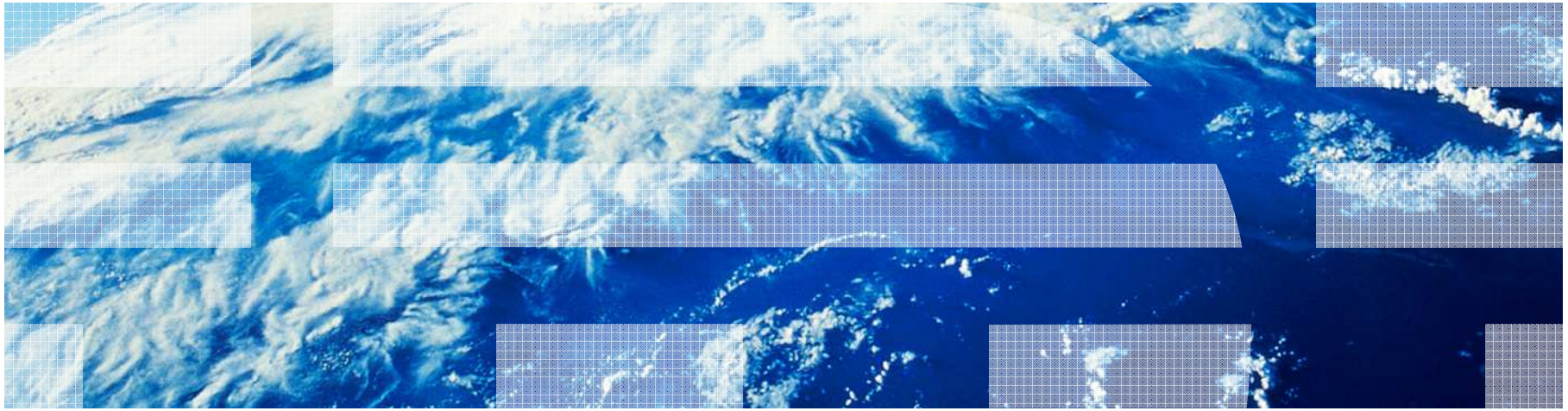
LPAR Weight, Entitlement, and Capacity

Revision 2014-04-11.1

Brian K. Wade, Ph.D.
IBM z/VM Development, Endicott, NY
bkw@us.ibm.com



Agenda

- What the z/VM administrator sees
- Basic mode and LPAR mode
- Weight, entitlement, and logical CPU count
- Ways to go wrong
- Some examples of doing it right
- z/VM Performance Toolkit reports

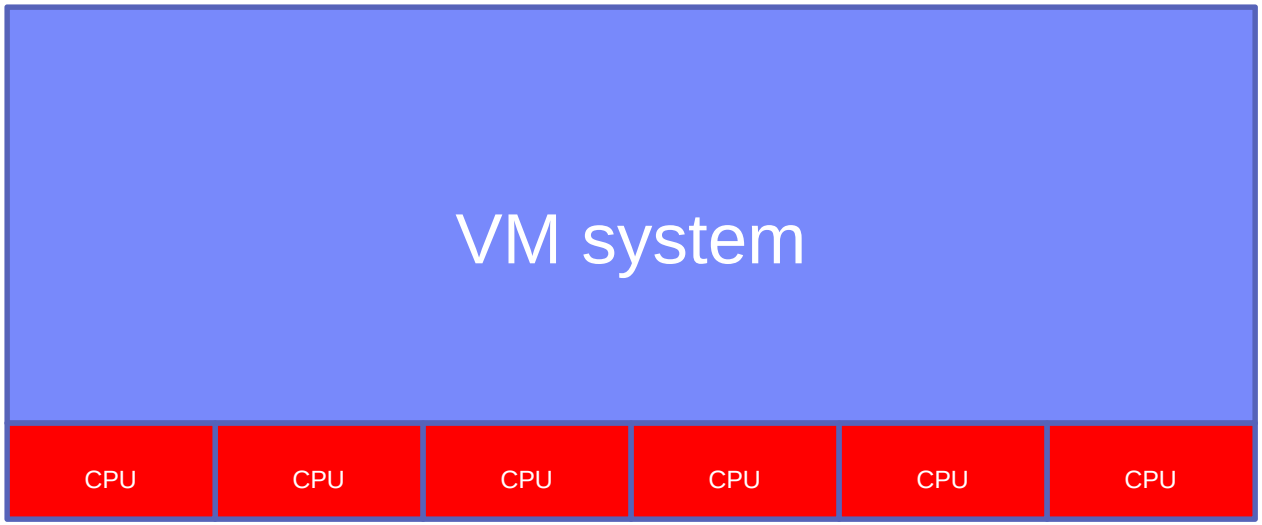
Our z/VM System Administrator, Jane

```
cp query proc
PROCESSOR 00 MASTER CP
PROCESSOR 01 ALTERNATE CP
PROCESSOR 02 ALTERNATE CP
PROCESSOR 03 ALTERNATE CP
PROCESSOR 04 ALTERNATE CP
PROCESSOR 05 ALTERNATE CP
Ready; T=0.01/0.01 13:46:02
```

Hey, I have a six-way! I know that's enough for my workload, so I'm golden!

In a moment we are going to find out just how wrong that conclusion is!

The Machine in Basic Mode

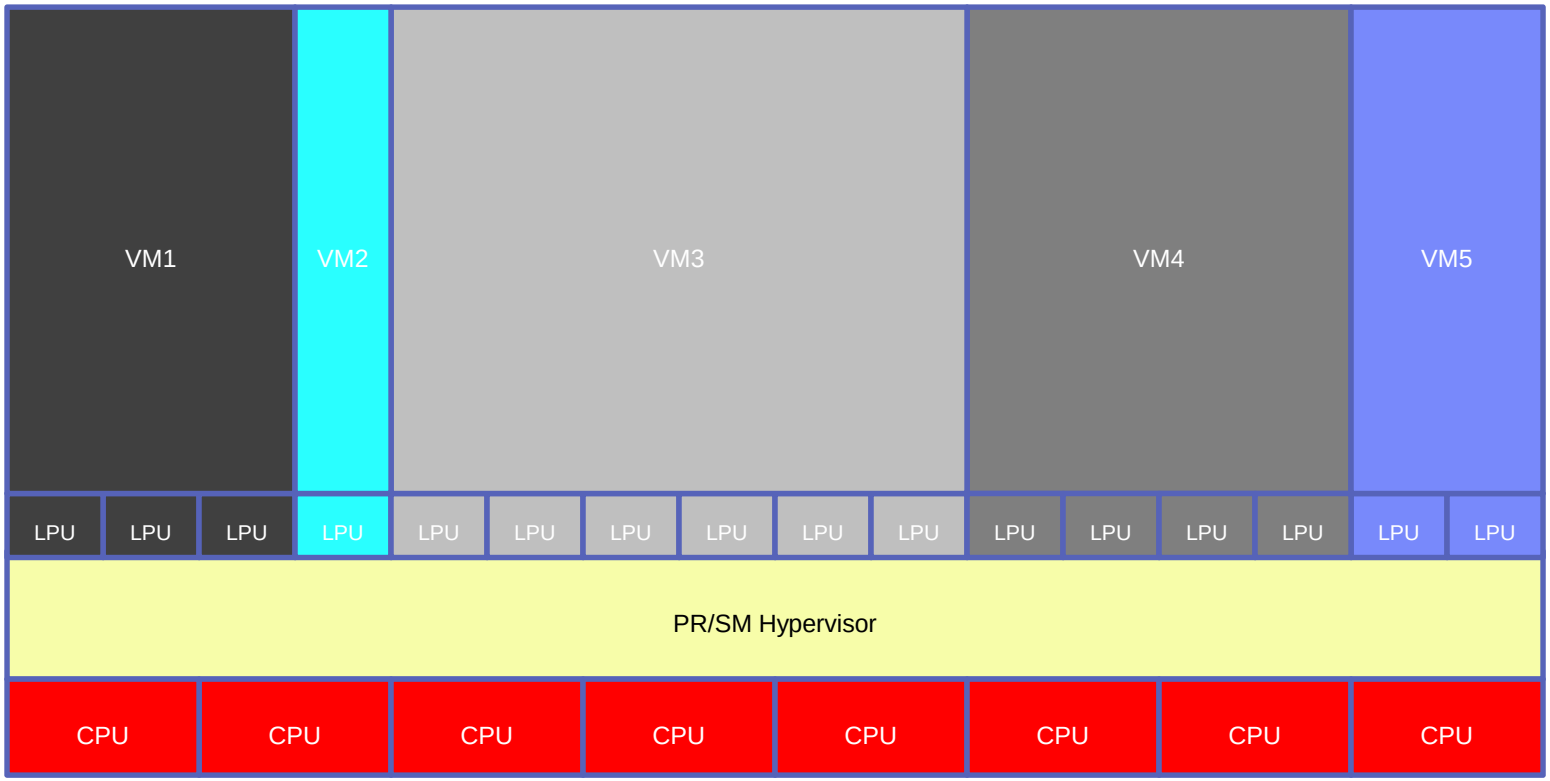


In the old days, VM ran right on the hardware. There was no such thing as the PR/SM hypervisor or an LPAR.

If CP QUERY PROC said you had six CPUs, you had six real, physical, silicon CPUs.

Those six CPUs were all yours, all the time.

The Machine in LPAR Mode

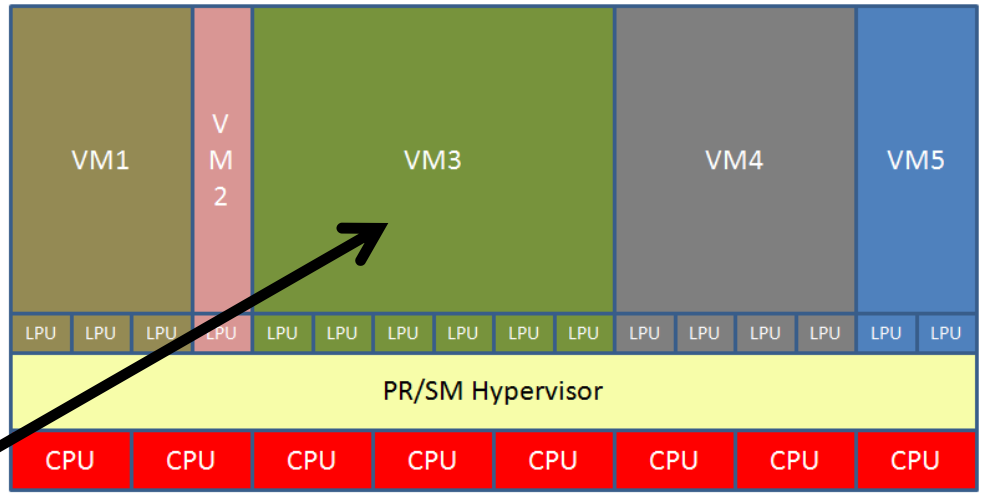


Processor Resource/System Manager (PR/SM) owns the physical machine.
PR/SM carves the machine into zones called *partitions.*
PR/SM timeslices partitions' logical CPUs onto physical CPUs.

A logical CPU is **not** a source of capacity. It is a **consumer** of capacity**.**

Poor Jane!

```
cp query proc
PROCESSOR 00 MASTER CP
PROCESSOR 01 ALTERNATE CP
PROCESSOR 02 ALTERNATE CP
PROCESSOR 03 ALTERNATE CP
PROCESSOR 04 ALTERNATE CP
PROCESSOR 05 ALTERNATE CP
Ready; T=0.01/0.01 13:46:02
```



16 logical CPUs (consumers of power)
-- running on --
8 physical CPUs (sources of power)

Jane's six-way system is now running in a partition.
She is now competing with many other partitions for the machine's eight CPUs' worth of power.
Jane has no idea that she might not get six CPUs' worth of power.

How Does PR/SM Decide?

The CEC administrator assigns each partition a *weight*.

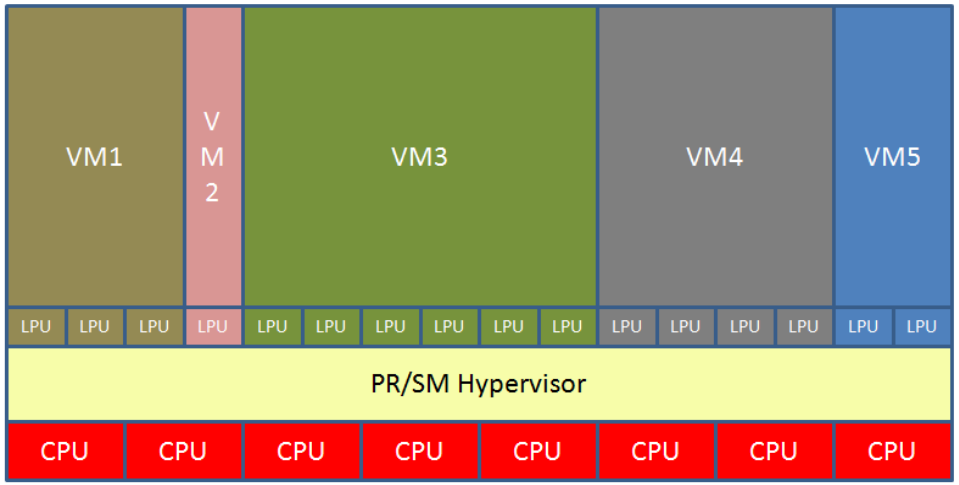
Weight expresses *relative importance* in the distribution of CPU power.

The weights determine the partitions' *entitlements.*

A partition's *entitlement* is the minimum power it can generally expect to be able to get whenever it wants it.

Entitlements come into play only when there is not enough power to satisfy all partitions' demands.

As long as the physical CPUs have some spare power, all partitions can use whatever they want.



S = number of shared physical CPUs = 8

$$\text{my E} = 100 * S * \frac{\text{(my weight)}}{\text{(sum of weights)}}$$

Notice:
1. Σ E = 100 * S. (the entitlements sum to the capacity)
2. E is **not** a function of the number of logical CPUs.

Entitlement: A Really Simple Example

Assume this machine has 18 shared physical engines.

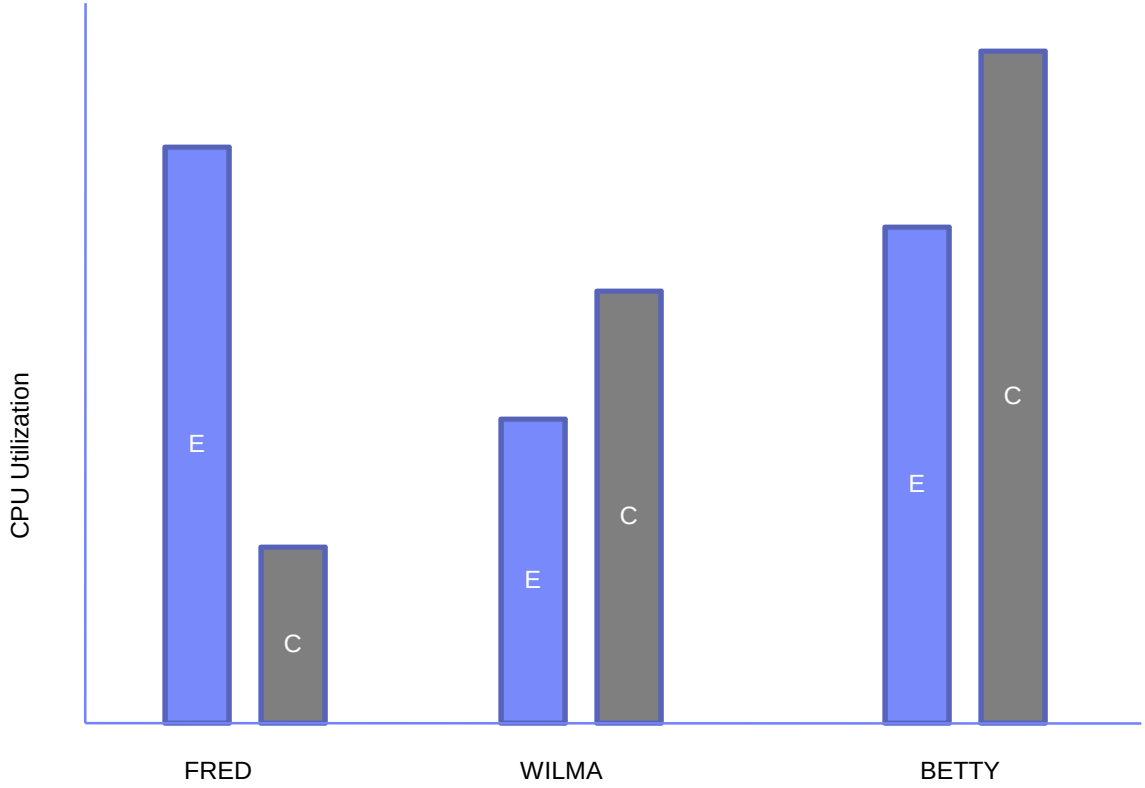
Partition	Weight	Weight-Sum	Calculation	Entitlement
FRED	35	90	100 * 18 * (35 / 90)	700%
BARNEY	55	90	100 * 18 * (55 / 90)	1100%
			SUM →	1800%

Notice:

1. The entitlements sum to the capacity of the shared physical engines.
2. The number of logical CPUs is NOT a factor in calculating entitlement.

By the way: "100%" means "one physical engine's worth of power".

Entitlement (E) vs. Consumption (C)



WILMA and BETTY can use over their entitlements only because FRED is using under his entitlement.

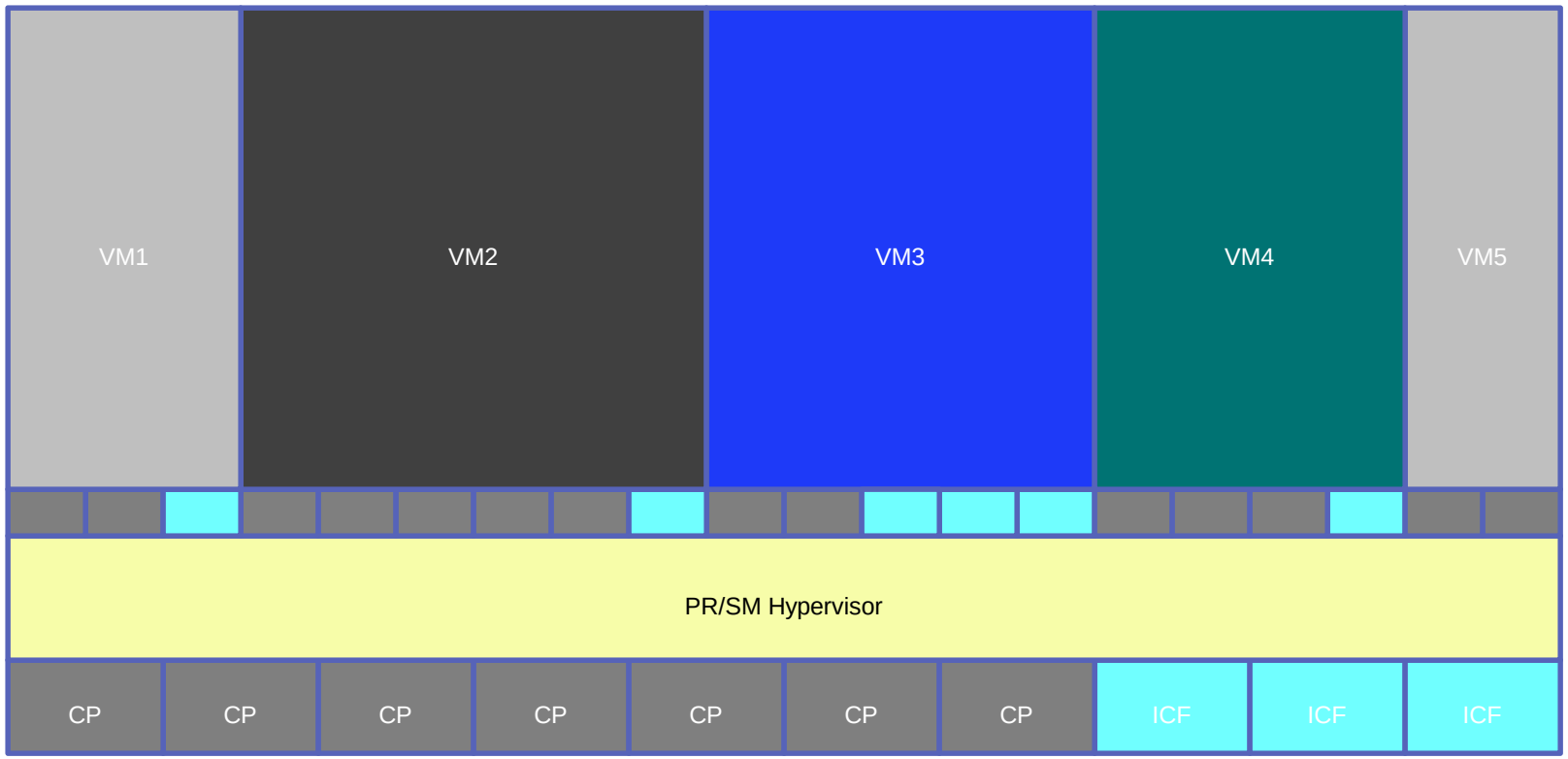
This can happen whether or not the physical CPUs are saturated.

FRED can use his entitlement whenever he wants.

If there is not enough spare power available to let FRED increase to his entitlement, PR/SM will divert power away from WILMA and BETTY to satisfy FRED.

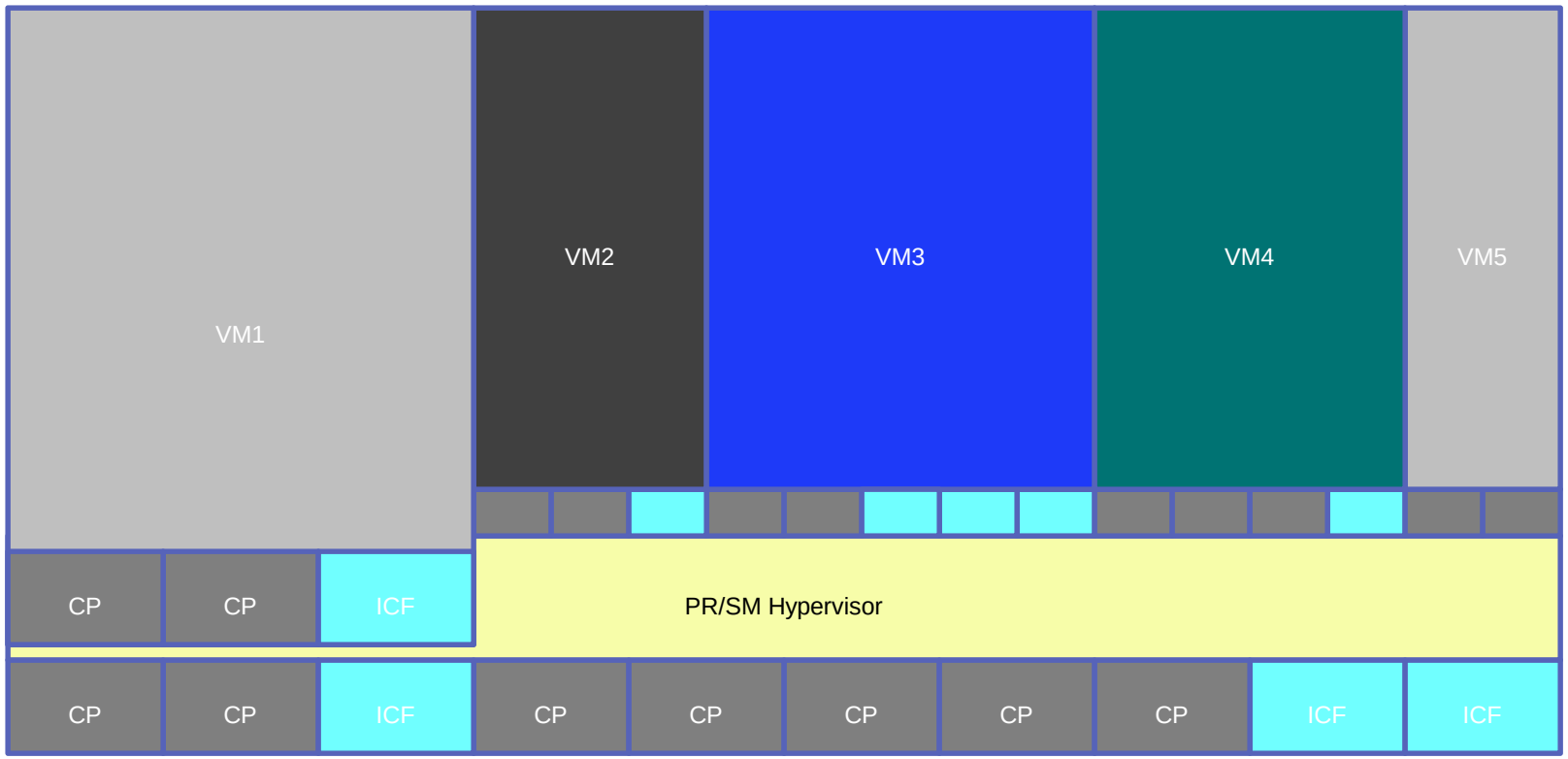
Three partitions: FRED, WILMA, and BETTY

Mixed-Engine Configurations



Weight and entitlement are type-specific attributes.
Each LPAR having logical CPs has a CP weight and entitlement.
Each LPAR having logical ICFs has an ICF weight and entitlement.

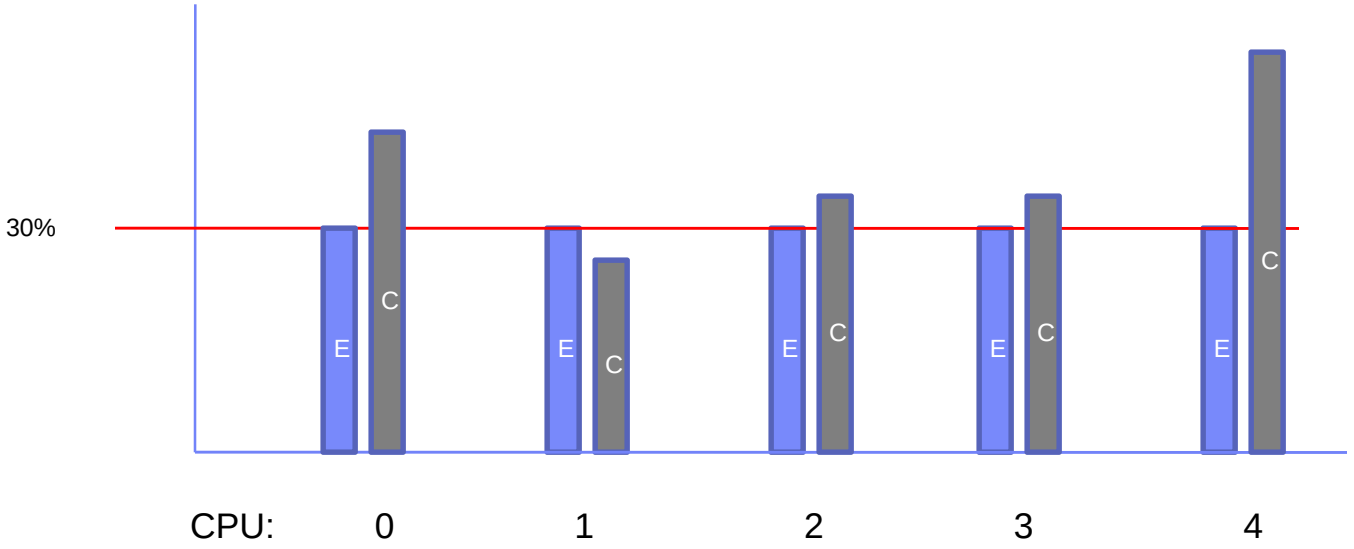
Dedicated Partitions



VM1's logical CPUs are directly assigned to physical CPUs.
Those physical CPUs are not used for any other partitions.
If you're running in VM1, life is really good! E=100%*n for all of your types.
(Q: what are the differences between this and a shared logical N-way with E=100%*N?)

Entitlement and Consumption within a Partition
(Horizontal Mode Partitions… z/VM 6.2 or earlier)

Within a single partition, the entitlement is distributed equally across the logical CPUs.

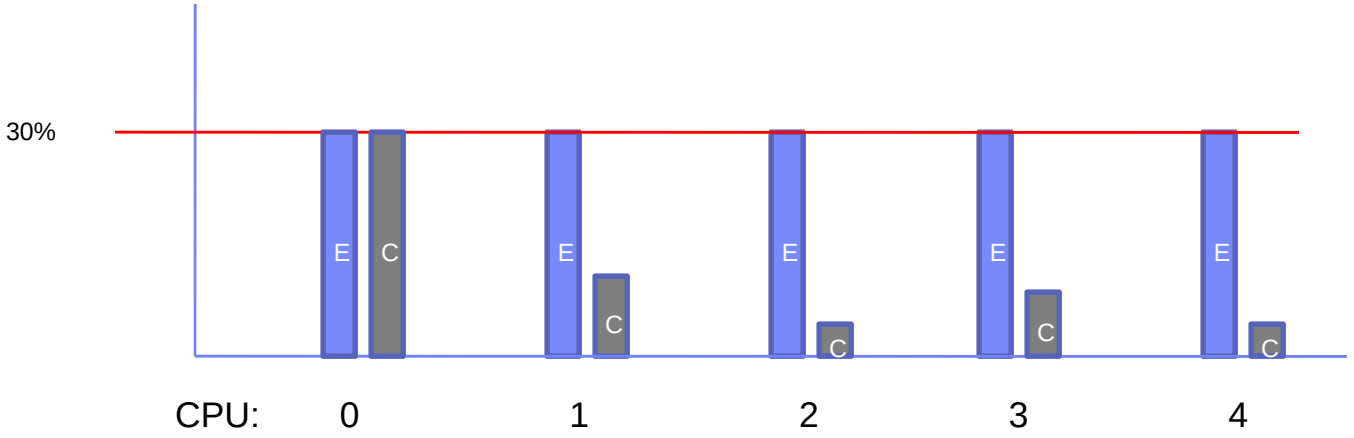


Suppose E = 150% and the partition has 5 logical CPUs.
Each logical CPU is entitled to (150% / 5) or 30% of a physical CPU.
The logical CPUs might actually consume more, depending on the availability of spare power.

Thinking about vertical mode? We can talk about HiperDispatch some other time.

What if the Partition is Capped?
(Horizontal Mode Partitions… z/VM 6.2 or earlier)

CAPPED: every logical CPU is held back to its share of the partition's entitlement.



Suppose E = 150% and the partition has 5 logical CPUs and is capped.
Each logical CPU is entitled to (150% / 5) or 30% of a physical CPU.
No logical CPU will ever run more than 30% busy. Availability of excess power is irrelevant.

In mixed-engine environments, capping is a type-specific concept.
For example, a partition's logical CPs can be capped and its logical ICFs not capped.

Some Guidelines to Use

1. Know your workloads. Especially, know how much power they need!
   If you need help, get help from workload sizing experts.
2. For each shared partition, what is its workload's bare minimum power requirement? Use that requirement as the partition's entitlement E.
3. Sum of the E values = bare minimum physical engines needed.
4. Add in some spare engines: for PR/SM itself and for comfort, or growth, or emergencies.
5. If in addition you want some dedicated LPARs, add in for those.
6. Set the shared partitions' weights in proportion to the entitlements you calculated above.
7. For each partition, what is the maximum power you want it ever to be able to consume?
8. Using those maxima, set the logical CPU counts.

A More Complete Example

From thorough study of our workloads we determined:
1. FRED needs at least 4.25 engines' worth of power, and never more than 8,
2. WILMA needs at least 6.75 engines' worth of power, and never more than 11,
3. BARNEY needs at least 8.00 engines' worth of power, and never more than 10.

Sum of the needs = 4.25 + 6.75 + 8.00 = 19.00 engines.
We chose a safety factor of 20% => 23 shared engines.
Also we have partition BETTY, a 4-way dedicated.
So we bought a CEC with 27 physical engines and then did this:

Partition	Shr/Ded	E needed	Weight	E calculation	E achieved	LPUs
FRED	shared	425%	43	2300 * (43/191)	518%	8
WILMA	shared	675%	68	2300 * (68/191)	819%	11
BARNEY	shared	800%	80	2300 * (80/191)	963%	10
BETTY	dedicated	-	-	-	-	4

Logical CPU overcommit ratio = (8+11+10) / 23 = 1.26.

Ways Things Go Wrong, part 1

- Failure to set entitlement high enough.
  - 4-member z/OS virtual sysplex running on z/VM
  - Each z/OS guest is a virtual 2-way
  - How much power does each z/OS guest realistically minimally require?
  - What will happen if the partition's entitlement is well below the workload's requirement?

Answer: if correct operation of the workload requires that the partition consume beyond its entitlement, the workload is exposed to failing if the CEC becomes constrained.

Ways Things Go Wrong, part 2

Shared partition with 12 logical CPUs

When other partitions are quiet, this partition could run 1200% busy.
When other partitions are active, this partition might get as little as 150%.

The system might perform erratically.
Users might be confused and unhappy.
Some workloads might fail.

Ways Things Go Wrong, part 3

Consider a CEC with 12 shared physical CPUs. S = 12

22 partitions.
205 logical CPUs. L = 205

What are the problems?

Q1: If the weights are about equal, about how much entitlement would each partition get?
A1: About (12/22) or about 50%.

Q2: About how many logical CPUs are in each partition?
A2: About (205/22) or about 10.

Q3: Do you see anything wrong with a logical 10-way having entitlement 50%?

High L/S is a cause of high overhead in PR/SM and of suspend time for the logical CPUs.

Ways Things Go Wrong, part 4

Consider this shared partition:

1. 12 logical CPUs
2. Entitlement 150%
3. Horizontal
4. Capped

Each logical CPU has an entitlement of (150%/12) = 12.5% of a physical CPU.

Because of the cap, each logical CPU will be held back to 12.5% busy.

Q1: What if a virtual 1-way guest wants to run 20% busy? Can it do so?

Q2: What if a virtual 2-way guest wants to run 30% busy? Can it do so?

Q3: What if there are many such 2-way guests on the system? What would happen?

z/VM Performance Toolkit Reports

- These reports are your friends:
  - LSHARACT report: tabulates entitlements
  - PHYSLOG report: tabulates physical CPU use
  - LPARLOG report: tabulates use by LPARs

Causes for Concern

```
1FCX306  Run 2013/12/13 09:12:36          LSHARACT
                                          Logical Partition Share
 From 2013/12/13 Initial
 To   2013/12/13 09:10:56
 For  (Not applicable)                    Result of LPARS Run
```

LPAR Data, Collected in Partition EPRF1

Physical PUs, Shared: CP- 34 ZAAP- 2 IFL- 16 ICF- 1 ZIIP- 3
Dedicated: CP- 8 ZAAP- 0 IFL- 0 ICF- 0 ZIIP- 0

Proc Type	Partition Name	LPU Count	LPAR Weight	Entlment	TypeCap	<LPU Total,%> Busy	Excess	LPU Conf	
CP	ECT2	4	10	188.9	---	...	...	o	
CP	EPLX1	8	60	1133.3	---	...	...	u	← why?
CP	EST1	8	10	188.9	---	...	...	o	
CP	EST2	6	10	188.9	---	...	...	o	
CP	EST3	6	30	566.7	---	...	...	-	
CP	FCFT	16	40	755.6	---	...	...	o	← exposed?
CP	K4	6	10	188.9	---	...	...	o	
CP	PHOS	5	10	188.9	---	...	...	o	
IFL	EPLX1	2	50	313.7	---	...	...	u	
IFL	EPLX2	8	45	282.4	---	...	...	o	
IFL	EPLX3	6	45	282.4	---	...	...	o	
IFL	ESTL1	7	50	313.7	---	...	...	o	← exposed?
IFL	EST3	4	25	156.9	---	...	...	o	
IFL	FCFT	2	40	251.0	---	...	...	u	

Changing the weights of EPLX1 and FCFT would move entitlement to FCFT without harming EPLX1.

The LSHARACT report tabulates entitlements. New in z/VM 6.3 Perfkit!

More Causes for Concern

```
1FCX306  Run 2013/11/15 13:43:01          LSHARACT
                                          Logical Partition Share
 From 2013/11/13 23:09:54
 To   2013/11/13 23:39:04
 For   1751 Secs 00:29:11                 Result of xxxxxxxx Run
```

LPAR Data, Collected in Partition xxxx1A

Physical PUs, Shared: CP- 12 ZAAP- 0 IFL- 0 ICF- 4 ZIIP- 0
Dedicated: CP- 0 ZAAP- 0 IFL- 0 ICF- 7 ZIIP- 0

Proc Type	Partition Name	LPU Count	LPAR Weight	Entlment	TypeCap	<LPU Total,%> Busy	Excess	LPU Conf	
CP	xxxx0A	12	20	12.4	---	25.4	13.0	o	
CP	xxxx0B	7	100	61.8	---	53.4	.0	o	
CP	xxxx0E	7	20	12.4	---	27.8	15.4	o	
CP	xxxx01	12	20	12.4	---	2.8	.0	o	
CP	xxxx04	7	10	6.2	---	6.8	.6	o	
CP	xxxx05	12	200	123.6	---	130.4	6.8	o	
CP	xxxx07	12	4	2.5	---	4.0	1.5	o	
CP	xxxx08	7	10	6.2	---	4.5	.0	o	
CP	xxxx09	12	100	61.8	---	76.1	14.3	o	
CP	xxxx1A	12	250	154.5	---	23.6	.0	o	← CAPPED
CP	xxxx11	12	500	309.0	---	95.4	.0	o	
CP	xxxx14	7	10	6.2	---	9.8	3.6	o	
CP	xxxx15	7	10	6.2	---	18.5	12.3	o	
CP	xxxx16	7	6	3.7	---	4.2	.5	o	
CP	xxxx17	12	500	309.0	---	49.0	.0	o	
CP	xxxx18	1	13	8.0	---	3.5	.0	-	
CP	xxxx19	7	20	12.4	---	16.6	4.2	o	
CP	xxxx21	9	75	46.3	---	15.6	.0	o	
CP	xxxx22	12	10	6.2	---	8.5	2.3	o	
CP	xxxx24	12	30	18.5	---	7.8	.0	o	
CP	xxxx25	7	4	2.5	---	4.3	1.8	o	
CP	xxxx28	12	30	18.5	---	.0	.0	o	

There are only 12 shared CPs but there are 22 partitions and 205 logical CPUs.

Entitlement is heavily diluted because there are so many partitions for only 12 shared physical CPs.

In addition, logical CPU counts are high, which dilutes entitlement within partitions and also causes excess PR/SM overhead.

Partition xxxx1A is especially in danger because it is capped (see FCX126 LPAR).

Effect of High L:S Ratio

```
1FCX302  Run 2013/11/15 13:37:38          PHYSLOG
                                          Real CPU Utilization Log
 From 2013/11/14 21:09:24
 To   2013/11/14 21:20:14
 For    650 Secs 00:10:50                  Result of xxxxxxxx Run
```

Interval End Time	Type	<PU Num> Conf	Ded	Total Weight	%LgclP	%Ovrhd	LpuT/L	%LPmgt	%Total	TypeT/L	
>>Mean>>	CP	12	0	1942	764.42	13.447	1.018	40.548	818.41	1.071	<- 7% PR/SM overhead on CPs
>>Mean>>	ICF	11	7	1249	727.70	.811	1.001	7.389	735.90	1.011	<- 1% PR/SM overhead on ICFs
>>Mean>>	>Sum	23	7	3191	1492.1	14.258	1.010	47.937	1554.3	1.042	
21:09:44	CP	12	0	1942	667.86	14.968	1.022	45.666	728.49	1.091	
21:09:44	ICF	11	7	1249	716.72	1.229	1.002	8.435	726.38	1.013	
21:09:44	>Sum	23	7	3191	1384.6	16.197	1.012	54.100	1454.9	1.051	
21:10:06	CP	12	0	1942	758.51	15.297	1.020	46.163	819.97	1.081	
21:10:06	ICF	11	7	1249	748.41	1.238	1.002	8.027	757.68	1.012	
21:10:06	>Sum	23	7	3191	1506.9	16.535	1.011	54.190	1577.7	1.047	

On CPs, this CEC's L:S ratio is (205/12) or 17.1.
Usually we see ratios in the neighborhood of 1.5 to 2.0.

New in z/VM 6.3 Perfkit!
You can see CEC busy, by physical CPU type, as a function of time.

High L:S Ratio Plus Capping

```
1FCX126  Run 2013/11/15 13:37:38          LPAR
                                          Logical Partition Activity
 From 2013/11/14 21:09:24
 To   2013/11/14 21:20:14
 For    650 Secs 00:10:50                  Result of xxxxxxxx Run
```

LPAR Data, Collected in Partition xxxx1A

Processor type and model : 2097-712
Nr. of configured partitions: 45
Nr. of physical processors : 23
Dispatch interval (msec) : dynamic

Partition	Nr.	Upid	#Proc	Weight	Wait-C	Cap	%Load	CPU	%Busy	%Ovhd	%Susp	%VMld	%Logld	Type	TypeCap
xxxx1A	16	26	14	250	NO	YES	2.9	0	6.0	.2	13.2	5.7	6.6	CP	---
				250		YES	...	1	5.0	.1	14.0	4.9	5.7	CP	---
				250		YES	...	2	5.1	.1	15.1	5.0	5.9	CP	---
				250		YES	...	3	4.8	.1	12.1	4.7	5.4	CP	---
				250		YES	...	4	4.8	.1	14.4	4.7	5.5	CP	---
				250		YES	...	5	4.5	.0	13.8	4.4	5.1	CP	---
				250		YES	...	6	4.2	.1	10.7	4.1	4.5	CP	---
				250		YES	...	7	3.9	.0	11.4	3.8	4.3	CP	---
				250		YES	...	8	4.0	.1	11.7	3.9	4.4	CP	---
				250		YES	...	9	4.1	.1	12.3	4.0	4.6	CP	---
				250		YES	...	10	4.0	.0	11.7	3.9	4.5	CP	---
				250		YES	...	11	4.1	.0	12.0	4.0	4.6	CP	---
				250		NO	...	12	6.3	.2	.3	6.1	6.1	ICF	---
				250		NO	...	13	6.5	.1	.2	6.4	6.4	ICF	---

Nicely Done!

```
1FCX306  Run 2013/11/20 11:20:55          LSHARACT
                                          Logical Partition Share
 From 2013/11/20 15:01:48
 To   2013/11/20 15:14:48
 For    780 Secs 00:13:00                  Result of xxxxxx Run
```

LPAR Data, Collected in Partition xxxxxxxx

Physical PUs, Shared: CP- 0 ZAAP- 0 IFL- 7 ICF- 0 ZIIP- 0
Dedicated: CP- 0 ZAAP- 0 IFL- 10 ICF- 0 ZIIP- 0

Proc Type	Partition Name	LPU Count	LPAR Weight	Entlment	TypeCap	<LPU Total,%> Busy	Excess	LPU Conf
IFL	xxxxxx	1	10	9.9	---	.5	.0	-
IFL	xxxxxxxx	3	300	295.8	---	159.7	.0	-
IFL	xxxxxxxx	5	400	394.4	---	335.8	.0	o
IFL	xxxxxxxx	10	DED	...	---	...	...	.

LPU counts match up well to entitlements.
L/S = 9/7 = 1.29.

I Thought This Was Brilliant

```
1FCX306  Run 2013/08/27 11:02:00          LSHARACT
                                          Logical Partition Share
 From 2013/08/24 23:52:00
 To   2013/08/25 00:50:00
 For   3480 Secs 00:58:00                 Result of PP2 Run
 ------------------------------------------------------------------------------

  LPAR Data, Collected in Partition PP2

  Physical PUs, Shared: CP-  3  ZAAP-  0  IFL- 11  ICF-  1  ZIIP-  0
             Dedicated: CP-  0  ZAAP-  0  IFL-  0  ICF-  0  ZIIP-  0

  ____ .             .      .        .       .      .      . .
  Proc Partition   LPU   LPAR                  <LPU Total,%> LPU
  Type Name      Count Weight Entlment TypeCap   Busy Excess Conf
  IFL  PPU           1      2     22.0     ...    1.8     .0 -
  IFL  PP1           2      8     88.0     ...    8.7     .0 o
  IFL  PP2           7     60    660.0     ...  374.5     .0 -
  IFL  PP6          11     30    330.0     ...  246.8     .0 o
```

PPU, PP1, and PP2 have fairly tight leashes… small distance from E to potential.
PP6 has a small requirement, but it gets to use everything no one else is using.
In other words, PP6 runs mostly on spare power.

What Are the Partitions Using?

```
1FCX202  Run 2013/11/20 11:20:55          LPARLOG
                                          Logical Partition Activity Log
 From 2013/11/20 15:01:48
 To   2013/11/20 15:14:48
 For    780 Secs 00:13:00                  Result of xxxxxxxx Run
```

Interval End Time	<Partition-> Name	Nr.	Upid	#Proc	Weight	Wait-C	Cap	%Load	<- Load per Log. Processor --> %Busy	%Ovhd	%Susp	%VMld	%Logld	Type	TypeCap
>>Mean>>	xxxxxxx1	10	27	1	10	NO	NO	.0	.5	.0	...	...	...	IFL	---
>>Mean>>	xxxxxxx2	12	24	3	300	NO	NO	9.4	53.2	.5	...	...	...	IFL	---
>>Mean>>	xxxxxxx3	18	25	10	DED	YES	NO	58.8	100.0	.0	...	...	...	IFL	---
>>Mean>>	xxxxxxx4	19	23	5	400	NO	NO	19.8	67.2	.8	1.8	66.1	67.3	IFL	---
>>Mean>>	Total	..	..	17	710	..	..	88.3	78.7	.3	...	...	...	..	---
15:02:18	xxxxxxx1	10	27	1	10	NO	NO	.0	.5	.0	...	...	...	IFL	---
15:02:18	xxxxxxx2	12	24	3	300	NO	NO	12.2	69.3	.4	...	...	...	IFL	---
15:02:18	xxxxxxx3	18	25	10	DED	YES	NO	58.8	100.0	.0	...	...	...	IFL	---
15:02:18	xxxxxxx4	19	23	5	400	NO	NO	22.7	77.1	.5	7.2	76.5	82.4	IFL	---
15:02:18	Total	..	..	17	710	..	..	94.0	83.9	.2	...	...	...	..	---
15:02:48	xxxxxxx1	10	27	1	10	NO	NO	.0	.5	.1	...	...	...	IFL	---
15:02:48	xxxxxxx2	12	24	3	300	NO	NO	9.9	56.0	.3	...	...	...	IFL	---
15:02:48	xxxxxxx3	18	25	10	DED	YES	NO	58.8	100.0	.0	...	...	...	IFL	---
15:02:48	xxxxxxx4	19	23	5	400	NO	NO	17.1	58.0	.9	1.2	56.8	57.5	IFL	---
15:02:48	Total	..	..	17	710	..	..	86.1	76.7	.3	...	...	...	..	---

Notes:

1. %Load: what fraction of the machine's physical capacity is being used by this partition?
2. %Busy: how busy is the average logical CPU of this partition?
3. This report is not terribly useful in mixed engine environments. Use FCX126 LPAR.

Summary

- Know your workloads' needs.
- Translate those needs into entitlements.
- Plan enough physical CPU to fulfill them.
- Add in a little spare.
- Add in your dedicated LPARs.
- Calculate those weights correctly.
- Be careful with capping!
- Use z/VM Performance Toolkit. It is your friend!