z16 Capacity Planning
(Part 2)

Fabio Massimo Ottaviani – EPV Technologies

May 2022

7 Processor cache architecture

From the logical point of view, z15 and z16 have very similar architectures.
If the data and instructions to be processed are found in the Level 1 cache (L1) dedicated to each processor, this is called a "cache hit".
In this case the processor clock speed can be exploited well.

If data and instructions cannot be found in L1 then the hardware tries to load them, in this order:

- from the Level 2 (L2) cache, which is still a cache dedicated to each processor[1],
- from the Level 3 (L3) cache, which is a cache serving all the processors of the same chip,
- from the Level 4 cache (L4), which is a cache serving all the processors of the same drawer,
- from the L4 cache of a remote drawer,
- from local memory (of the same drawer),
- from remote memory (of another drawer).

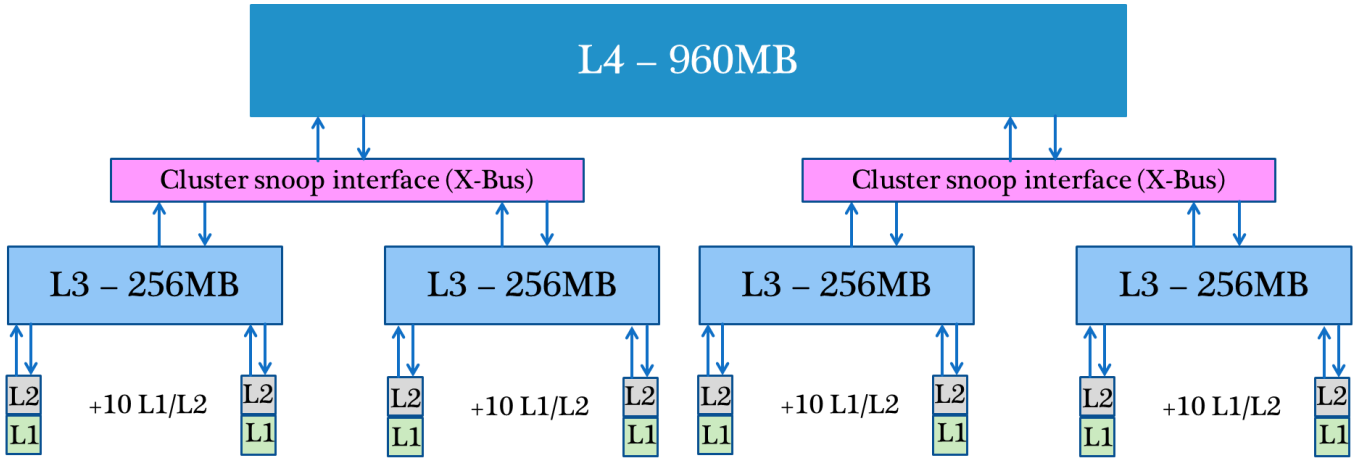
This is a "cache miss" and clock cycles are lost while waiting for data and instructions to be loaded into the L1 cache.
The number of lost cycles depends on the cache level accessed: it can range from a few cycles when the miss is resolved in L2 to hundreds of cycles if an access to remote memory is needed.

On the contrary, from a physical point of view there are many significant differences between z15 and z16:

- z16 has a maximum of 4 drawers instead of 5 drawers as in z15;
- in each drawer there are four dual chip modules (DCM) instead of 2 clusters as in z15;
- in each chip there are 8 cores instead of 12 as in z15;
- the L2 cache size is much bigger in z16 (32MB) than in z15;
- the L2 cache includes both instruction and data in z16, while separate L2 caches for data and instructions (4MB each) are provided in z15;
- the L2 cache is semi-private in z16; it means that the private part of the cache is dedicated to the processor and the part that is not private is used to virtualize the L3 (VL3) and L4 (VL4) caches[2];
- L3 and L4 are virtualized in z16; physical L3 and L4 caches are instead provided in z15.

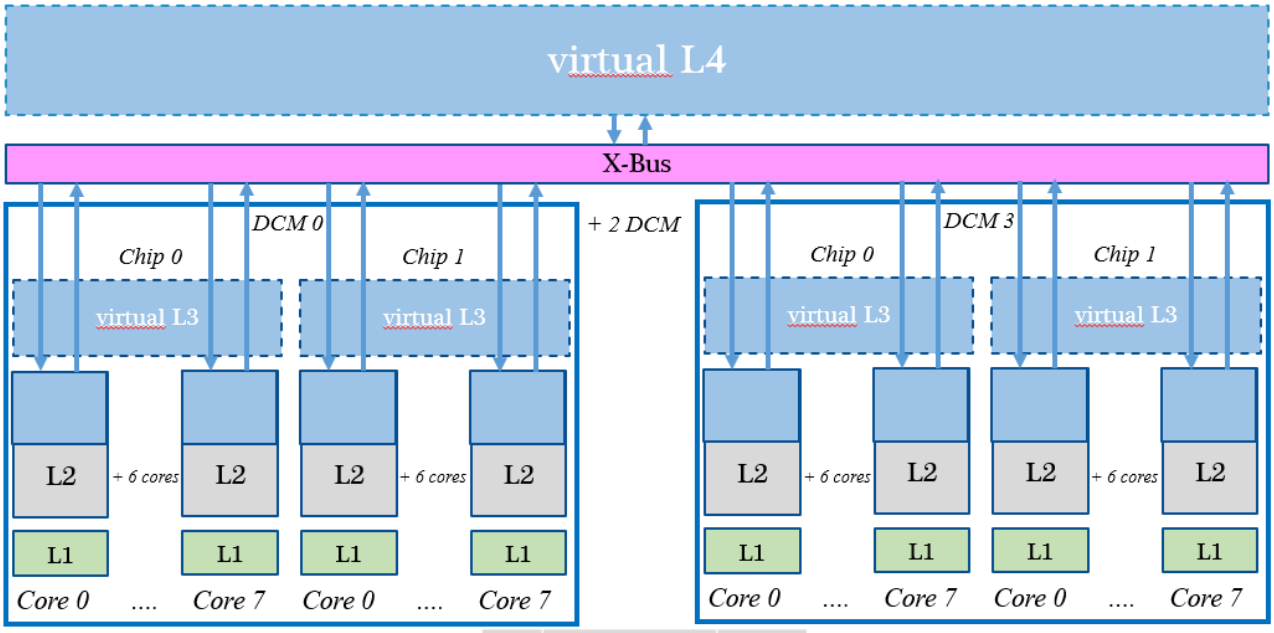
The schemas in Figure 6 and 7 shows some of the differences between the z15 and z16 processor cache architectures.

[1] Only part of it in z16.
[2] 16MB are dedicated to the associated core, and 16MB are used to virtualize L3 and L4 caches (the 50/50 split is adjustable).



L2	Instructions	4MB
	Data	4MB
L1	Instructions	128K
	Data	128K

Figure 6



L2		32MB
L1	Instructions	128K
	Data	128K

Figure 7

In IBM documentation you will find that the VL3 cache can use up to 256MB per chip and the VL4 cache up to 2048MB per drawer.
Honestly, this statement seems a bit misleading. In theory it is true but in practice it looks impossible even if you consider that also the L2 cache of not characterized cores is used for VL3 and VL4.[3]

It is also important to consider that in production systems running on z15 the L1 miss are normally 3% or lower. Of this 3%, about 70% are in-sourced from L2.
With this architecture IBM is trying to significantly improve the hit in the L2 cache. Anyway, the performance of z16, for what concerns the virtualized L3 and L4 caches, needs to be carefully evaluated.[4]

The bottom line is always the same: *"Workload capacity performance will be quite sensitive to how deep into the memory hierarchy the processor must go to retrieve the workload's instructions and data for execution. Best performance occurs when the instructions and data are found in the cache(s) nearest the processor so that little time is spent waiting prior to execution; as instructions and data must be retrieved from farther out in the hierarchy, the processor spends more time waiting for their arrival.[5]"*

As in z15, the two main factors determining workload performance in z16 are:

- Percentage of L1 misses over total searches;
- Percentage of L1 misses satisfied by each cache level (including memory).

In the next chapter we will show how to calculate them for z16 machines.

[3] The number of cores in a z16 drawer is 64 but only part of them can be characterized and, of course, customers may need only part of the characterizable cores.
[4] Many other improvements have been introduced in z16. They are described in the "IBM z16 (3931) Technical Guide" manual.
[5] From IBM Large Systems Performance Reference

9 SMF 113 counters

The CPU Measurement Facility (CPU MF), introduced with z10 machines, provides the ability to obtain measurements (counters) on processor cache effectiveness.
Collected information are recorded in SMF 113 subtype 1.[6]

The most important groups of counters are:

- basic counters; which should be used to calculate the percentage of L1 misses;
- extended counters; which should be used to calculate the percentage of L1 misses sourced by each cache level and, starting from them, the RNI value.

BASIC COUNTERS

Six metrics are provided in the Basic Counters section:

- B0, CYCLE COUNT
- B1, INSTRUCTION COUNT
- B2, L1 I-CACHE DIRECTORY-WRITE COUNT
- B3, L1 I-CACHE PENALTY CYCLE COUNT
- B4, L1 D-CACHE DIRECTORY-WRITE COUNT
- B5, L1 D-CACHE PENALTY CYCLE COUNT

The basic counters' meaning is the same whatever the machine model is (z10, z196, z114, zBC12, zEC12, z13, z13s, z14, z15 and z16).

Starting from these measurements the percentage of L1 misses over total searches can be calculated for z16 by using the usual formula (valid for all the models):

$$\%L1\ Miss = ((B2 + B4) / B1) * 100$$

z16 EXTENDED COUNTERS

The extended counters' meaning normally depends on the machine model. The SMF113_1_CtrVersion2 field allows identification of the model: it is 1 for z10, 2 for z196 and z114, 3 for zEC12 and zBC12, 4 for z13 and z13s, 5 for z14, 6 for z15 and 7 for z16.

Due to the new processor cache architecture, many more extended counters are provided and the meaning of almost all of them has changed.

Based on information available at the time of this writing, the number of L1 misses sourced by each cache levels should be calculated for z16 as follows:[7]

- L2d, data sourced from L2 = E145 + E146;
- L2i, instructions sourced from L2 = E169 + E170;
- L3d, data sourced from VL3 = E147 + E149 + E150 + E151;
- L3i, instructions sourced from VL3 = E171 + E173 + E174 + E175;
- L4Ld, data sourced from VL4 Local = E148 + E152 + E153 + E154;
- L4Li, instructions sourced from VL4 Local = E172 + E176 + E177 + E178;

[6] SMF 113 subtype 2 is frozen, and all the new information will be added to subtype 1.
[7] These calculations need to be considered as provisional. We will eventually publish an addendum to correct them.

- L4Lm, data or instructions sourced from VL4 Local = E160 + E161 + E162 + E163 + E164 + E165;
- L4Rd, data sourced from VL4 Remote = E155;
- L4Ri, instructions sourced from VL4 Remote = E177;
- L4Rm, data or instructions sourced from VL4 Remote = E166 + E167 + E168;
- MEMLd, data sourced from Local Memory = E156 + E157 + E158;
- MEMRd, data sourced from Remote Memory = E159;
- MEMLi, instructions sourced from Local Memory = E180 + E181 + E182;
- MEMRi, instructions sourced from Remote Memory = E183.

Starting from these measurements, the percentage of L1 misses sourced by each cache level can be calculated by using the following formulas[8]:

- %L2 = (L2d + L2i) / (B2 + B4) * 100
- %L3 = (L3d + L3i) / (B2 + B4) * 100
- %L4L = (L4Ld + L4Li + L4Lm) / (B2 + B4) * 100
- %L4R = (L4Rd + L4Li + L4Rm) / (B2 + B4) * 100
- %MEM = (MEMLd + MEMLi + MEMRd + MEMRd) / (B2 + B4) * 100

The Relative Nest Intensity (RNI) of a workload is a measure of the activity to the nest. The nest is composed by shared caches (L3 and L4) and memory; it is the most performance sensitive area of the memory hierarchy.

The following formula[9] allows you to calculate the RNI of a workload running on a z16 machine, starting from the amount of activity to the nest:

IBM z16 RNI=4,3 x (0,45 x %L3 + 1,3 x %L4L + 5,0 x %L4R + 6,1 x %MEM) / 100

where:

- %L3 is the percentage of L1 misses sourced from the shared chip-level VL3 cache,
- %L4L is the percentage of L1 misses sourced from the local drawer VL4 cache,
- %L4R is the percentage of L1 misses sourced from a remote drawer VL4 cache,
- %MEM is the percentage of L1 misses sourced from memory.

The coefficients (in bold) are used to weight cache and memory accesses; in the above formula:

- accessing the VL3 cache (%L3) is weighted 0,45;
- accessing the local drawer VL4 cache (%L4L) is weighted 1,3;
- accessing the remote drawer VL4 cache (%L4R) is weighted 5,0:
- accessing memory (%MEM), including both local and remote memory, is weighted 6,1;
- an additional coefficient (4,3) is used to adjust the resulting RNI value.

As already discussed, workload capacity performance is quite sensitive to how deep into the memory hierarchy the processor must go to retrieve the workload's instructions and data to be executed. The higher the %L1 Miss and RNI, the worse will be the workload capacity performance.

[8] More details in "The CPU-Measurement Facility Extended Counters Definition for z10, z196/z114, zEC12/zBC12, z13, z14, z15 and z16" manual (SA23-2261-07).
[9] Please note that this formula is not the same as the one used for z15.

By using the %L1 Miss and RNI values, together with the rules in the next figure, you can understand which benchmark best represents the workload running in each system.

%L1 Miss	RNI	Benchmark
< 3%	>= 0,75	AVG RNI
< 3%	< 0,75	LOW RNI
3% to 6%	> 1,00	HIGH RNI
3% to 6%	0,60 to 1,00	AVG RNI
3% to 6%	< 0,60	LOW RNI
> 6%	>= 0,75	HIGH RNI
> 6%	< 0,75	AVG RNI

Figure 9

In practical terms, the machine will look less powerful on a workload represented by a HIGH RNI benchmark than on a workload represented by an AVG or LOW RNI benchmark.

10 The CPI index

The CPI index represents the average number of cycles needed per instruction. It can be calculated by using basic counters and the following simple formula (valid for all the models including z16):

$$CPI = B0 / B1$$

As you can imagine there is not a Rule of Thumb for the ideal CPI value. However, it is intuitive that to exploit the processor power the CPI value should be as low as possible.

Measuring this index on a regular basis will allow you to evaluate the effect of changes in:

- hardware configuration;
- microcode;
- exploitation of HiperDispatch;
- LPAR configuration such as weights, number of logical processors, number of LPARs, etc.;
- system and subsystem levels;
- workload mixture.

You can use CPI to evaluate the performance benefits when moving to a new machine generation, but you must normalize the CPI values to the processor speed to make a meaningful comparison.[10]

$$\textit{normalized new machine CPI} = \frac{\textit{old machine cycle}}{\textit{new machine cycle}} * \textit{new machine CPI}$$

You can also get a deeper understanding of CPI by splitting it into:

- finite_CPI; cycles needed because L1 cache is not infinite; it indicates which portion of CPI is due to data and instructions coming from L2 and shared caches (Nest);

[10] It is not an issue if moving from z14 or z15 to z16 because the clock speed is the same.

- instruction_complexity_CPI; cycles needed even with an infinite L1 cache; it indicates which portion of CPI is due to the effectiveness of the microprocessor design with your workload.

They can be estimated for z16 by using the following simple formulas, which are the same used in z15 and z14:

$$finite\_CPI^{11} = E143 / B1$$

$$instruction\_complexity\_CPI = CPI - intercept$$

11 Conclusions

z16 looks to be a very powerful machine with up to 200 CP and a maximum capacity of more than 215.000 MIPS.

Compared to z15, other benefits are that:

- single processor capacity has been increased by about 10%;
- capacity variability, due to workload characteristics, seems to be lower.

Many relevant changes have been introduced in the z16 processor cache architecture. The impression is that the z16 machine is the first step of a new development cycle.
For this reason, particular care should be used in capacity planning studies.

There are many more extended counters provided in SMF 113 for z16 and they have not the same meaning as for z15; the RNI formula, to be used to choose the right benchmark, is also specific for z16.

[11] The E143 extended counter provides the number of cycles where a level-1 cache or level-2 TLB miss is in progress.