POWER7 Affinity & Performance Part 1

Nigel Griffiths
IBM Power Systems
Advanced Technology Support, Europe

tinyurl.com/AIXpert Blog

Local, Near & Far Memory – Blog Article Series

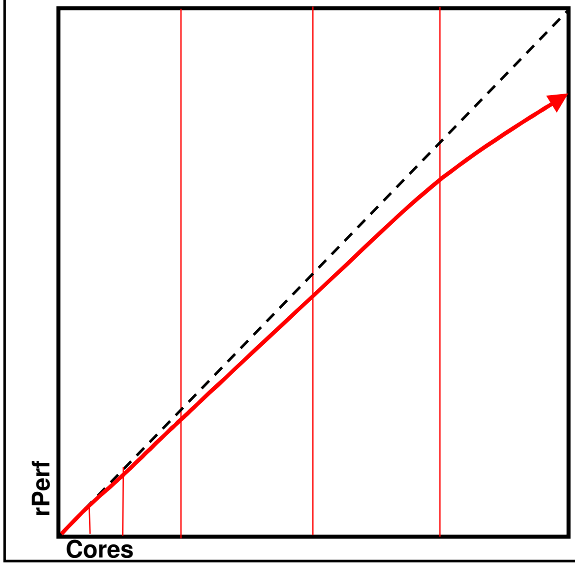
1. Large Power7 boxes more local memory
2. Virtual Machine CPU & Memory Lay Out
3. Scheduling processes to SMT & Virtual Processors
4. Aggressive Intelligent Threads
5. Low Entitlement has a Bad Side Effect
6. Too High a Virtual Processor number has a Bad Side Effect
7. VM placement also needs RAM
8. Dynamic LPAR changes can mess up your placement
9. Firmware Updates for better Affinity
10. Final of the table by Model
11. Why Local+Far on Lower End machines?
    - POWER7 Affinity Nine Conclusions
12. I have a 10 core POWER7 chip, eh!

- Plus AIX Virtual Processor Folding is Misunderstood
- Many others: nmon, Systems Director, VIOS, …

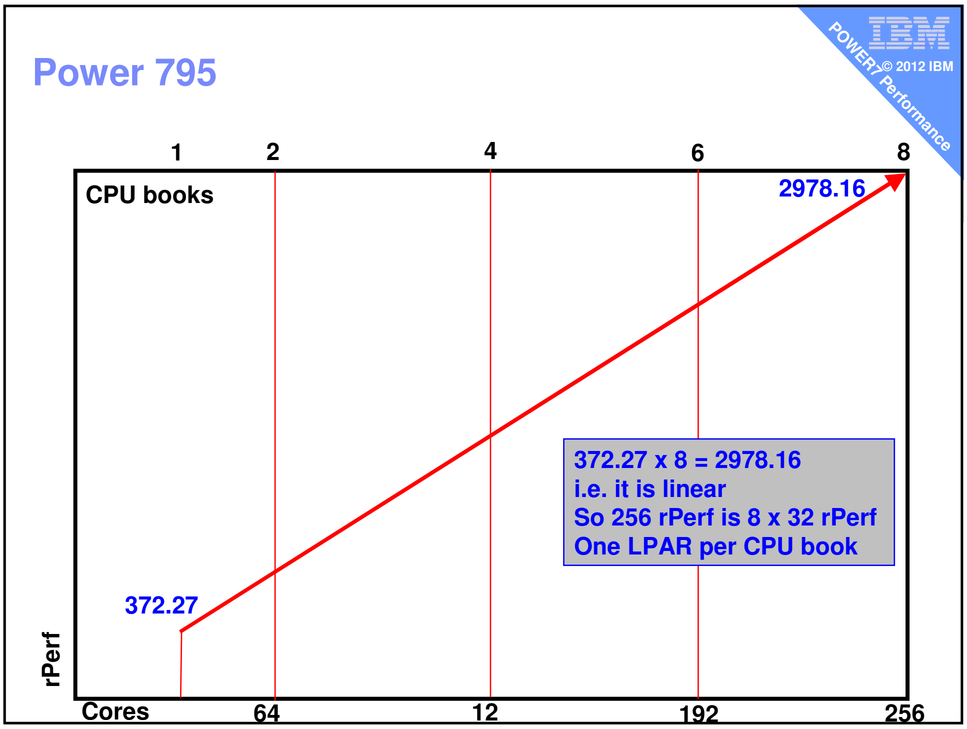
The SMP curve & Cache Boundaries

- Recap – we all know this!

Well Known SMP Curve

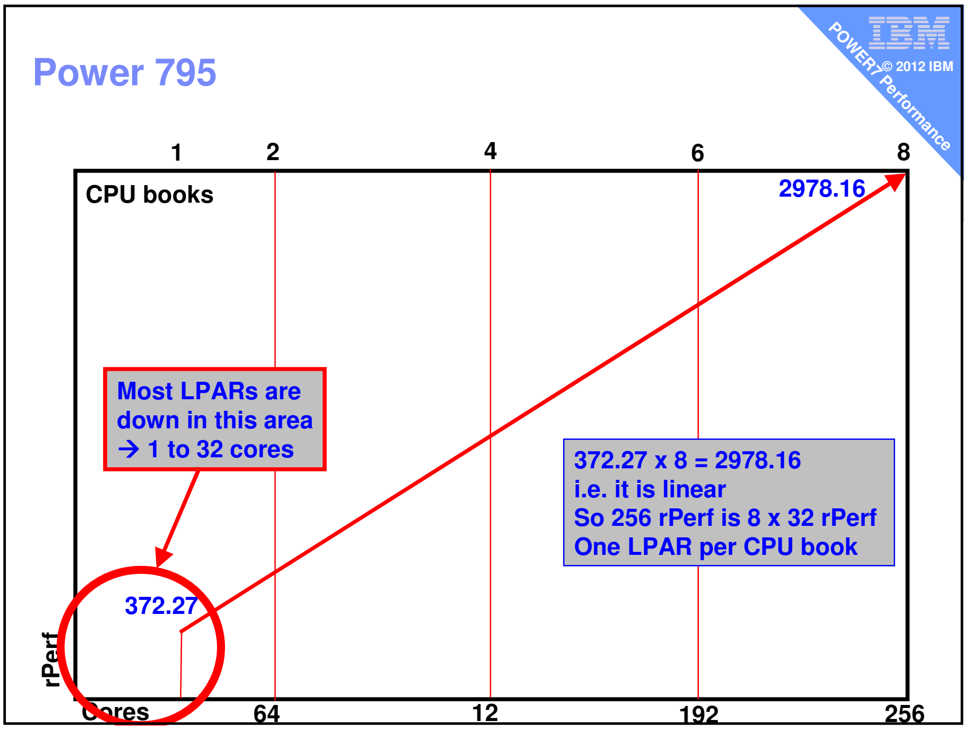


- Standard large machine don’t scale perfectly linearly
- Distances & comm’s & sharing caches
- Localising your VM can give you a “free boost”
- Not to scale and drawn by hand!



Power 795

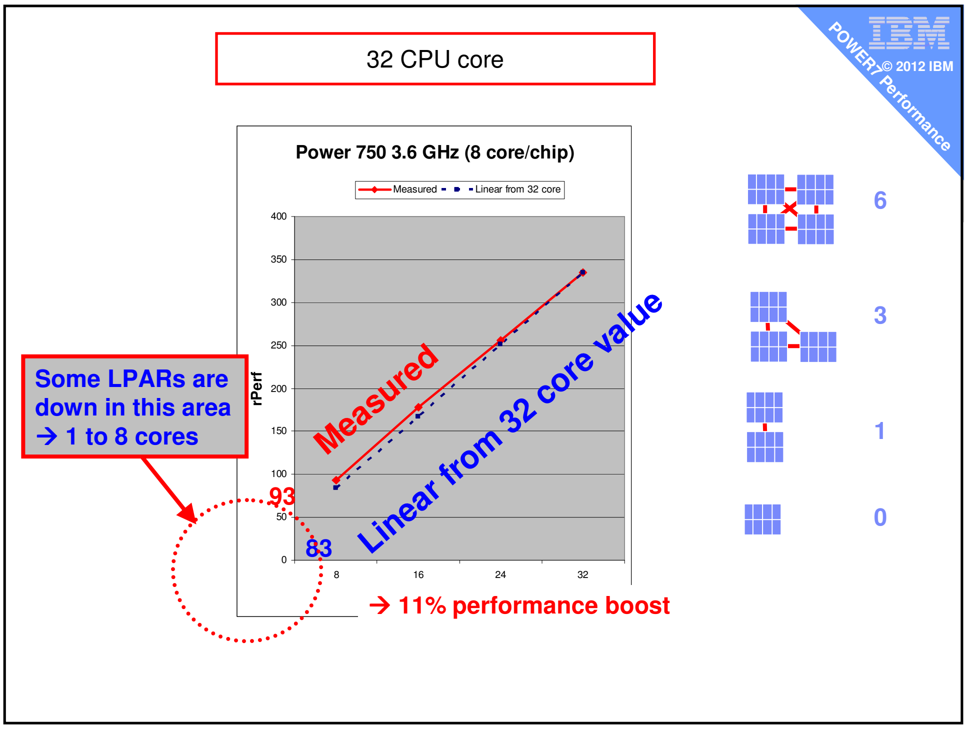
372.27 x 8 = 2978.16
i.e. it is linear
So 256 rPerf is 8 x 32 rPerf
One LPAR per CPU book



Power 795

Most LPARs are down in this area → 1 to 32 cores

372.27 x 8 = 2978.16
i.e. it is linear
So 256 rPerf is 8 x 32 rPerf
One LPAR per CPU book

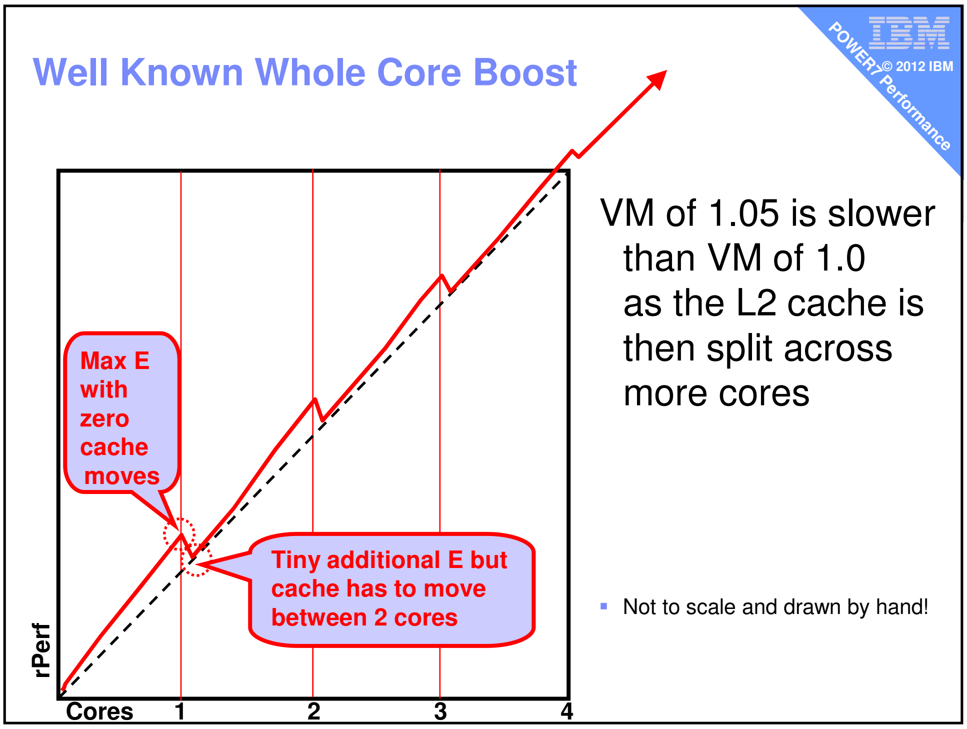


32 CPU core

Power 750 3.6 GHz (8 core/chip)

Some LPARs are down in this area → 1 to 8 cores

→ 11% performance boost



Well Known Whole Core Boost

Max E with zero cache moves

Tiny additional E but cache has to move between 2 cores

VM of 1.05 is slower than VM of 1.0 as the L2 cache is then split across more cores

- Not to scale and drawn by hand!

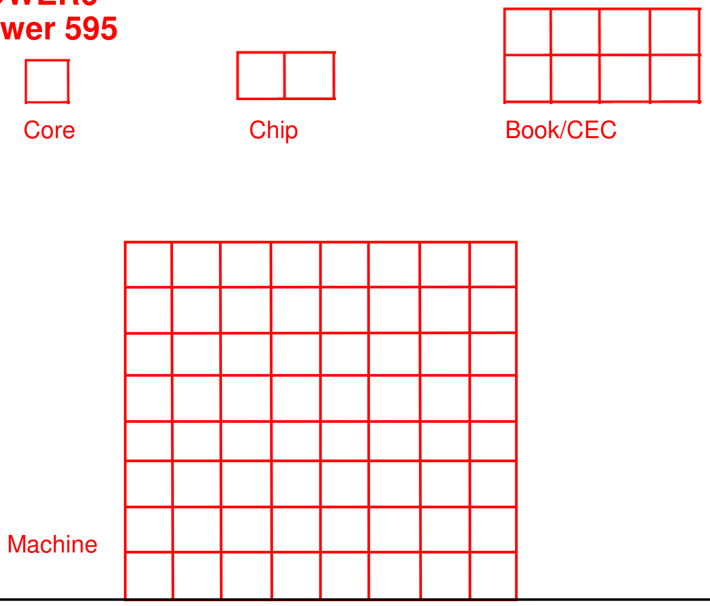
POWER7

- 8 cores per chip
- Also 4 core & 6 core chips
- Increases the strain on the memory bus

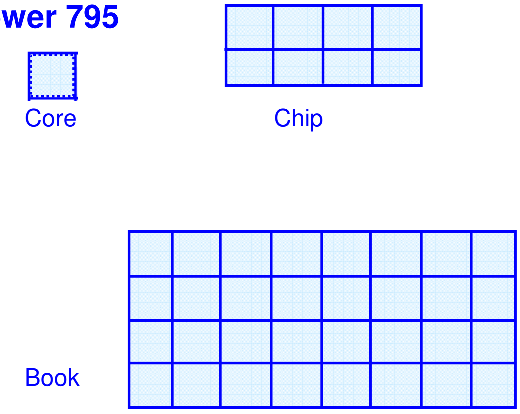
POWER6

- 2 cores per chip

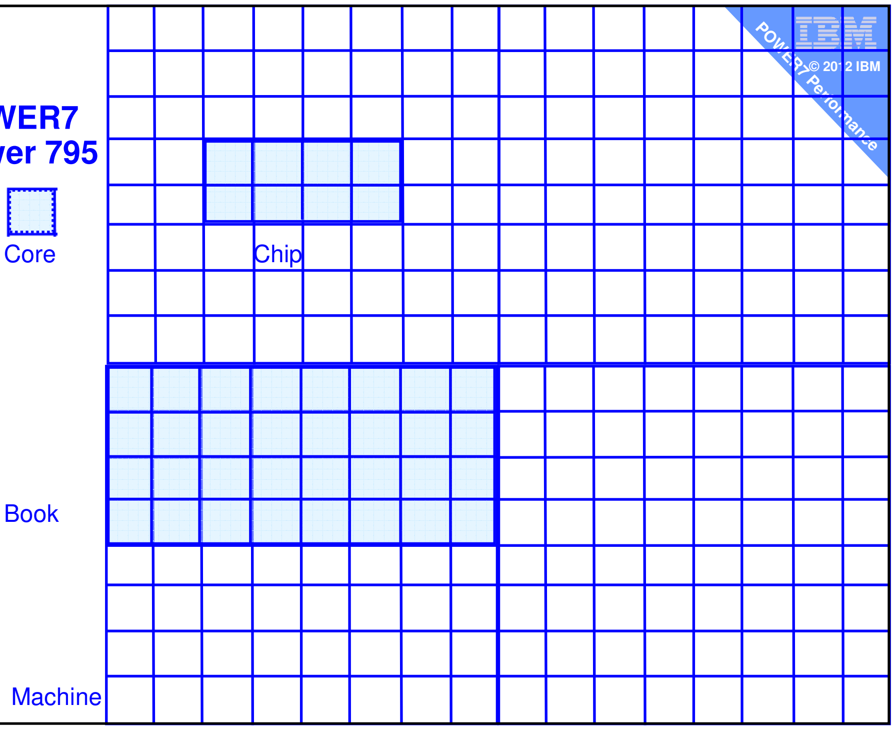
POWER6
Power 595



POWER7
Power 795



POWER7
Power 795



Reminders about rPerf for sizing

- Relative Performance measurement
- For comparing POWER machines and “bangs per buck”
- But a common source if misconception

rPerf for Sizing

- The assumptions have been forgotten = causes serious Sizing issues
- POWER6 10 CPU VM with rPerf=100
- POWER7 10 CPU VM with rPerf=150

Which is true?

A. So application is 50% faster
B. Utilisation will drop by 33%
C. Batch will finish 33% quicker
D. All of the above
E. None of the above

rPerf for Sizing

- The assumptions have been forgotten
  = causes serious Sizing issues

- POWER6 10 CPU VM with rPerf=100
- POWER7 10 CPU VM with rPerf=150

Which is true?

A. So application is 50% faster
B. Utilisation will drop by 33%
C. Batch will finish 33% quicker
D. All of the above
E. None of the above

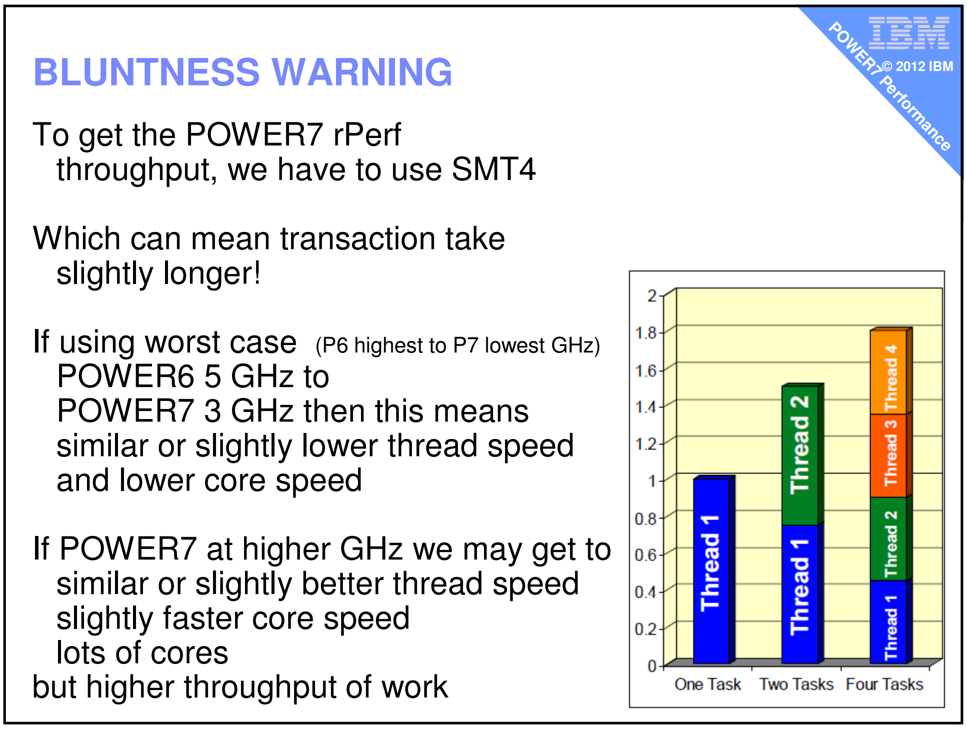
←These are speed statements but rPerf is all about Throughput. Also comes with many assumptions …

rPerf for Sizing - Ten Golden Rules

1. Highly threaded workloads – CPUs x SMT4 x 2

2.	Well tuned system	– retuned from scratch
3.	Full Spec RAM	– no empty slots + lots of GBs
4.	No Disk Issues	– 100’s of disks, no bottlenecks
5.	No Network Issues	– tuned to zero bottlenecks
6.	Current App. software	– not previous generation
7.	Latest AIX6/7	– latest TL + all service packs
8.	Large LPARs	– no micro-partitions
9.	Firmware	– latest
10.	Bug Free	– user willing to fix

AND the workload can be proved to be the same before and after = same number of transactions, dialogue steps, same size database, same SQL, record batch processed per minute/hour.



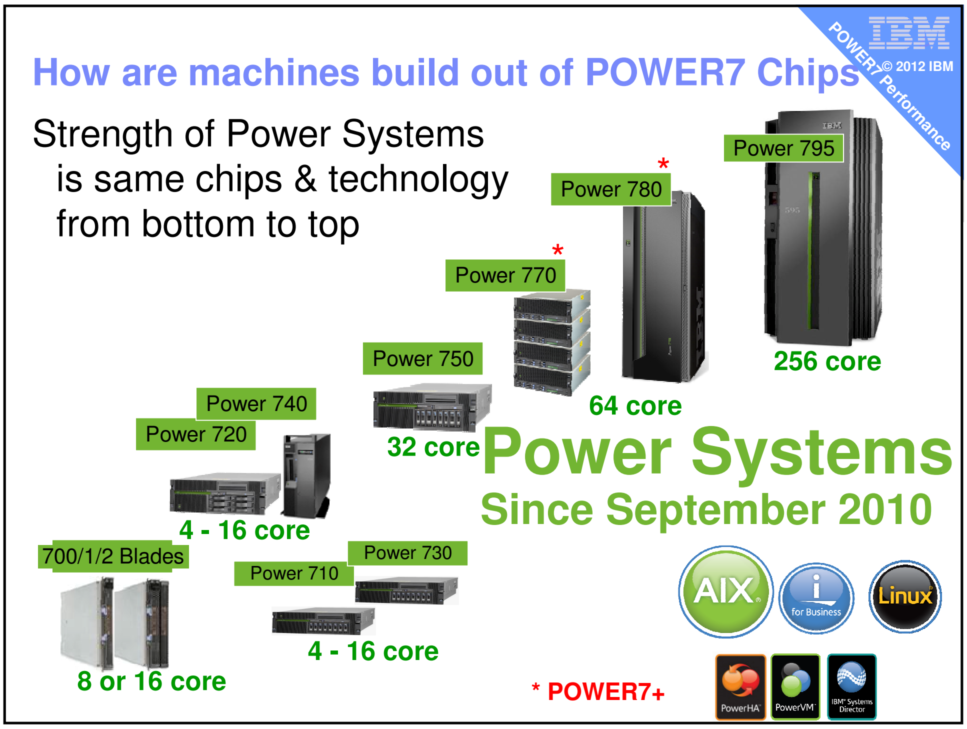
BLUNTNESS WARNING

To get the POWER7 rPerf
throughput, we have to use SMT4

Which can mean transaction take
slightly longer!

If using worst case (P6 highest to P7 lowest GHz)
POWER6 5 GHz to
POWER7 3 GHz then this means
similar or slightly lower thread speed
and lower core speed

If POWER7 at higher GHz we may get to
similar or slightly better thread speed
slightly faster core speed
lots of cores
but higher throughput of work



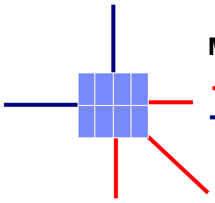
How are machines build out of POWER7 Chips

Strength of Power Systems
is same chips & technology
from bottom to top

Power Systems
Since September 2010

* POWER7+

Bandwidth and Buses

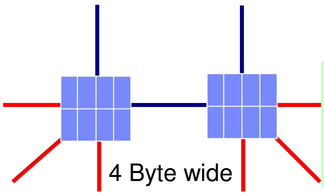


Memory Buses

3 XYZ bus

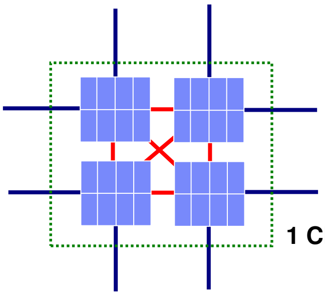
2 AB bus

2, 4 or 8 Byte wide



Lower End = Far memory
Two POWER7 chips
directly connected using AB
XYZ not used

New POWER7+ and
Power 795 book
Four POWER7 chips directly
Eight links to the
other CPU books



3 XYZ bus

8 AB bus

8 Byte wide

CPU & Memory Affinity by POWER7 Model

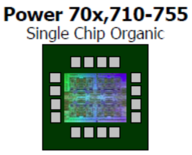
Model/RAM Access	**POWER7 Chips/CPU**	**Local**	**Near**	**Far**
Power7 blades	1	Same Chip		-
Power7 blades	2	Same Chip		Other Chip
Power 710	1	Same Chip		-
Power 730	2	Same Chip		Other Chip
Power 720	1	Same Chip		-
Power 740	2	Same Chip		Other Chip
Power 750/755	1 to 4	Same Chip		Other Chip
Power 770/780	2 to 8 / 64	Same Chip	Other 1 Chip but same CEC	Different CEC
Power 770/780 POWER7+ (dual core sockets)	4 to 16 / 128	Same Chip	Other 3 Chips but same CEC	Different CEC
Power 795	4 to 32 / 256	Same Chip	Other Chip but same CPU Book	Different CPU Book

Bandwidth and busses

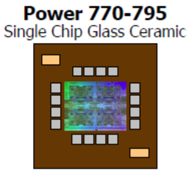
- Local memory
  - Memory 68GB/s per memory controller (P7 has 2)
  - Power 770/780/795 uses both – the rest uses one
- Near memory bus → XYZ Intra-node
  - Only on 770/780 two chips* & 795 four chips, 8 byte wide
  - ~40 - 50 GB/s depends on the model
- Far memory bus → AB Inter-node
  - Power710-750 between chips = 4 byte wide (reduced cost)
  - Power 770/780 CEC & 795 book = 8 byte wide
  - 23 – 26 GB/s depends on the model

* POWER7+ models use the 4 core per chip

POWER7 mounting



1 Memory Controller = 68 GB/s
4 Byte AB memory buses
between chips = 23 GB/s each



2 Memory Controllers total 136 GB/s
8 Byte XYZ memory buses in the node
8 Byte AB busses between nodes

Perspective & Perception

Distance	NOT	but more like
Local	Good	**Blisteringly fast**
Near	Bad	**Excellent**
Far	Ugly	**Good**

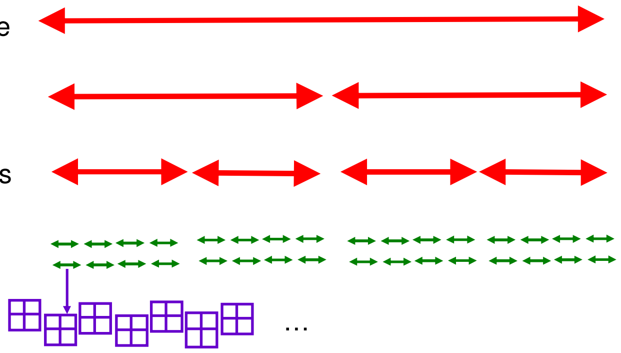
How does this effect me?

- Memory better if VM is:
  - Inside the 8 core POWER7 size
  - Inside the 16 core Power 770/780 CEC drawer *
  - Inside the 32 core Power 795 CPU book
- Note the VM size is Virtual Processor not Entitlement
- Used to determine the SRADs …

* 32 core POWER7+ Power 780

SRAD – eh!

- Scheduler Resource Affinity Domains
  - Groups of efficient CPU + memory
  - Hypervisor decides what you get
  - AIX works within SRADs to place processes with fastest RAM
  - Hierarchy of resources
  - Whole machine
  - CEC / Book
  - POWER7 chips
  - Cores
  - Threads



Blog Part 2

In POWER Logical Virtual Machines

- So how can we investigate
  - lssrad -av
  - mpstat –d
  - topas –M
  - svmon …
  - Lets see some examples

lssrad -av

```
# lssrad -av
REF1    SRAD          MEM         CPU
0
            0    29224.00        0-27
            1     2490.00        28-31
```

lssrad -av

- Only options that make sense!

REF1

- backplane, CEC drawer, or CPU book
- Why REF1? = Reference !!!!

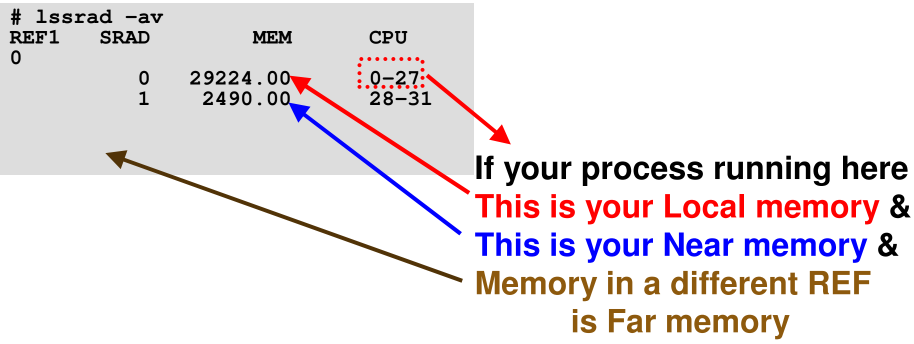
SRAD → CPU+RAM group

MEM → Megabytes!

CPU

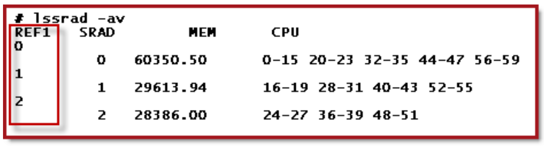
- Logical CPU number
- Assuming SMT=4

lssrad -av



```
# lssrad -av
REF1    SRAD          MEM         CPU
0
            0    29224.00        0-27
            1     2490.00        28-31
```

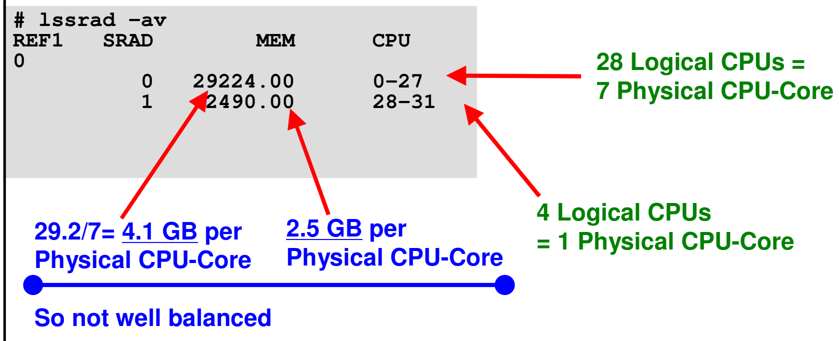
If your process running here
This is your Local memory &
This is your Near memory &
Memory in a different REF
is Far memory



```
# lssrad -av
REF1   SRAD        MEM        CPU
0
          0    60350.50       0-15 20-23 32-35 44-47 56-59
1
          1    29613.94       16-19 28-31 40-43 52-55
2
          2    28386.00       24-27 36-39 48-51
```

Example from Power 750 Local + Far

lssrad -av



```
# lssrad -av
REF1   SRAD        MEM      CPU
0
          0   29224.00      0-27
          1    2490.00      28-31
```

28 Logical CPUs =
7 Physical CPU-Core

29.2/7= 4.1 GB per
Physical CPU-Core

2.5 GB per
Physical CPU-Core

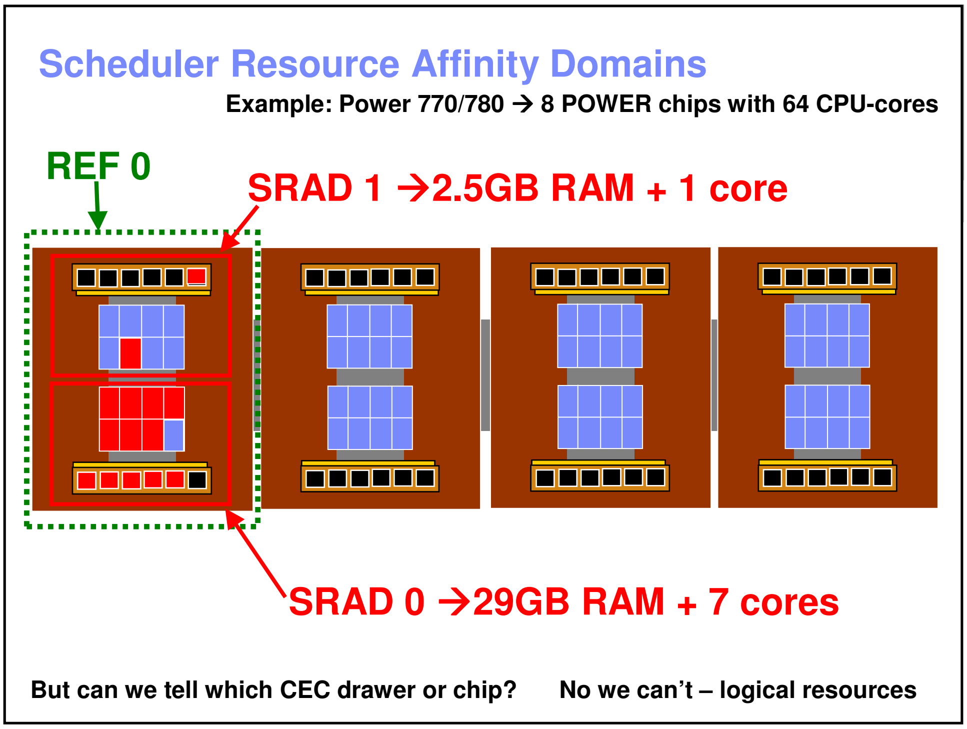
4 Logical CPUs
= 1 Physical CPU-Core

So not well balanced

Total memory here= 31714 MB
but HMC 32 GB= 32768 MB
~3% less

BLOG 2

Investigating commands



Scheduler Resource Affinity Domains

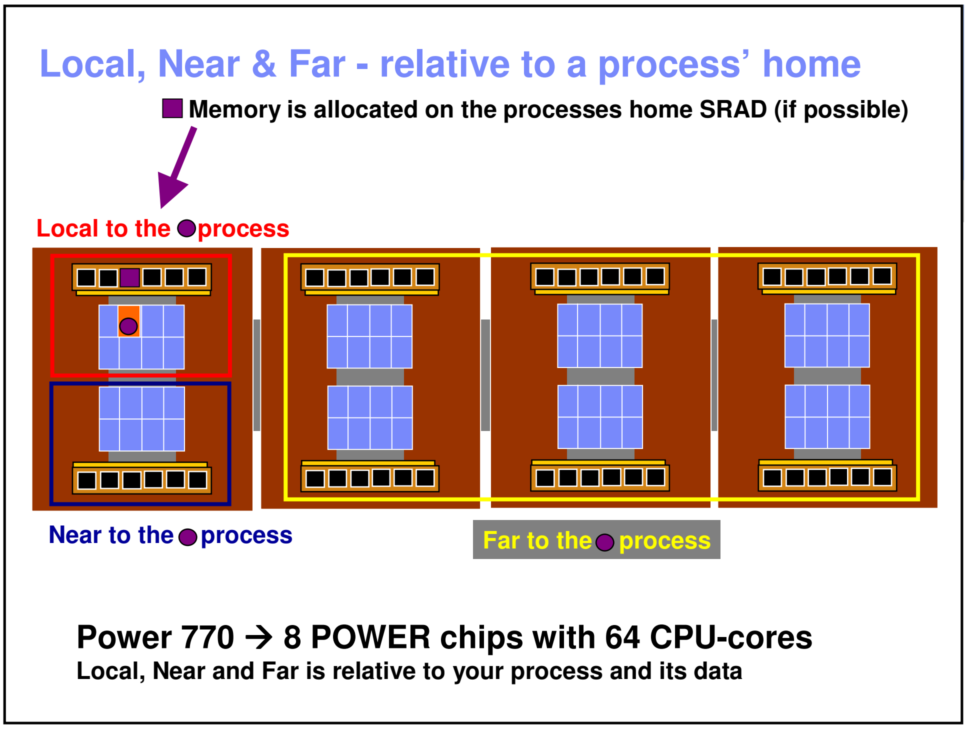
Example: Power 770/780 → 8 POWER chips with 64 CPU-cores

REF 0

SRAD 1 →2.5GB RAM + 1 core

SRAD 0 →29GB RAM + 7 cores

But can we tell which CEC drawer or chip? No we can't – logical resources



Local, Near & Far - relative to a process' home

🟪 Memory is allocated on the processes home SRAD (if possible)

Local to the ● process

Near to the ● process

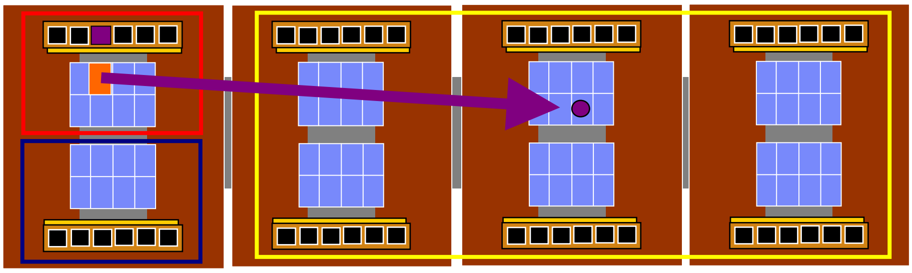
Far to the ● process

Power 770 → 8 POWER chips with 64 CPU-cores

Local, Near and Far is relative to your process and its data

Local, Near & Far - relative to a process' home

Local to the ● process



Near to the ● process

Far to the ● process

If a process is schedule away from it's home SRAD

Dedicated CPU 256 way Power 795 + 4TB RAM

```
# lparstat -i
Node Name : test1
Partition Name : test1new
Partition Number : 2
Type : Dedicated
Mode : Capped
Entitled Capacity : 256.00
Partition Group-ID : 32770
Shared Pool ID : -
Online Virtual CPUs : 256
Maximum Virtual CPUs : 256
Minimum Virtual CPUs : 1
Online Memory : 4121088 MB
Maximum Memory : 4194304 MB
Minimum Memory : 256 MB
Variable Capacity Weight : -
Minimum Capacity : 1.00
Maximum Capacity : 256.00
Capacity Increment : 1.00
Maximum Physical CPUs in system : 256
Active Physical CPUs in system : 256
Active CPUs in Pool : -
Shared Physical CPUs in system : 0 [Note: no SMT number shown as it is off]
Maximum Capacity of Pool : 0
Entitled Capacity of Pool : 0
Unallocated Capacity : -
Physical CPU Percentage : 100.00%
Unallocated Weight : -
Memory Mode : Dedicated
Total I/O Memory Entitlement : -
Variable Memory Capacity Weight : -
Memory Pool ID : -
Physical Memory in the Pool : -
Hypervisor Page Size : -
Unallocated Variable Memory Capacity Weight: -
Unallocated I/O Memory entitlement : -
Memory Group ID of LPAR : -
Desired Virtual CPUs : 256
Desired Memory : 4121088 MB
Desired Variable Capacity Weight : -
Desired Capacity : 256.00
Target Memory Expansion Factor : -
Target Memory Expansion Size : -
Power Saving Mode : Disabled
```

```
# lssrad -av
REF1 SRAD MEM CPU
0
    0 94341.00 0 4 8 12 16 20 24 28
    1 94711.00 32 36 40 44 48 52 56 60
    2 94711.00 64 68 72 76 80 84 88 92
    3 94711.00 96 100 104 108 112 116 120 124
1
    4 94711.00 128 132 136 140 144 148 152 156
    5 94695.00 160 164 168 172 176 180 184 188
    6 94695.00 192 196 200 204 208 212 216 220
    7 94695.00 224 228 232 236 240 244 248 252
2
    8 94695.00 256 260 264 268 272 276 280 284
    9 94695.00 288 292 296 300 304 308 312 316
    10 94695.00 320 324 328 332 336 340 344 348
    11 94695.00 352 356 360 364 368 372 376 380
3
    12 94695.00 384 388 392 396 400 404 408 412
    13 94695.00 416 420 424 428 432 436 440 444
    14 94695.00 448 452 456 460 464 468 472 476
    15 94695.00 480 484 488 492 496 500 504 508
4
    16 93970.94 512 516 520 524 528 532 536 540
    17 45421.00 544 548 552 556 560 564 568 572
    18 94695.00 576 580 584 588 592 596 600 604
    19 94695.00 608 612 616 620 624 628 632 636
5
    20 94695.00 640 644 648 652 656 660 664 668
    21 94695.00 672 676 680 684 688 692 696 700
    22 94695.00 704 708 712 716 720 724 728 732
    23 94695.00 736 740 744 748 752 756 760 764
6
    24 94695.00 768 772 776 780 784 788 792 796
    25 94695.00 800 804 808 812 816 820 824 828
    26 94695.00 832 836 840 844 848 852 856 860
    27 94864.00 864 868 872 876 880 884 888 892
7
    28 94896.00 896 900 904 908 912 916 920 924
    29 94880.00 928 932 936 940 944 948 952 956
    30 94896.00 960 964 968 972 976 980 984 988
    31 94309.00 992 996 1000 1004 1008 1012 1016 1020
```

**Note: SMT=1 CPU numbers by first
Logical CPU number 0, 4, 8, 12, …**

mpstat -d

```
mpstat -d 1 6

System configuration: lcpu=8 ent=1.0 mode=Uncapped

cpu      cs     ics   bound      rq    push  S3pull   S3grd  S0rd  S1rd  S2rd  S3rd  S4rd  S5rd    ilcs    vlcs S3hrd S4hrd S5hrd
  0     162      70       0       0       0       0       0  99.5   0.5   0.0   0.0   0.0   0.0      14      43 100.0   0.0   0.0
  1      42      28       0       0       0       0       0 100.0   0.0   0.0   0.0   0.0   0.0      15      19 100.0   0.0   0.0
  2       0       0       0       0       0       0       0     -     -     -     -     -     -       0       0     -     -     -
  3       0       0       0       0       0       0       0     -     -     -     -     -     -       0       0     -     -     -
  4       0       0       0       0       0       0       0     -     -     -     -     -     -       0       0     -     -     -
  5       0       0       0       0       0       0       0     -     -     -     -     -     -       0       0     -     -     -
  7       6       6       0       0       0       0       0     -     -     -     -     -     -       0      19 100.0   0.0   0.0
ALL     210     104       0       0       0       0       0  99.6   0.4   0.0   0.0   0.0   0.0      29      81 100.0   0.0   0.0
---------------------------------------------------------------------------------------------------------------------
  0     180      73       0       0       0       0       0  99.6   0.0   0.0   0.4   0.0   0.0      19      42 100.0   0.0   0.0
  1      38      29       0       0       0       0       0  96.4   3.6   0.0   0.0   0.0   0.0      18      19 100.0   0.0   0.0
  2       1       1       0       0       0       0       0   0.0   0.0   0.0 100.0   0.0   0.0       0       0 100.0   0.0   0.0
  3       1       1       0       0       0       0       0   0.0 100.0   0.0   0.0   0.0   0.0       0       0 100.0   0.0   0.0
  4       1       1       0       0       0       0       0   0.0   0.0   0.0 100.0   0.0   0.0       0       0 100.0   0.0   0.0
  5       1       1       0       0       0       0       0   0.0 100.0   0.0   0.0   0.0   0.0       0       0 100.0   0.0   0.0
  6       1       1       0       0       0       0       0   0.0   0.0   0.0 100.0   0.0   0.0       0       1 100.0   0.0   0.0
  7       8       8       0       0       0       0       0   0.0 100.0   0.0   0.0   0.0   0.0       0      27 100.0   0.0   0.0
ALL     231     115       0       0       0       0       0  97.1   1.5   0.0   1.5   0.0   0.0      37      89 100.0   0.0   0.0
---------------------------------------------------------------------------------------------------------------------
```

s[0-5]rd = % of thread re-dispatches within a scheduling affinity domain

s[3-5]hrd % of thread dispatches on this logical processor

“-” → intelligent threads in action

Blog Part 2

mpstat –d

- s[0-5]rd = % of thread re-dispatches within a scheduling affinity domain
  - Processes home on this Logical Processor are running in these places
  - S0rd – same Logical processor (SMT) **These are currently undocumented**
  - S1rd – same Core
  - S2rd – MCM ?? Only on the Power 775
  - S3rd – same POWER7 chip
  - S4rd – same CEC
  - S5rd – other CEC
- s[3-5]hrd % of thread dispatches on this Logical processor
  - This logical processor is running other SRADs workload
  - S3hrd – Local
  - S4hrd – Near
  - S5hrd – Far
- “-” = intelligent threads in action – logical processor is off

mpstat –d 1 999 (thread dispatch = memory access)

Thread dispatches

- S3hrd → Local
- S4hrd → Near
- S5hrd → Far

Relative to home SRAD if away from home then Home memory access is also Local, Near, Far

Note: vlcs highlights
how often used

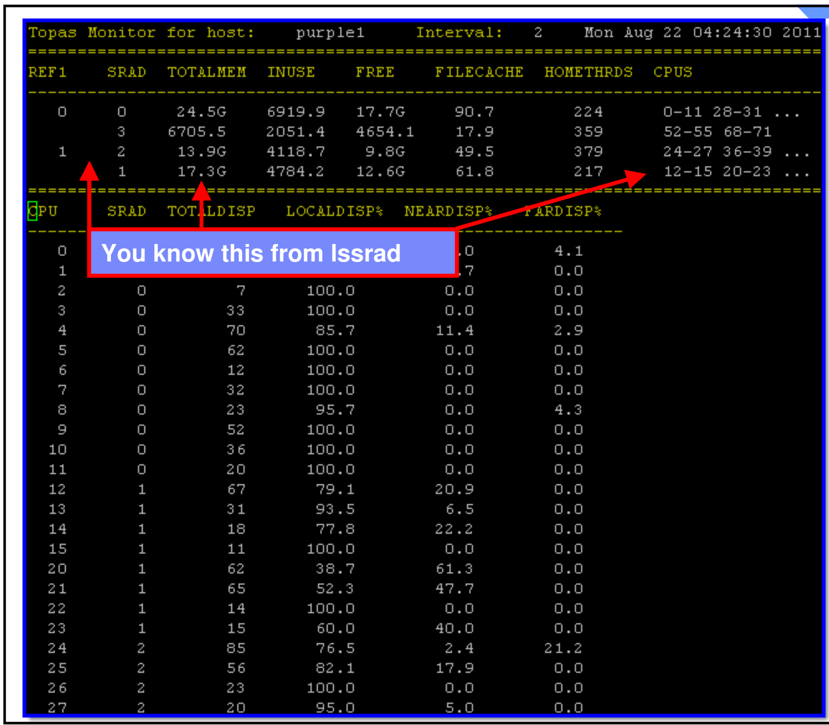
```
# mpstat -d 1 999
. . .
cpu      . . . S3hrd S4hrd S5hrd
 16      . . . 100.0   0.0   0.0
 17      . . .  80.0   0.0  20.0
 18      . . .     -     -     -
 19      . . .     -     -     -
 20      . . .  96.9   3.1   0.0
 21      . . . 100.0   0.0   0.0
 22      . . .     -     -     -
 23      . . .     -     -     -
 24      . . . 100.0   0.0   0.0
 25      . . . 100.0   0.0   0.0
 26      . . .     -     -     -
 27      . . .     -     -     -
 28      . . . 98.0   2.0   0.0
 29      . . .    -     -     -
 30      . . .    -     -     -
 31      . . .    -     -     -
 32      . . . 100.0   0.0   0.0
 33      . . . 100.0   0.0   0.0
 34      . . .     -     -     -
 35      . . .     -     -     -
 36      . . .  98.4   1.6   0.0
 37      . . .     -     -     -
 38      . . .     -     -     -
 39      . . .     -     -     -
 40      . . . etc
```

topas –M or
topas and the hit M or
nmon then ~ then M

Power 770 …

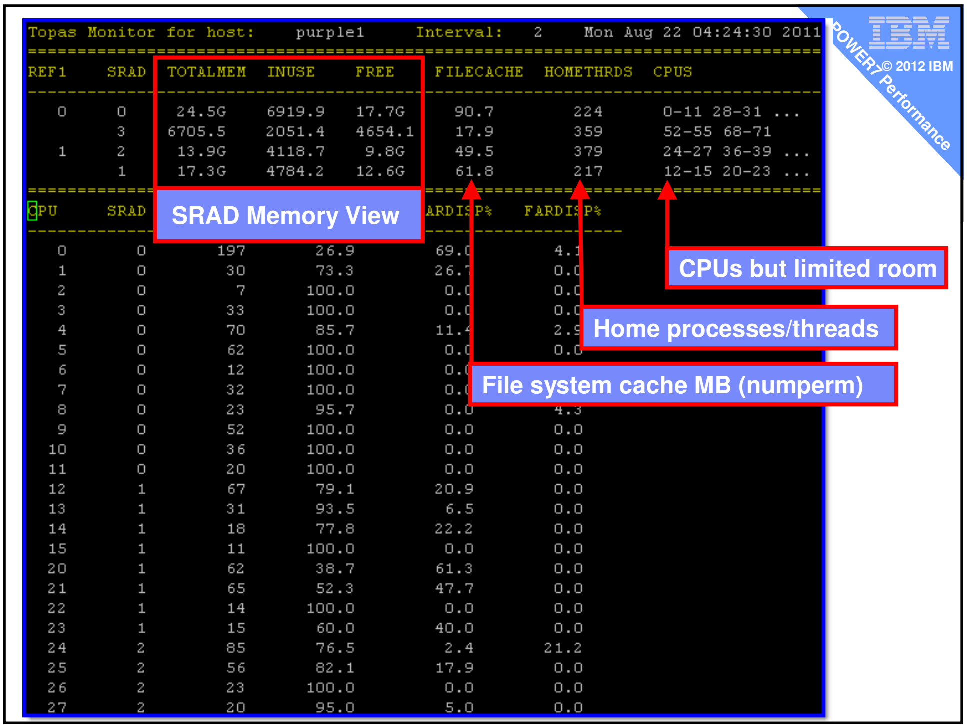
```
# lssrad -av
REF1    SRAD         MEM       CPU
0
           0    25054.75       0-11 28-31 40-43 56-59 72-75
           3     6705.50       52-55 68-71
1
           1    17679.00       12-15 20-23 32-35 44-47 60-63 76-79 104-107
           2    14193.00       24-27 36-39 48-51 64-67
```

Yes it is a bit of a mess to highlight some things



```
Topas Monitor for host:     purple1      Interval:   2    Mon Aug 22 04:24:30 2011
===================================================================================
REF1     SRAD  TOTALMEM  INUSE     FREE    FILECACHE  HOMETHRDS  CPUS
-----------------------------------------------------------------------------------
   0       0     24.5G   6919.9    17.7G     90.7        224      0-11 28-31 ...
           3    6705.5   2051.4   4654.1     17.9        359      52-55 68-71
   1       2     13.9G   4118.7     9.8G     49.5        379      24-27 36-39 ...
           1     17.3G   4784.2    12.6G     61.8        217      12-15 20-23 ...
===================================================================================
CPU      SRAD  TOTALDISP   LOCALDISP%  NEARDISP%   FARDISP%
----------------------------------------------------------
   0                                          .0       4.1
   1                                          .7       0.0
   2         0         7      100.0          0.0       0.0
   3         0        33      100.0          0.0       0.0
   4         0        70       85.7         11.4       2.9
   5         0        62      100.0          0.0       0.0
   6         0        12      100.0          0.0       0.0
   7         0        32      100.0          0.0       0.0
   8         0        23       95.7          0.0       4.3
   9         0        52      100.0          0.0       0.0
  10         0        36      100.0          0.0       0.0
  11         0        20      100.0          0.0       0.0
  12         1        67       79.1         20.9       0.0
  13         1        31       93.5          6.5       0.0
  14         1        18       77.8         22.2       0.0
  15         1        11      100.0          0.0       0.0
  20         1        62       38.7         61.3       0.0
  21         1        65       52.3         47.7       0.0
  22         1        14      100.0          0.0       0.0
  23         1        15       60.0         40.0       0.0
  24         2        85       76.5          2.4      21.2
  25         2        56       82.1         17.9       0.0
  26         2        23      100.0          0.0       0.0
  27         2        20       95.0          5.0       0.0
```

You know this from lssrad



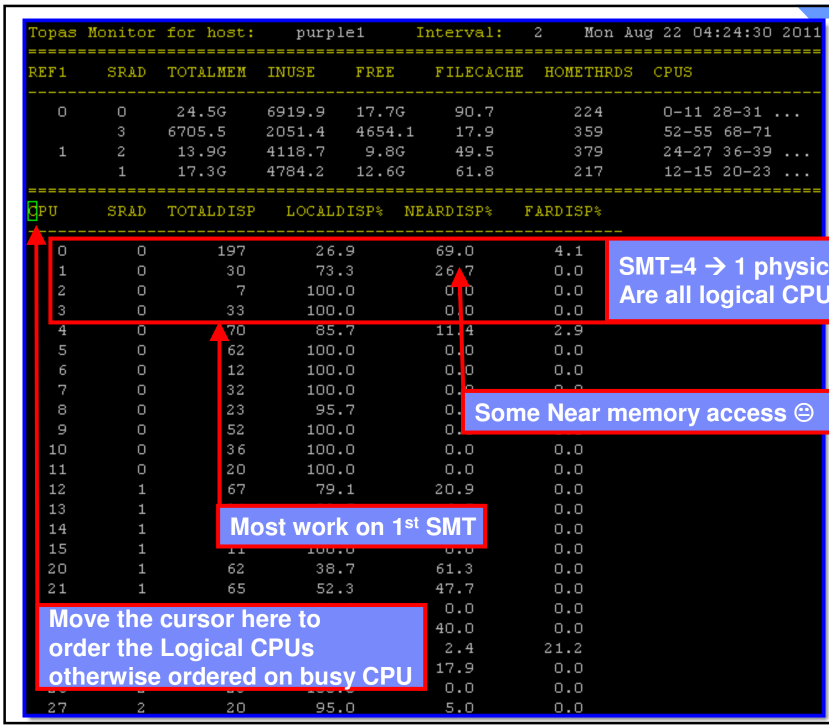
```
Topas Monitor for host:     purple1      Interval:   2    Mon Aug 22 04:24:30 2011
===================================================================================
REF1     SRAD  TOTALMEM  INUSE     FREE    FILECACHE  HOMETHRDS  CPUS
-----------------------------------------------------------------------------------
   0       0     24.5G   6919.9    17.7G     90.7        224      0-11 28-31 ...
           3    6705.5   2051.4   4654.1     17.9        359      52-55 68-71
   1       2     13.9G   4118.7     9.8G     49.5        379      24-27 36-39 ...
           1     17.3G   4784.2    12.6G     61.8        217      12-15 20-23 ...
===================================================================================
CPU      SRAD                     NEARDISP%   FARDISP%
----------------------------------------------------------
   0         0       197       26.9         69.0       4.1
   1         0        30       73.3         26.7       0.0
   2         0         7      100.0          0.0       0.0
   3         0        33      100.0          0.0       0.0
   4         0        70       85.7         11.4       2.9
   5         0        62      100.0          0.0       0.0
   6         0        12      100.0          0.0
   7         0        32      100.0          0.0
   8         0        23       95.7          0.0       4.3
   9         0        52      100.0          0.0       0.0
  10         0        36      100.0          0.0       0.0
  11         0        20      100.0          0.0       0.0
  12         1        67       79.1         20.9       0.0
  13         1        31       93.5          6.5       0.0
  14         1        18       77.8         22.2       0.0
  15         1        11      100.0          0.0       0.0
  20         1        62       38.7         61.3       0.0
  21         1        65       52.3         47.7       0.0
  22         1        14      100.0          0.0       0.0
  23         1        15       60.0         40.0       0.0
  24         2        85       76.5          2.4      21.2
  25         2        56       82.1         17.9       0.0
  26         2        23      100.0          0.0       0.0
  27         2        20       95.0          5.0       0.0
```

SRAD Memory View

CPUs but limited room

Home processes/threads

File system cache MB (numperm)



```
Topas Monitor for host:     purple1      Interval:    2     Mon Aug 22 04:24:30 2011
==================================================================================
REF1     SRAD  TOTALMEM  INUSE     FREE     FILECACHE  HOMETHRDS  CPUS
----------------------------------------------------------------------------------
   0       0     24.5G   6919.9   17.7G       90.7        224       0-11 28-31 ...
           3    6705.5   2051.4   4654.1      17.9        359       52-55 68-71
   1       2     13.9G   4118.7    9.8G       49.5        379       24-27 36-39 ...
           1     17.3G   4784.2   12.6G       61.8        217       12-15 20-23 ...
==================================================================================
CPU      SRAD  TOTALDISP   LOCALDISP%  NEARDISP%   FARDISP%
----------------------------------------------------------------
   0       0        197        26.9        69.0        4.1
   1       0         30        73.3        26.7        0.0
   2       0          7       100.0         0.0        0.0
   3       0         33       100.0         0.0        0.0
   4       0         70        85.7        11.4        2.9
   5       0         62       100.0         0.0        0.0
   6       0         12       100.0         0.0        0.0
   7       0         32       100.0         0.0        0.0
   8       0         23        95.7         0.
   9       0         52       100.0         0.
  10       0         36       100.0         0.0        0.0
  11       0         20       100.0         0.0        0.0
  12       1         67        79.1        20.9        0.0
  13       1                                           0.0
  14       1                                           0.0
  15       1         11       100.0         0.0        0.0
  20       1         62        38.7        61.3        0.0
  21       1         65        52.3        47.7        0.0
                                            0.0        0.0
                                           40.0        0.0
                                            2.4       21.2
                                           17.9        0.0
                                            0.0        0.0
  27       2         20        95.0         5.0        0.0
```

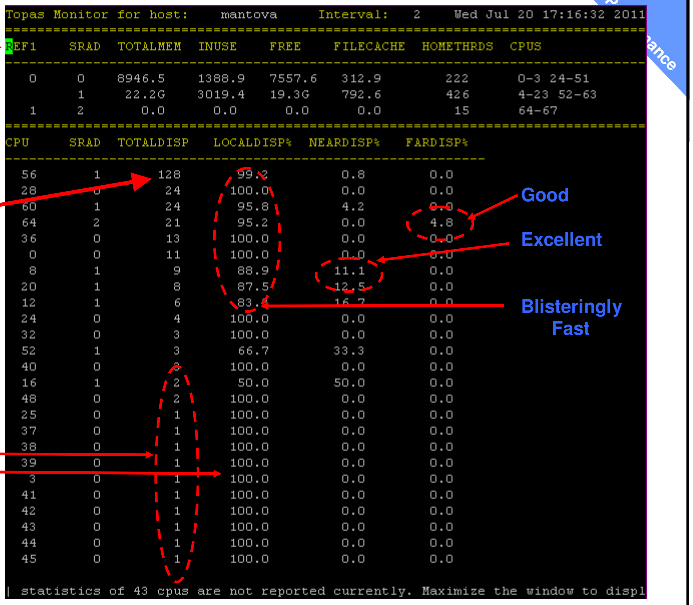
SMT=4 → 1 physical core
Are all logical CPUs in use?

Some Near memory access 😐

Most work on 1st SMT

Move the cursor here to order the Logical CPUs otherwise ordered on busy CPU

In CPU Busy Order



```
Topas Monitor for host:     mantova      Interval:    2     Wed Jul 20 17:16:32 2011
==================================================================================
REF1     SRAD  TOTALMEM  INUSE     FREE     FILECACHE  HOMETHRDS  CPUS
----------------------------------------------------------------------------------
   0       0    8946.5   1388.9   7557.6    312.9        222      0-3 24-51
           1     22.2G   3019.4   19.3G     792.6        426      4-23 52-63
   1       2      0.0      0.0      0.0       0.0         15      64-67
==================================================================================
CPU      SRAD  TOTALDISP   LOCALDISP%  NEARDISP%   FARDISP%
----------------------------------------------------------------
  56       1        128        99.2         0.8        0.0
  28       0         24       100.0         0.0        0.0
  60       1         24        95.8         4.2        0.0
  64       2         21        95.2         0.0        4.8
  36       0         13       100.0         0.0        0.0
   0       0         11       100.0         0.0        0.0
   8       1          9        88.9        11.1        0.0
  20       1          8        87.5        12.5        0.0
  12       1          6        83.3        16.7        0.0
  24       0          4       100.0         0.0        0.0
  32       0          3       100.0         0.0        0.0
  52       1          3        66.7        33.3        0.0
  40       0          2       100.0         0.0        0.0
  16       1          2        50.0        50.0        0.0
  48       0          2       100.0         0.0        0.0
  25       0          1       100.0         0.0        0.0
  37       0          1       100.0         0.0        0.0
  38       0          1       100.0         0.0        0.0
  39       0          1       100.0         0.0        0.0
   3       0          1       100.0         0.0        0.0
  41       0          1       100.0         0.0        0.0
  42       0          1       100.0         0.0        0.0
  43       0          1       100.0         0.0        0.0
  44       0          1       100.0         0.0        0.0
  45       0          1       100.0         0.0        0.0

| statistics of 43 cpus are not reported currently. Maximize the window to displ
```

Good

Excellent

Blisteringly Fast

Very little use so 100% not that important

svmon – just for reference

- Report global affinity domains more detail than lssrad
  - svmon -G -O affinity=on,unit=MB
- Report memory statistics
  - svmon -P [PID] -O threadaffinity=on and -O affinity=detail

# svmon -G -O unit=auto,timestamp=on,pgsz=on,affinity=detail

```
# svmon  -G -O unit=auto,timestamp=on,pgsz=on,affinity=detail
Unit: auto                                                    Timestamp: 21:18:40
---------------------------------------------------------------------------------------
               size       inuse        free         pin     virtual   available   mmode
memory        64.0G       4.57G       59.4G       3.28G       4.53G       59.4G     Ded
pg space    512.00M     203.60M

               work        pers        clnt       other
pin           1.42G          0K          0K       1.86G
in use        4.37G          0K     204.76M

PageSize   PoolSize       inuse        pgsp         pin     virtual
s    4 KB         -       2.75G     203.60M       2.45G       2.71G
    Domain affinity        used
                  3      136447
                  0       44655
                  2       21703
                  1       31757
m   64 KB         -       1.82G          0K     847.12M       1.82G
    Domain affinity        used
                  3       13280
                  0       16208
                  2       11328
                  1       13136

Domain affinity        free       used      total      lcpus
              0       23.4G      1.06G      24.5G       78.9M     0  1  2  3  4  5  6  7  8  9
   10 11 28 29 30 31 40 41 42 43 56 57 58 59 72 73 74 75
              1       16.5G    792.95M      17.3G       53.1M    12 13 14 15 20 21 22 23 32 33
   34 35 44 45 46 47 60 61 62 63 76 77 78 79 104 105 106 107
              2       13.3G    574.69M      13.9G       42.7M    24 25 26 27 36 37 38 39 48 49
   50 51 64 65 66 67
              3       6.24G    319.82M      6.55G       14.4M    52 53 54 55 68 69 70 71
```

Missing! File Cache heading on my beta AIX version

BLOG 3

Process Thread Scheduling to SMT Threads

Scheduling to SMT

- POWER5 & 6
  - SMT=2 and modes on or off

- POWER7
  - Up to SMT=4 and modes 1, 2, 4
  - And intelligent threads – auto switching mode
  - 1 runnable thread → SMT=1 mode
  - 2 runnable threads → SMT=2 mode
  - 3 or 4 runnable threads → SMT=4 mode

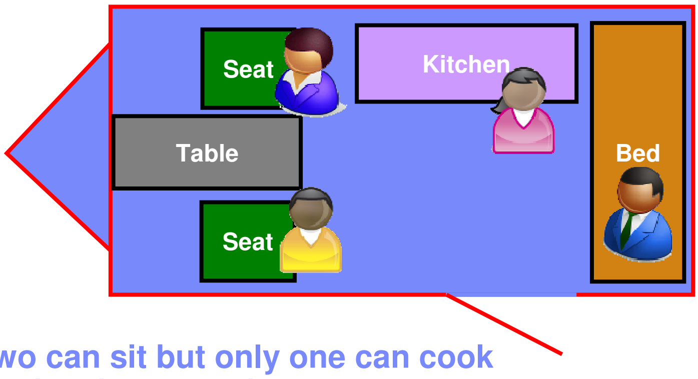
- A physical processor core is running one LPAR at a time
  - So all four SMT's in one LPAR at a time
  - AIX schedules work on these are four logical processors via run queues

Power7 12 execution units

- 2 integer units
- 2 load-store units
- 4 double-precision floating-point units
- 1 branch unit
- 1 condition register unit
- 1 vector unit
- 1 decimal floating-point unit

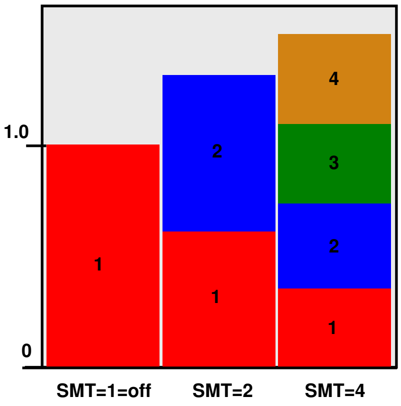
Four people in a small caravan

Works fine provided they don't all want to do the same thing at the same time



Two can sit but only one can cook
So they have to take turns

Simultaneous Multi-Threading (SMT)



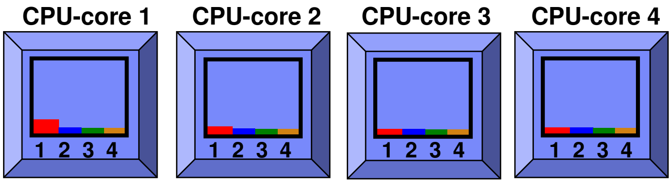
Relative Performance by SMT mode

- Over all more gets done
- Individual threads go a bit slower
- Good for throughput of many transactions
- Response time a little longer than SMT=1

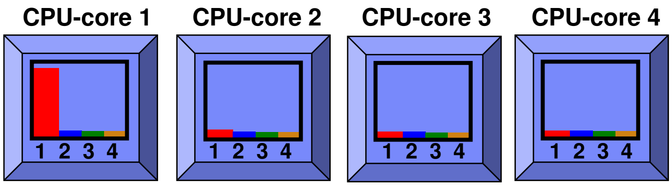
Example

- Virtual machine (LPAR)
- Entitlement of 1.5 (Uncapped)
- Virtual processor count of 4
- AIX is set to SMT=4 on our POWER7 machine
- So each CPU-cores has four SMT threads
  = 16 logical CPUs

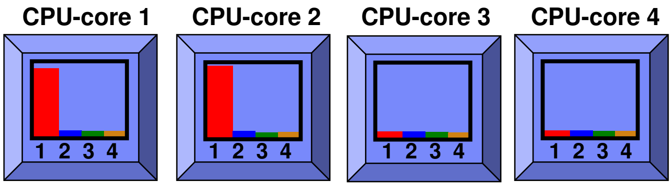
Tested on Firmware 730 & AIX 7.1 TL1



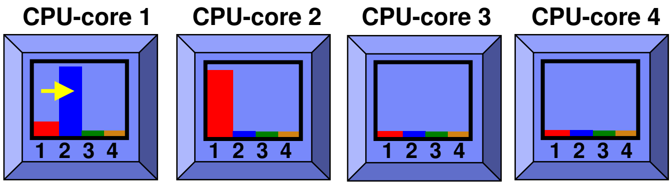
One busy (running 100% of the time) program

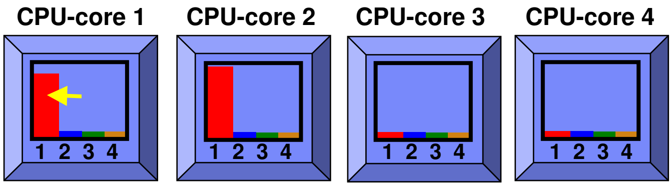


Where does the 2nd program go?



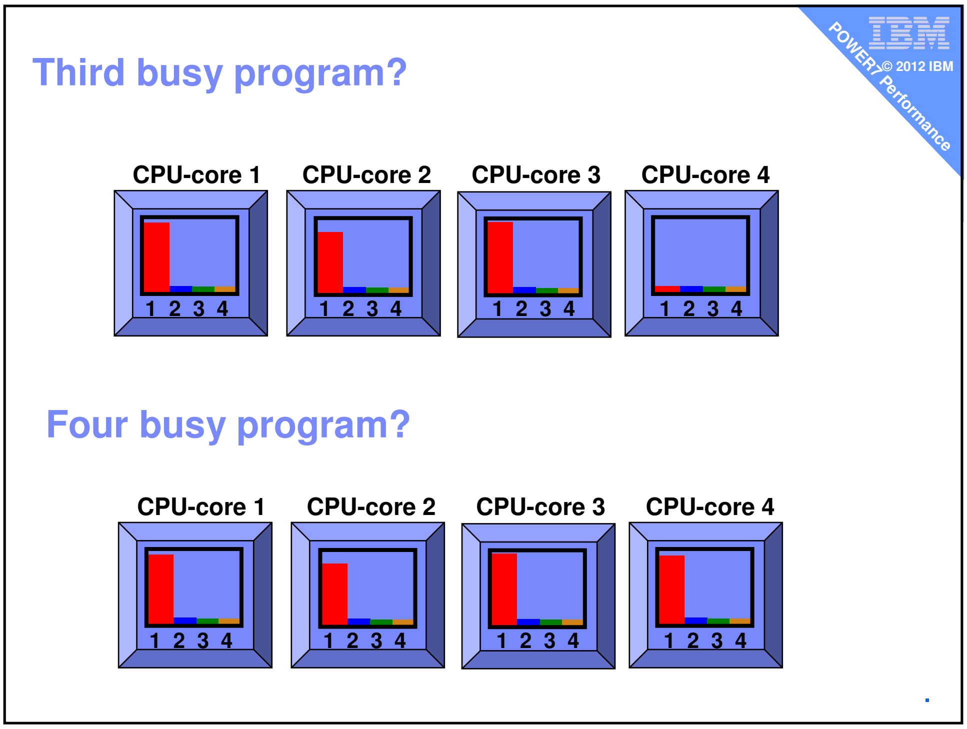
Some times you see SMT2 used but then return to SMT1





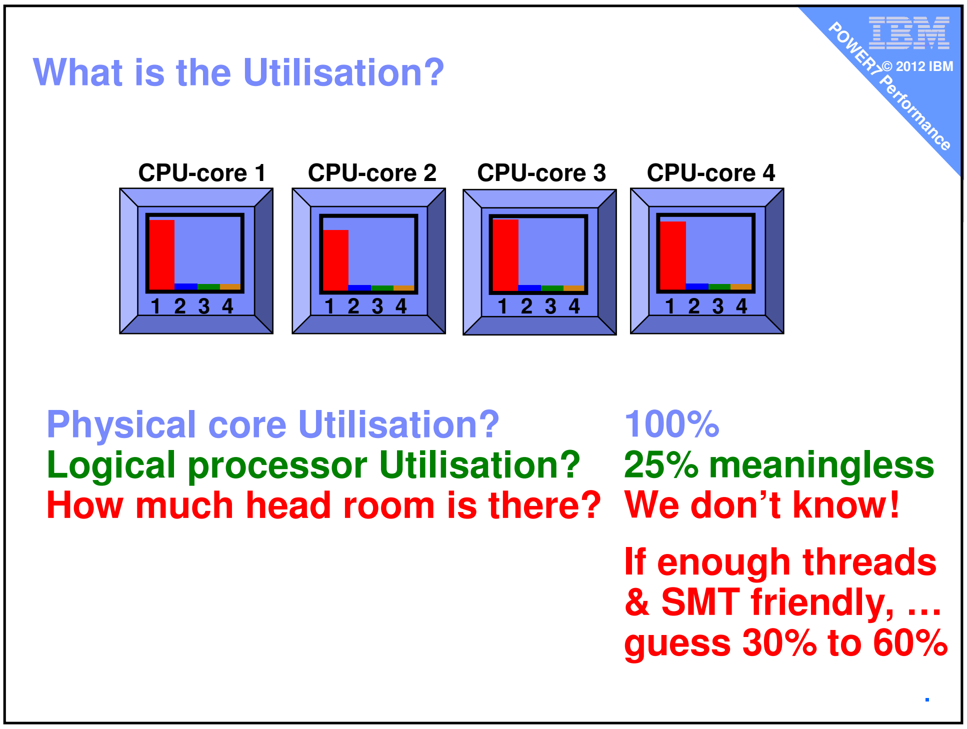
POWER7 is proactive
moving to SMT1

Jitter



Third busy program?

Four busy program?



What is the Utilisation?

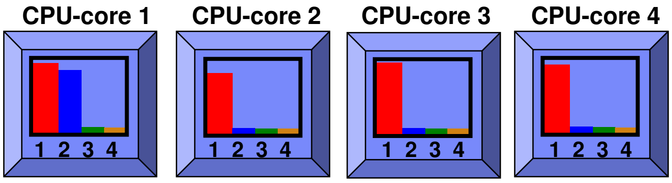
Physical core Utilisation? 100%

Logical processor Utilisation? 25% meaningless

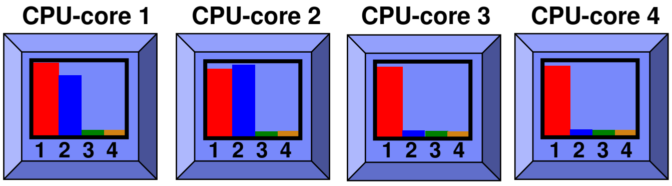
How much head room is there? We don’t know!

If enough threads
& SMT friendly, …
guess 30% to 60%

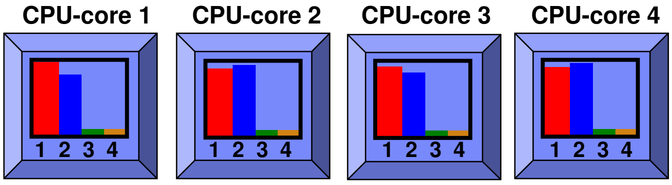
One more busy program – where?



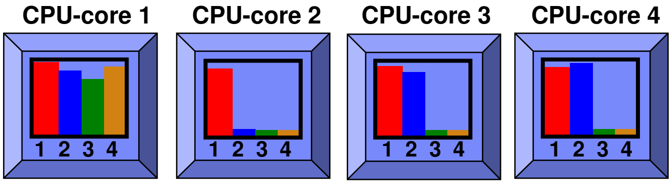
Sixth busy program – where?



Eight busy programs



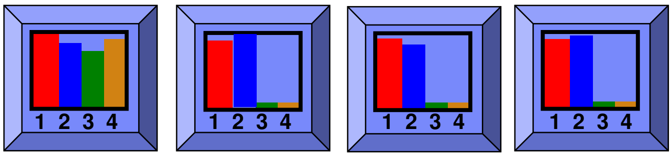
Ninth busy program – where?



No SMT=3, 1st core goes SMT4 runs four threads,
so 2nd core goes SMT=1 Note: now we have SMT=4 + SMT=1 and SMT=2

Ten busy programs

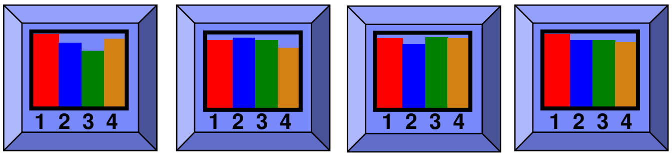
CPU-core 1 CPU-core 2 CPU-core 3 CPU-core 4



CPU-core 1 CPU-core 2 CPU-core 3 CPU-core 4

Eventually 16 programs

CPU-core 1 CPU-core 2 CPU-core 3 CPU-core 4



Where does the 17th go?

Classic time sharing on the logical CPUs

Advanced points in using POWER7 SMT=4

1. With 4 programs, 8 programs or 16 programs ALL physical cores were 100% busy

2. If you don't have enough runnable processes (run queue), you can't use SMT=4 and you don't get the full POWER7 rPerf

3. Not enough processes (or process threads)
   - Tune app or middleware to use more
   - Reduce your VP count! NOT the Entitlement
   - Get the users to work faster!!

HOWEVER

- Above was for **Spinning** processes = 100% busy

- Typical workloads 100's of processes taking factions of a second, so harder to determine if it needs more CPU resources

- AIX uses % thresholds to determine when to switch:
  - On more cores or
  - On more SMT threads

- These % were subject to fixes for AIX6 TL5 +TL6

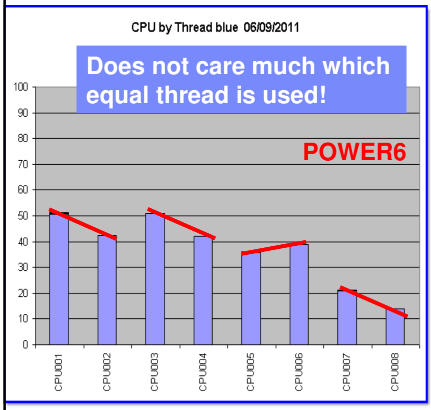
BLOG 4

POWER7 Aggressive Intelligent Threads

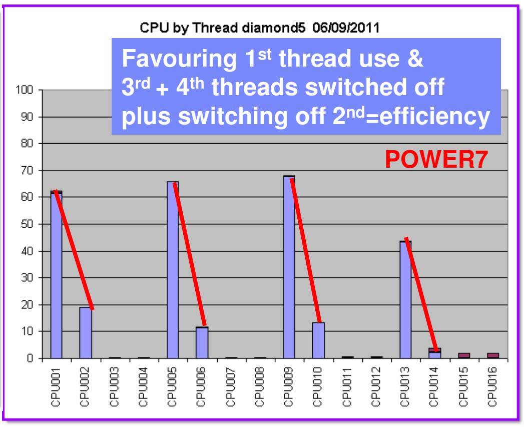
Comparing POWER6 and 7

Regular question: Is POWER7 broken?

– “fake” 8 program steady workload



Does not care much which equal thread is used!



Favouring 1st thread use & 3rd + 4th threads switched off plus switching off 2nd=efficiency

We see POWER Intelligent threads working & POWER7 is working very well

POWER7 Aggressive Intelligent Threads

- POWER6 in SMT=2 not bothered which thread

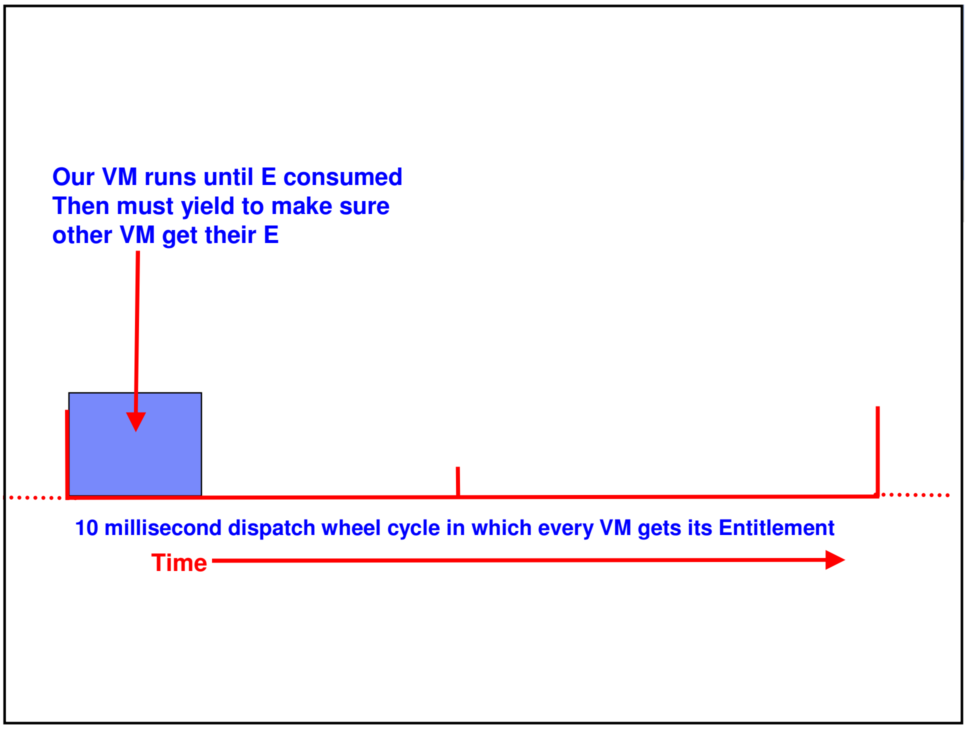
- POWER7 moves process threads to SMT1 (or 2) proactively and switches to SMT=1 (or 2) mode for higher efficiency

- In other stats POWER6 & 7 both using ~2.5 CPUs but on POWER7 its obvious we can remove a CPU

- You CAN'T find the 2.5 on the previous graphs as you can't average logical CPUs intermixing on the physical CPU. (70+10+0+0)/4=20%

BLOG 5

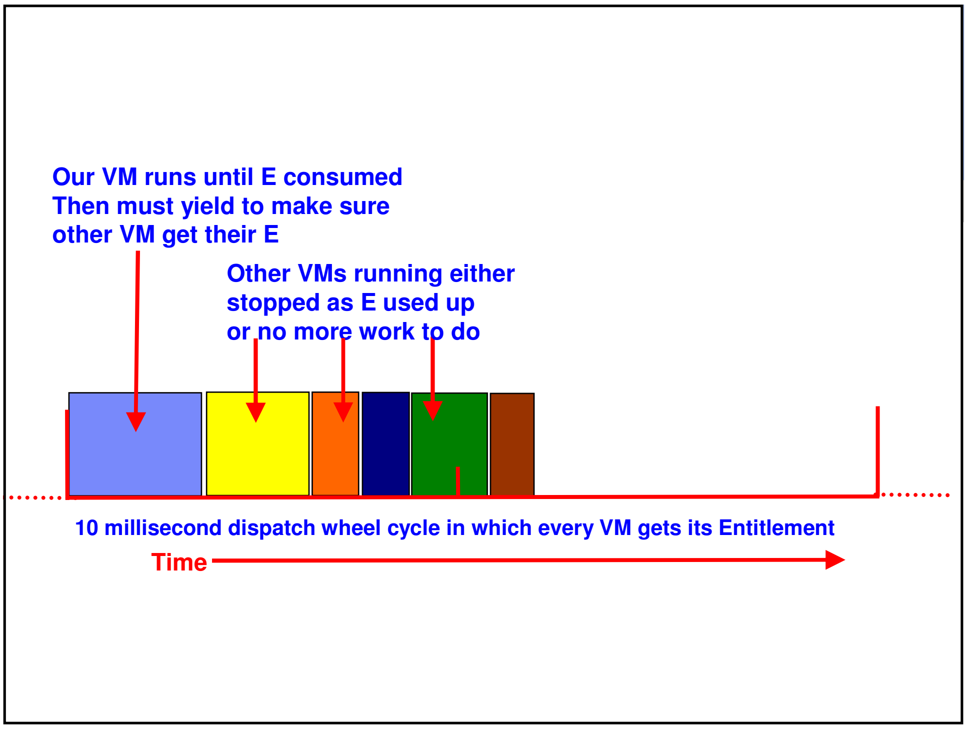
Low Entitlement →Bad Side Effect



Our VM runs until E consumed
Then must yield to make sure
other VM get their E

10 millisecond dispatch wheel cycle in which every VM gets its Entitlement

Time

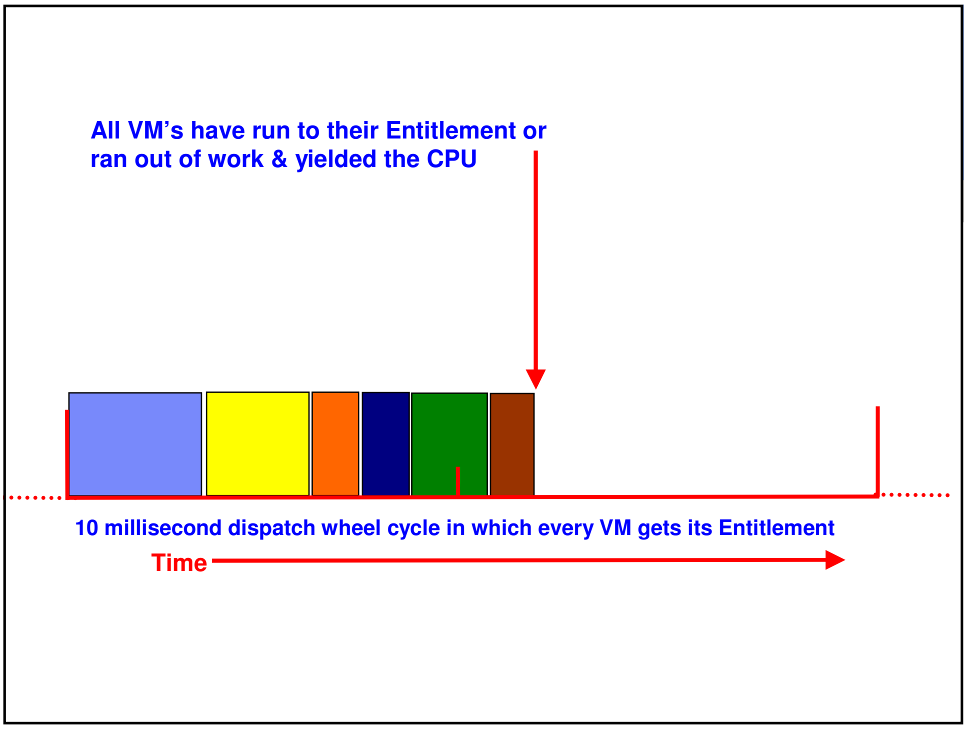


Our VM runs until E consumed
Then must yield to make sure
other VM get their E

Other VMs running either
stopped as E used up
or no more work to do

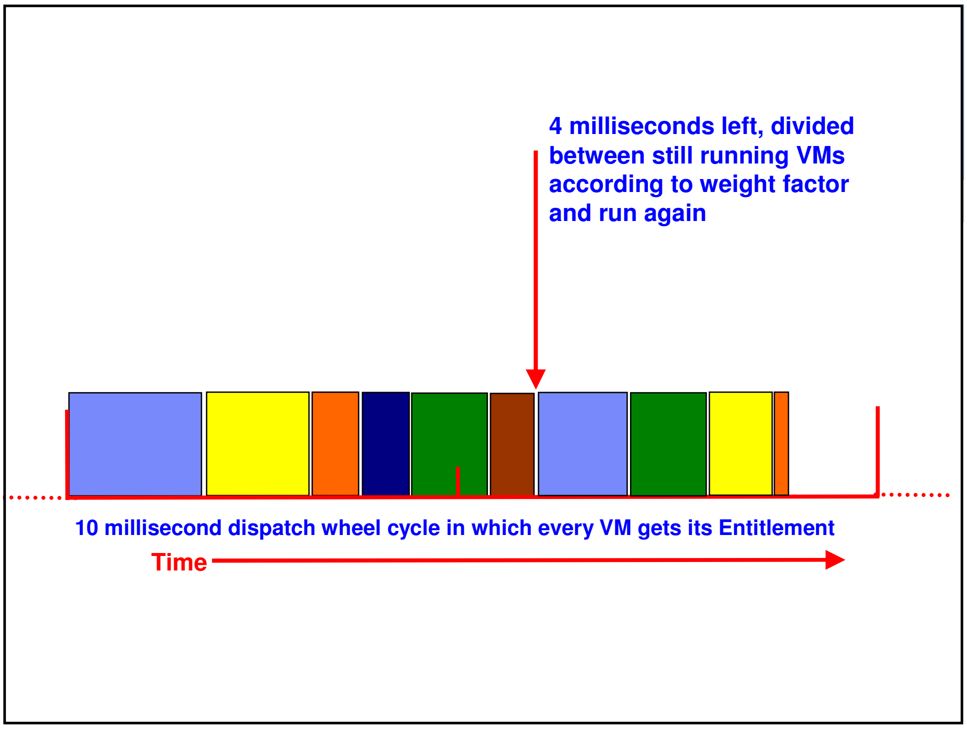
10 millisecond dispatch wheel cycle in which every VM gets its Entitlement

Time



All VM’s have run to their Entitlement or
ran out of work & yielded the CPU

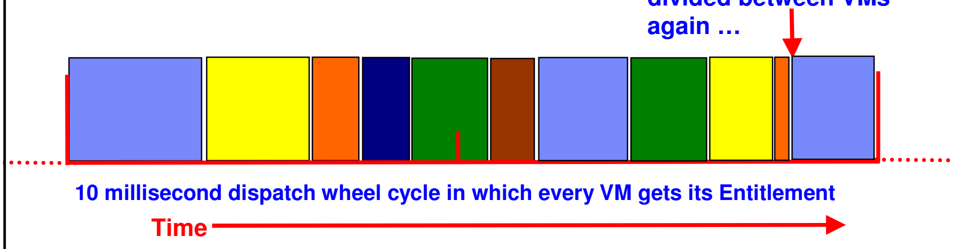
10 millisecond dispatch wheel cycle in which every VM gets its Entitlement



4 milliseconds left, divided
between still running VMs
according to weight factor
and run again

10 millisecond dispatch wheel cycle in which every VM gets its Entitlement

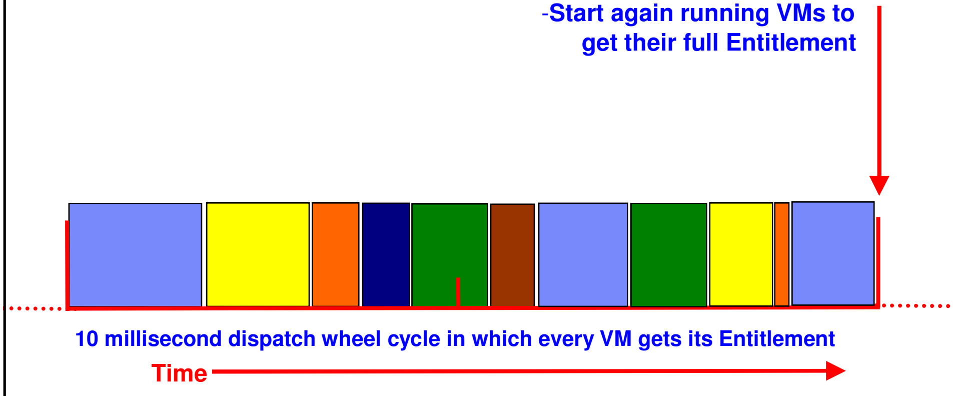
1 milliseconds left divided between VMs again …



10 millisecond dispatch wheel cycle in which every VM gets its Entitlement

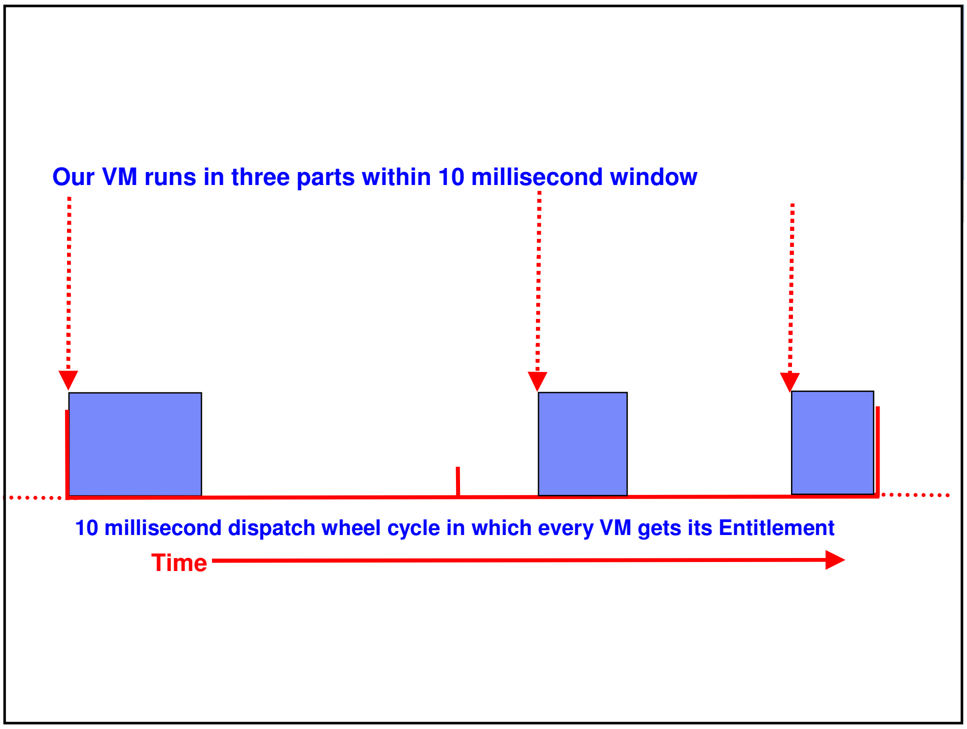
Time

Dispatch wheel cycle complete
-Start again running VMs to
get their full Entitlement



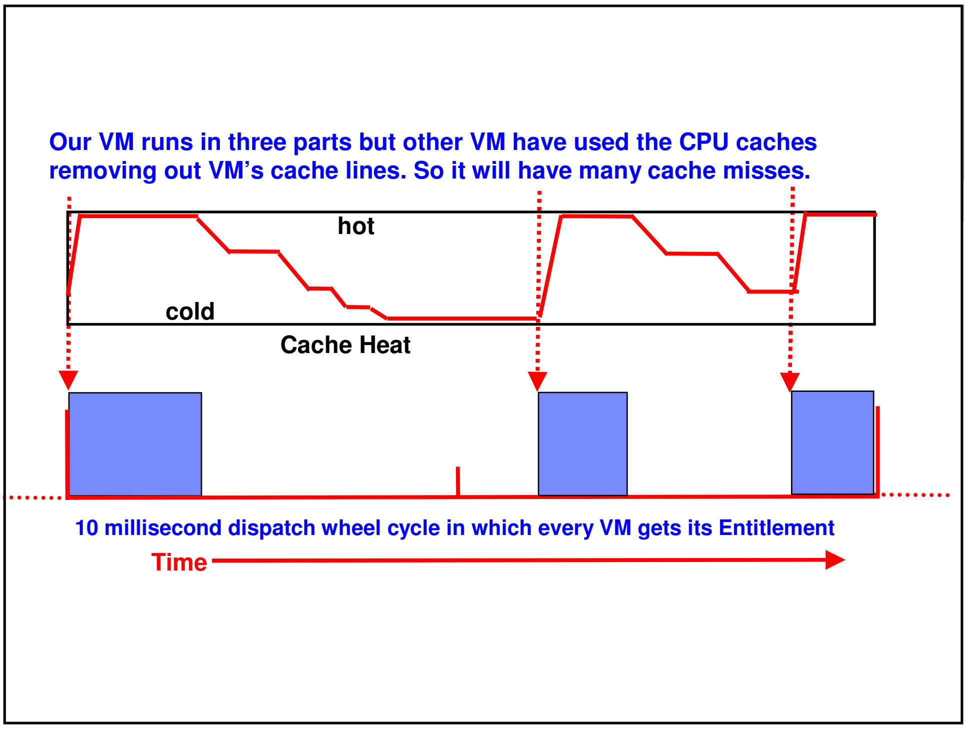
10 millisecond dispatch wheel cycle in which every VM gets its Entitlement

Time



Our VM runs in three parts within 10 millisecond window

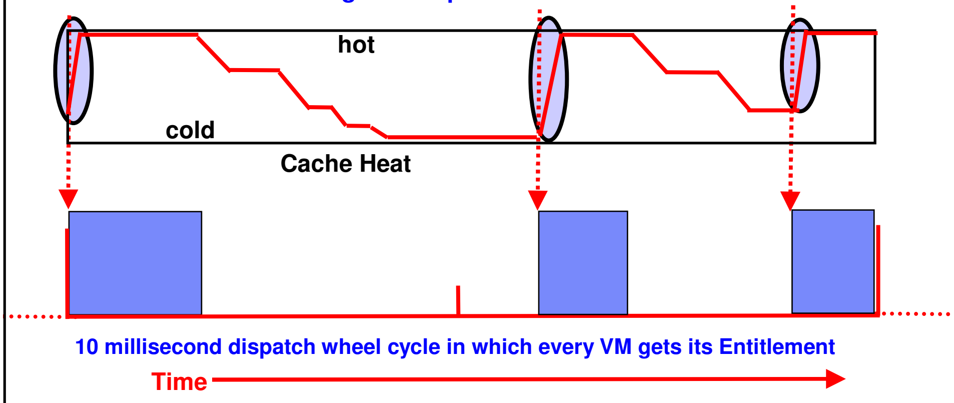
10 millisecond dispatch wheel cycle in which every VM gets its Entitlement



Our VM runs in three parts but other VM have used the CPU caches removing out VM’s cache lines. So it will have many cache misses.

10 millisecond dispatch wheel cycle in which every VM gets its Entitlement

As our VM runs it needs to warm up the CPU caches each time
– this means it not running at full speed for a while.



hot

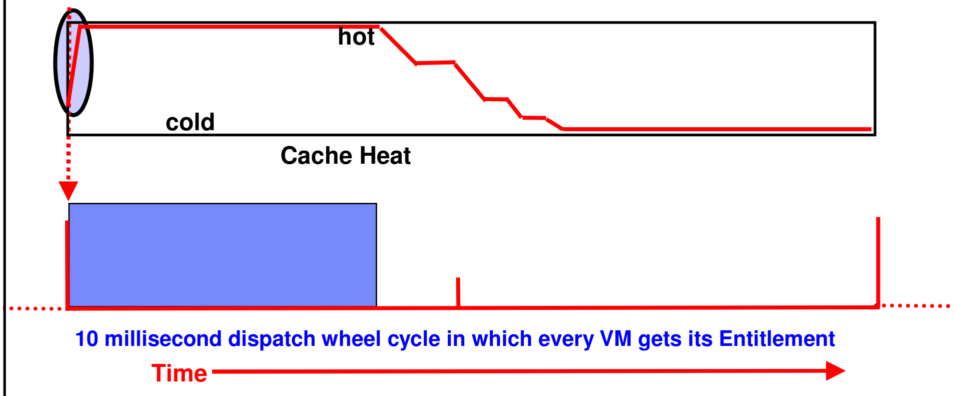
cold

Cache Heat

10 millisecond dispatch wheel cycle in which every VM gets its Entitlement

Time

Higher Entitlement lets It run continuously with less cache warm up time



hot

cold

Cache Heat

10 millisecond dispatch wheel cycle in which every VM gets its Entitlement

Time

Lessons

- Don't forget 10 milliseconds is a long time on a CPU
  - At 4 GHz = 400,000,000 instructions (assuming 1 op/cycle)
  - Illustration was grossly simplified by factor 10 or more
- Get the Entitlement "about right“

Next Week

POWER7 Affinity and Performance Part 2 “same time, same channel”

- Guru of the Month & YouTube
- Ten Top Techie Treats – information sources
- My Redbook Library
- Getting help from guru level tools
  - The Optimisers
  - The Advisors
- More Advanced Level and New stuff
  - Physical VM placement
  - VM “defrag”
  - Working out “space capacity”
  - Getting POWER7 to look more like POWER5/6

Nigel Griffiths
IBM Power Systems
Advanced Technology Support, Europe