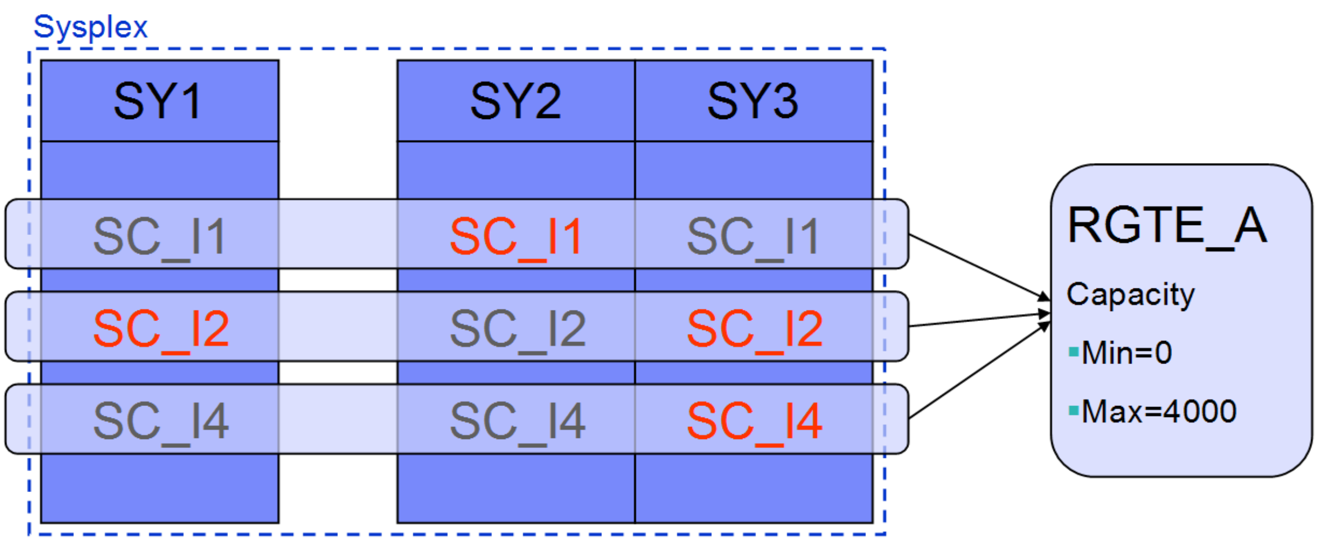
- Resource groups are a means to limit or protect work *when proper classification, goals and importance are not sufficient.*



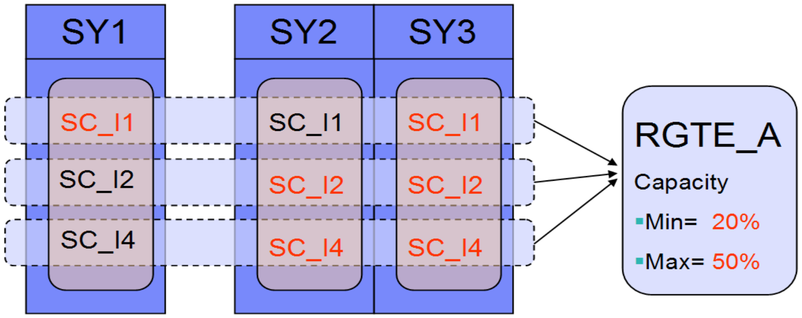
- A Resource Group is associated to one or more Service Classes
- Defines the service that the related Service Class(es) are managed to. Either
  - **limit the amount of processing capacity available to the service classes,**
  - or set a minimum processing capacity for the service classes in the event that the work is *not achieving its goals*

Type 1 Resource Groups

- Sysplex-wide defined in unweighted service units per second
  - "Unweighted" or "raw" meaning that the CPU and SRB service definition coefficients are not applied
- Sysplex-wide managed
- General Considerations
  - Multiple service classes may be assigned to a resource group
    - Different utilizations on the different systems and mix of importance levels make it difficult to predict actual consumption
  - Systems may have different capacities



Type 2 and 3 Resource Groups



- Sysplex-wide defined,
  but definition applies to each system
- Managed by each system
- General Considerations
  - Multiple service classes can be assigned to a resource group but this has no sysplex-wide effect
  - Definition is based on one of two possible units:
    - **Type 2: Percentage of LPAR capacity**
    - **Type 3: In number of processors (100 = 1 CP)**

Locating LPAR SU/sec Numbers

The service units that

- The Service Unit information can be located in the "z/OS MVS Planning: Workload Management" manual CPU Capacity Table
- Or on IBM Resource Link https://ibm.biz/BdFHFv :

IBM zEnterprise EC12

Processor	STIDP Type	STSI Model Name	CPs	SU/SEC	SRMsec/RealSec
2827-701	2827	701	1	78048.7805	1811.5932
2827-702	2827	702	2	73394.4954	1811.5932
2827-703	2827	703	3	71428.5714	1811.5932
2827-704	2827	704	4	69868.9956	1811.5932
2827-705	2827	705	5	68085.1064	1811.5932
2827-706	2827	706	6	66945.6067	1811.5932
2827-707	2827	707	7	65843.6214	1811.5932

A 4-way LPAR on a zEC12 model 7xx server can deliver approx.
4 * 69869
~ **279476 SU/sec**

Resource Group Management

- To implement capping, the elapsed time is divided into 256 or 64 (pre-z/OS V2.1) slices. Each cap slice then represents 1/256th or 1/64th of the total elapsed time.

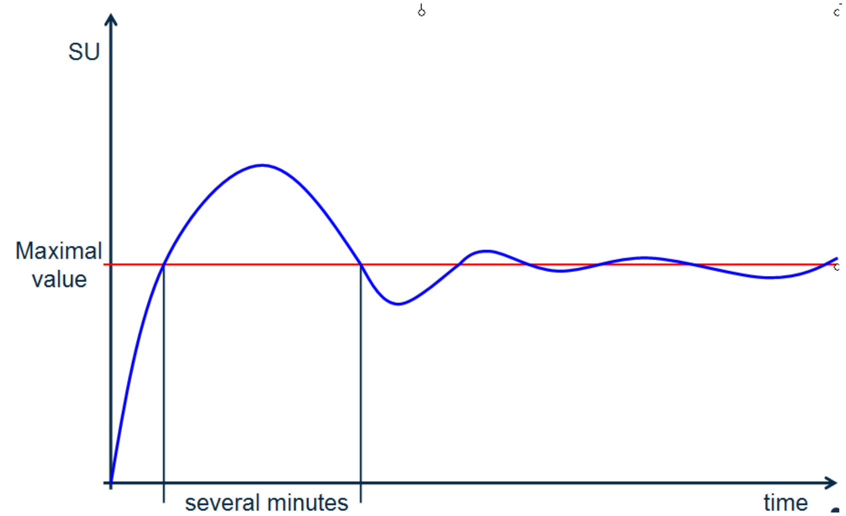
- Dispatchable units from address spaces or enclaves belonging to a resource group are made nondispatchable during some slices in order to reduce access to the CPU to enforce the resource group maximum.

- The time where address spaces or enclaves in a resource group are set non-dispatchable is called a CAP SLICE.

- The time where address spaces or enclaves in a resource group are set dispatchable is called an AWAKE SLICE.

Resource Group Maximum continued…

- Every 10 seconds the policy adjustment code re-evaluates the resource groups and adjusts the cap pattern accordingly
- The forecast for the next 10 seconds is based on the average data from the last minute
- Because of the 1 minute average data, during a ramp up period, the max may be exceeded. Also, during periods of workload oscillation WLM may tend to under cap on the up swing but over cap when the workload is dropping off.



Resource Group Maximum continued…

Under certain conditions work may continue consuming service even while being capped

- Any locked work will continue to be dispatched as long as the lock is held
  - Check promoted times in RMF workload activity report
- The region control task is exempt from this nondispatchability.
- The address space will not be marked nondispatchable until the next dispatch.

Resource Group Considerations with zAAP/zIIPs

- Resource Groups are managed based on their general purpose processor consumption (TCB+SRB)

- Difficult to predict result of assigning RGs to service classes that execute on specialty processors
  - Especially when IFAHONORPRIORITY=YES or IIPHONORPRIORITY=YES is in effect.

1	9	17	25	33	41	49	57
2	10	18	26	34	42	50	58
3	11	19	27	35	43	51	59
4	12	20	28	36	44	52	60
5	13	21	29	37	45	53	61
6	14	22	30	38	46	54	62
7	15	23	31	39	47	55	63
8	16	24	32	40	48	56	64

Other considerations for Resource Groups

- **Not valid for transaction oriented work, such as CICS or IMS transactions.**
  - In order to assign a minimum or maximum capacity to CICS or IMS transactions, the region service classes can be assigned to a resource group.
    - Such interactive work can respond harshly to CPU bottlenecks: Evaluate what cap level can be tolerated
- **Given the combination of the goals, the importance level, and the resource capacity, some goals may not be achievable when capacity is restricted.**
- Unless there is a specific need for limiting or protecting capacity for a group of work, it is best to not define resource groups and to just let workload management manage the processor resources to meet performance goals.

Identifying Resource Group Capping

- In the RMF Workload Activity report, RG capping is identified in the Execution Delays section as CAP delays
- CAP delays may also be incurred by service classes that have not been associated with resource groups
  ➔ Discretionary Goal Management (DGM)

```
GOAL: EXECUTION VELOCITY 20.0%     VELOCITY MIGRATION:   I/O MGMT  93.9%     INIT MGMT 90.1%

          RESPONSE TIME EX    PERF  AVG    --EXEC USING%--  ------------- EXEC DELAYS % -----------  -USING%-
SYSTEM                  VEL%  INDX ADRSP  CPU AAP IIP I/O  TOT CPU CAP I/O                           CRY CNT

SYS1          --N/A--   93.9   0.2   0.0   46 N/A N/A  43  5.8 4.3 1.2 0.3                           0.0 0.0
```

Discretionary Goal Management (DGM)

- Allows an *eligible over-achieving* service class to donate CPU to a discretionary period
  - Objective is to improve service that discretionary periods receive when no non-discretionary periods need help and goals are vastly overachieved

- The donation is implemented through resource group capping.

- To be considered as a donor a period must meet several requirements, including
  - Not a member of a Resource Group (RG)
  - Non-aggressive goal:
    - If it has a velocity goal, the goal must be ≤ 30
    - If it has a response time goal, the goal must be > 60 sec
  - The performance index PI must be < 0.7

- If a period should never donate due to DGM, define appropriately:
  - Velocity goal > 30 or response time goal ≤ 60 sec, or
  - Define resource group with MIN=MAX=0 and associate service classes to be protected with that RG

Agenda

- Overview of capping types
- Initial capping
- Absolute capping
- Defined capacity & group capacity
- Resource group capping
- **4HRA management**

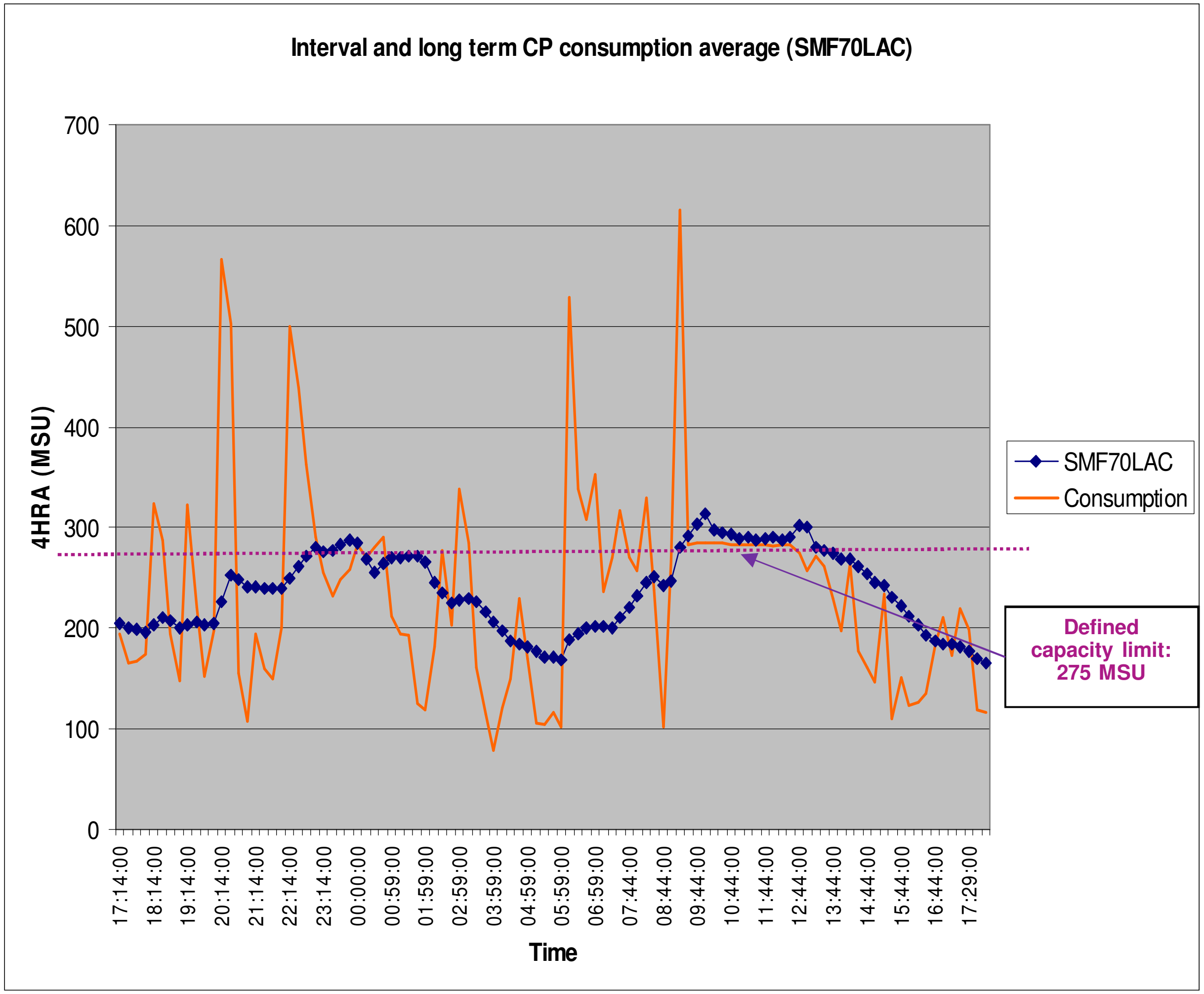
- Additional Material

64

4HRA business aspects

- Peak value of MIN(4HRA, defined capacity limit) over billing period determines software charges
  - 4HRA peaks may exceed the defined limit
- Periods of low utilization can be used to “save” capacity for subsequent peak times
  - No capping when 4HRA < limit
- Utilization peaks drive up the 4HRA
- From a cost perspective it is usually desirable to **limit the peak consumption**
- Seek for technical means to
  - Limit consumption (→peak consumption)
    - Primarily of less important work
    - Also during –previously uncapped– periods
  - Maintain service levels, responsiveness and system integrity
    - Especially for important work

Interval consumption and the 4 hour rolling average: A sample day



Techniques for managing the 4HRA

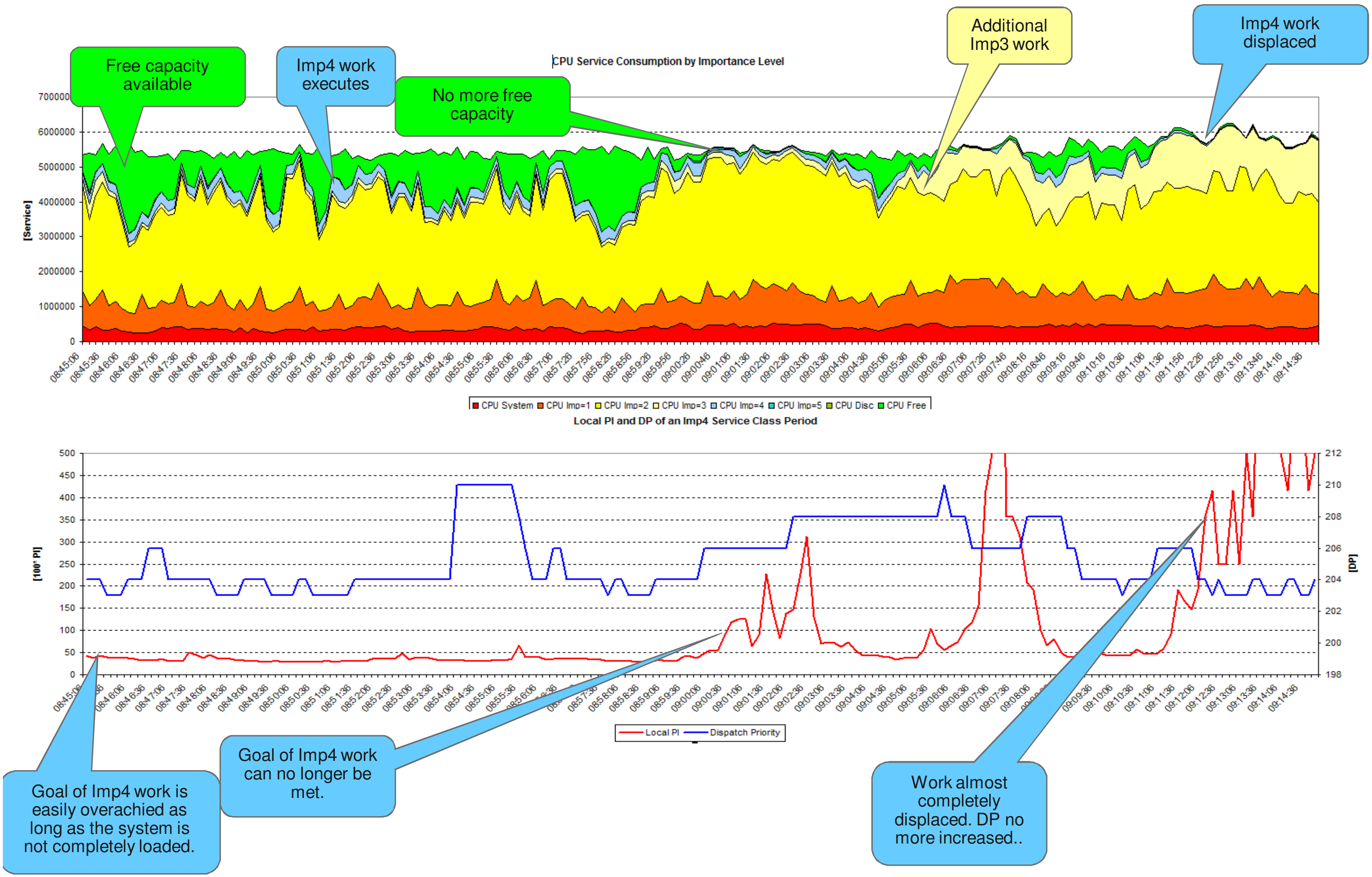
- Schedule work into off-peak hours

- Limit consumption at an LPAR level
  – Defined or group capacity
  – WLM importance level determines what work gets sacrificed first
    - 4HRA-wise irrelevant, but technically - beware of reduced preemption, promotion

- Selectively limit work within a system
  – Limit demand or parallelism
    - E.g. number of initiators
  – Resource groups
    - But not suitable for every work.

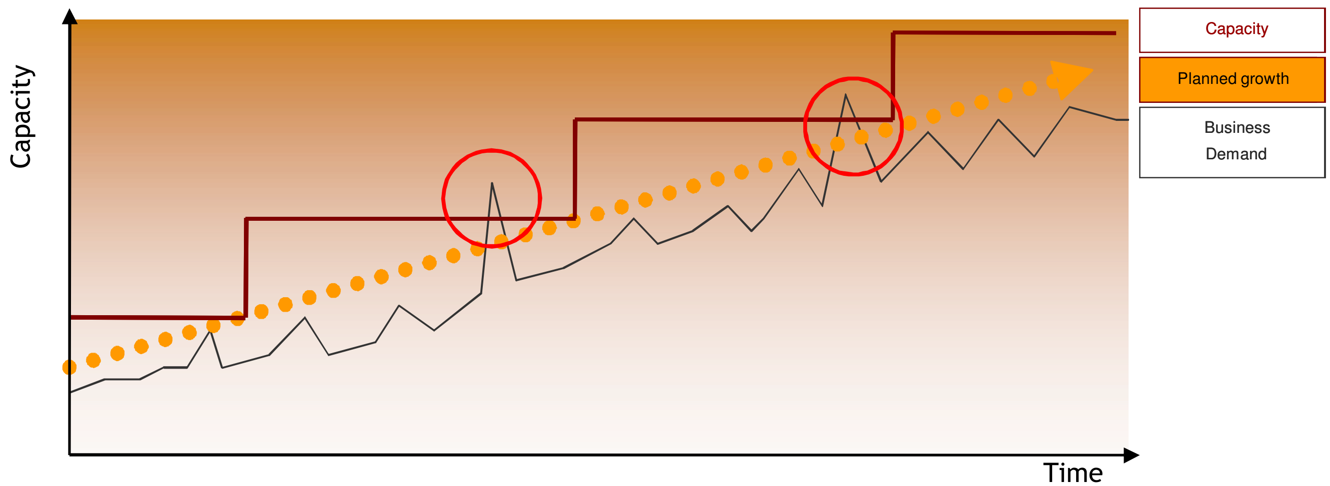
- Any combination of the above
  – Can also help to mitigate impacts of capping

Importance Distribution and Displacement of Work



IBM z/OS Capacity Provisioning Basics

- **Contained in z/OS base component free of charge**
  - Requires a monitoring component, such as z/OS RMF, or equivalent
  - Base element since z/OS V1.9
- **Exploits on System z On/Off Capacity on Demand Feature**
  - IBM zEnterprise System z10 or later
  - If On/Off CoD is not used CPM “analysis” mode may be used for monitoring and alerts
- **Exploits Defined Capacity and Group Capacity**
  - Defined Capacity with IBM System z10 or later
  - Group Capacity with IBM zEnterprise z196 or later



Capacity Provisioning Capabilities Overview

– The Capacity Provisioning Manager (CPM) can control additional capacity on IBM zEC12, z196, or z10 (plus BC10 and later)
  - Number of temporary zAAPs or zIIPs
  - Temporary general purpose capacity
- Considers different capacity levels (i.e. effective processor speeds) for subcapacity processors (general purpose capacity)
  - Can advise on logical processors
  - **Defined capacity and group capacity limits**
  - Can control one or more IBM zEnterprise or System z10 servers
    - Including multiple Sysplexes
  - Provides commands to control z196 and later static power save mode
  - Provides commands to control temporary IFLs

CPM allows for different types of provisioning requests:

- Manually at the z/OS console through Capacity Provisioning Manager commands
- Via user defined policy at specified schedules
- Via user defined policy by observing workload performance on z/OS

Policy Approach

The Capacity Provisioning policy defines the circumstances under which additional capacity may be provisioned:

- Three "dimensions" of criteria considered:
  - **When** is provisioning allowed
  - **Which** work qualifies for provisioning
  - **How much** additional capacity may be activated
- These criteria are specified as "rules" in the policy:

```
If
{   in the specified time interval
    the specified work "suffers"
}
Then up to
{       -  the defined additional capacity
        may be activated

}
```

- The specified rules and conditions are named and may be activated or deactivated selectively by operator commands

Key benefit of CPM is the real time in-depth analysis of bottlenecks

```
MODIFY CPOSERV,APPL=REPORT WORKLOAD TYPE=DETAILED

Workload is analyzed for 1 system(s)
Workload for system PROD1 of sysplex PRODPLEX on CPC CPC1
   CICSHIGH.1 PL/PD/DL/DD/S 1.8 5 1.2 12 System
       PI from 11/16/2012 07:43 is 2.76
           Last limit crossing was 12/16/2012 07:27
       Demand for additional physical zIIPs not recognized
           System zIIP-utilization too low
       Demand for additional physical zAAPs not recognized
           System zAAP-utilization too low
       Demand for additional defined capacity recognized
       Demand for additional physical CPs not recognized
       Demand for capacity level increase not recognized
       Demand for additional logical CPs not recognized
           CPC-wide CP-utilization too low
```

- Key benefit of CPM is the real time in-depth analysis of workload constraints and demands
  - Based on WLM-provided metrics
- Can identify what type of capacity (if any) will help
- Timely reaction, even before capping begins

Capacity Provisioning Policy Strategies… for cost optimization

- Baseline defined or group capacity (DC/GC) limit relatively low
  - but still realistic for periods of low to average utilization

- Use Capacity Provisioning Manager rules to increase DC/GC limit
  - only when required by a qualifying workload during a qualifying time period
    - Time & workload conditions:
      Allow for higher DC/GC limits as required by workload

  - unconditionally during a qualifying time period
    - Time conditions without workload conditions:
      Unconditionally provision full rule scope

- When needed, can differentiate between different systems, service definitions, or override policies

Capacity Provisioning Policy sample scenario for cost optimization with LPAR defined capacity

- Sample scenario defines two qualifying workloads
  - Important online work
    - Monday through Friday, 07:45 – 18:00
    - Comprised of two service classes
      - DB2HIGH
      - ONLSTC
    - Up to +300 MSU may be provided in addition
  - Early evening batch
    - Monday through Friday, 20:00 – 22:00
    - Comprised of one service classe
      - BATCRIT
    - Up to +70 MSU may be provided in addition

Capacity Provisioning Policy Sample…
… with LPAR defined capacity (1)

- Two workloads that may warrant higher DC limits during different times of day:

Maximum Processor Scope | Logical Processor Scope | Maximum Defined Capacity Scope | Maximum Group Capacity Scope | **Rules**

Actions

Name Filter	Description Filter	Default Status Filter
WeekNight	Weekdays DC pre midnight batch	Enabled
WeekdayDC	Weekdays DC for online work	Enabled

- WeekdayDC rule scope allows for up to +300 (additional) MSU:

Processor Scope | **Defined Capacity Scope** | Group Capacity Scope | Conditions

Actions

System Filter	Sysplex Filter	Max. Increase (MSU) Filter
SYS1	PLEX1	300

Capacity Provisioning Policy Sample…
… with LPAR defined capacity (2)

- Rule is enabled for all weekdays prime time

Nonrecurring Time Conditions | Recurring Time Conditions | Workload Conditions

Actions

Name	Start Date	End Date	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Start Time ▲	Deadline	End Time
Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
AllWeekD	Jan 2, 2014	Dec 31, 2014	✓	✓	✓	✓	✓			7:45 AM	6:00 PM	6:30 PM

- Workload is defined by specific service classes

mportance Filters | Included Service Classes | Excluded Service Classes

Actions

	Service Definition	Service Policy	Service Class	Period	Provisioning PI	Provisioning Duration (Minutes)	Deprovisioning PI	Deprovisioning Duration (Minutes)
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
	Any service definition	Any service policy	DB2HI	1	1.4	2	1.1	10
[x]	Any service definition	Any service policy	ONLSTC	1	1.5	5	1.1	10

Capacity Provisioning Policy Sample…
… with LPAR defined capacity (3)

- Similarly, another rule is defined to cover a batch workload
  - Up to +70 MSU for a single batch service class

Nonrecurring Time Conditions | **Recurring Time Conditions** | Workload Conditions

Actions

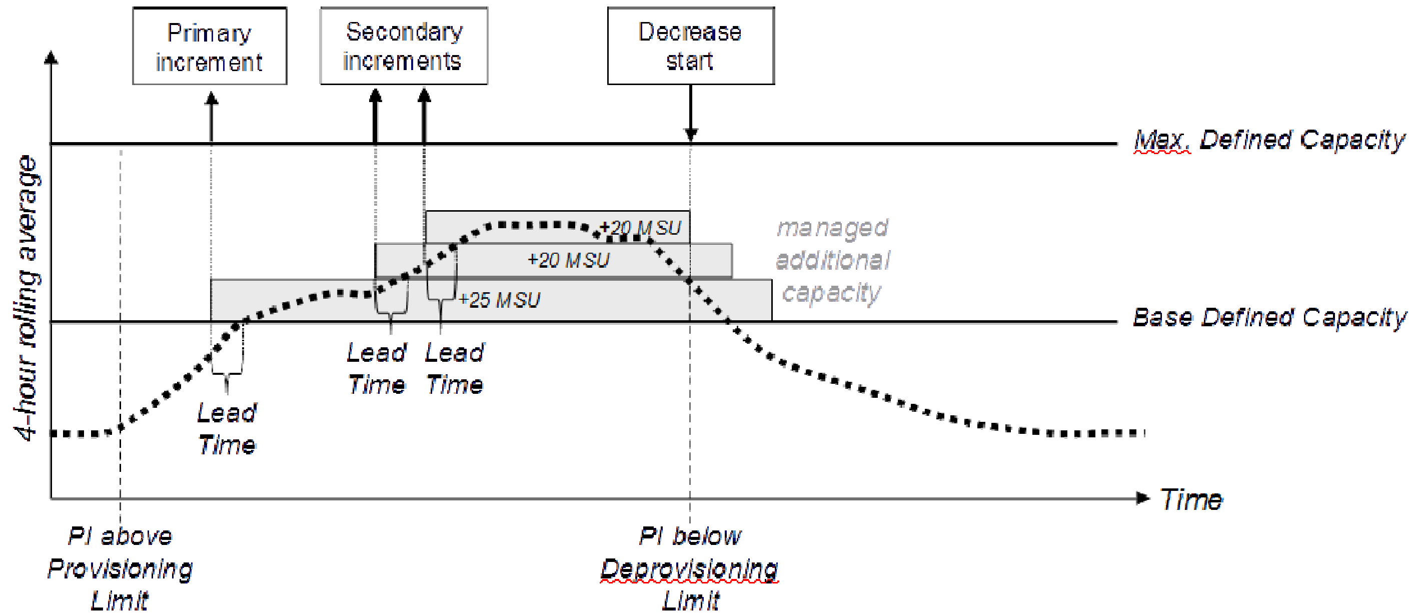
	Name Filter	Start Date Filter	End Date Filter	Mon Filter	Tue Filter	Wed Filter	Thu Filter	Fri Filter	Sat Filter	Sun Filter	Start Time ▲ Filter	Deadline Filter	End Time Filter
☐	AllWeekN	Jan 2, 2014	Dec 31, 2014	☑	☑	☑	☑	☑	☐	☐	8:00 PM	10:00 PM	10:00 PM

Importance Filters | **Included Service Classes** | Excluded Service Classes

Actions

	Service Definition Filter	Service Policy Filter	Service Class Filter	Period Filter	Provisioning PI Filter	Provisioning Duration (Minutes) Filter	Deprovisioning PI Filter	Deprovisioning Duration (Minutes) Filter	PI Scope Filter
☐	Any service definition	Any service policy	BATCRIT	1	1.8	5	1.3	10	System

Capacity Provisioning Defined Capacity Management



- When required by the defined workload the CPM will increase the defined capacity limit while the workload criteria are met
- The additional defined capacity will be managed down as the workload permits
  – Or deferred, based on user specification
- Additional user-initiated DC/GC activations are recognized and tolerated.

Conclusion

z/OS provides the tools to monitor
and tightly manage
the rolling 4 hour average consumption
for efficient cost management.

z/OS Capacity Provisioning Documentation

- For more information contact: IBMCPM@de.ibm.com
- z/OS Capacity Provisioning: Introduction and Update for z/OS V2.1, SHARE in Anaheim, Session 14210, 8/2013
- Website http://www.ibm.com/systems/z/os/zos/features/cpm
- z/OS MVS Capacity Provisioning User's Guide, SC34-2661, at http://publibz.boulder.ibm.com/epubs/pdf/iea3u110.pdf