z10 Capacity Planning Issues

Fabio Massimo Ottaviani – EPV Technologies

White paper

1 Introduction

IBM z10 machines present innovative architecture and features (HiperDispatch) designed to exploit the speed of the new 4.4 GHz processors. LSPR benchmarks have also been redesigned to estimate z10 capacity and to update the software configuration to the most current release.
These two elements introduce a number of Capacity Planning issues you have to be aware of before trying to estimate z10 capacity compared to old machines.
After a short introduction to the z10 design and to HiperDispatch, this paper discusses these issues. With the up to date information and advice provided you will be able to perform more reliable capacity planning studies.

2 Overview of the z10 design

One of the key elements of the z10 design is the usage of new 4.4 GHz processors (assembled in quad core chips) running much faster than the 1.7 GHz processors used previously in z9 machines.
In the z9 the processor speed can only be fully exploited when required data and instructions are found in the Level 1 cache associated with each processor. However when a Level 1 cache miss occurs, data and instructions have to be loaded from other cache levels (such as the Level 2 cache shared among all the physical processors packaged in a book) or from memory.
The number of cycles (latency) needed to get the data is proportional to the distance to travel and to the processor speed; so for the same distance a faster processor will actually lose more cycles than a slower one.

To reduce latency, the following changes have been introduced in the z10 design:

- a new microprocessor architecture based on three levels of cache (compared with the two used in z9 machines):
  - each microprocessor has its own Level 1 cache split into 128 KB for data and 64 KB for instructions; in z9 Level 1 cache is split into 256 KB for data and 256 KB for instructions[1];
  - a Level 1.5 cache, also dedicated to each microprocessor, which is 3 MB in size; this new intermediate cache level is not present in z9 machines and has been introduced to reduce the traffic to and from the Level 2 cache;
  - a 48 MB Level 2 cache versus 40 MB in z9;
- a different pipeline design;
- a point-to-point connection topology allowing every book to communicate with every other book instead of the ring topology employed on z9 machines.

[1] It's worth noting that the Level 1 cache is smaller in z10 than in z9.

3 HiperDispatch

Another innovative technology introduced by IBM to exploit the increased processor speed of z10 machine is a new feature called HiperDispatch.

z10 EC models scale up to 64 very powerful processors allowing the concentration of a much higher number of LPARs than before on a single machine. A side effect of this high number of LPARs is the likely increase in the number of defined logical processors compared to the number of physical CPs.

These factors tend to reduce the probability for a logical processor to be re-dispatched to the same physical processor and therefore reuse instructions and data previously loaded in the Level 1 cache thus increasing latency.

HiperDispatch has been designed to minimise the number of Level 1 (and Level 1.5) cache misses in z10 machines and, if a cache miss occurs, to maximise the probability of finding instructions and data in the Level 2 cache of the book where the logical processor is dispatched.

With HiperDispatch, z/OS and PR/SM communicate and work together, through a new metric called polarization weight, in order to re-dispatch a unit of work to the same physical processor or at least in the same group of physical processors previously used.

4 Capacity planning issues

The new design and the availability of HiperDispatch make the estimatation of z10 capacity a challenging task. To estimate correctly you have to download and use zPCR V5.1e and also to be aware of the following important issues:

a) **All z10 LSPR benchmarks have been performed with HiperDispatch turned on. This is the reason why the latest zPCR version is not able to estimate z10 capacity unless HiperDispatch is active.**

However customers could decide not to activate HiperDispatch on some or all the LPARs of a machine[2]. HiperDispatch mode has to be explicitly enabled by specifying HIPERDISPATCH=YES, in the IEAOPTxx member of SYS1.PARMLIB. The default is HIPERDISPATCH=NO which maintains the existing mode of dispatching.
If for whatever reason you run with HiperDispatch off, you will lose those benefits and you have to expect a lower capacity than reported in LSPR benchmarks.
According to IBM documentation the increase of capacity due to HiperDispatch activity is in the 0% – 10% range.

The benefits of activating HiperDispatch depend essentially on the following factors:

- size of LPARs; a configuration where the largest z/OS image fits within a book will see minimal improvement using HiperDispatch

[2] The most common reasons are missing software pre-requisites, open APARs which could impact system efficiency or simply prudent behavior.

- running workload; workloads which are CPU-intensive (batch applications) normally get a minimal improvement since they typically have long-running tasks that tend to stick on a logical engine anyway; workloads that tend to have common tasks and high dispatch rates (transactional applications) may see larger improvements;
- logical to physical CP ratios; LPAR configurations that have higher ratios may see bigger improvements especially if IRD or other automation techniques to reduce the number of online logical processors are not used.

b) **All the new LSPR benchmarks are based on z/OS 1.8[3] and more recent levels of subsystem and compiler software. These new benchmarks assign a lower capacity to z9 machines than the previous ones which were based on z/OS 1.6.**

The following graph shows the difference for the LoIO workload mix.

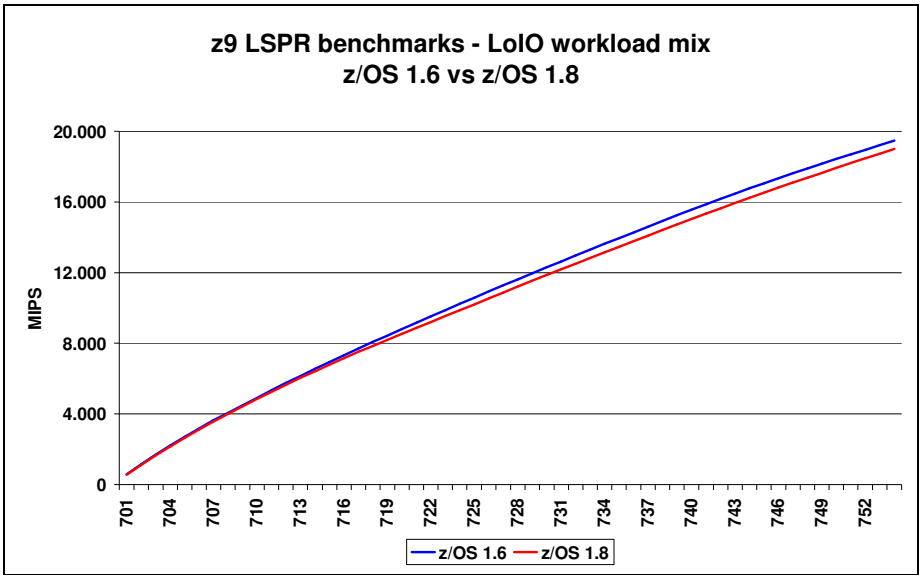

Figure 1

zPCR FAQs suggest use of a 2094-701 as a reference CPU.

The new suggested capacity scaling factor[4] (MIPS in IBM language) are 568,288 when using the LSPR Multi-Image Capacity Ratios table and 602 when using LPAR Configuration Capacity Planning[5].

So if you have to estimate the capacity delta when migrating from z9 (or older machines) to z10 you don't have to use old zPCR studies or MIPS table values but you do have to perform a new zPCR study for the old machine or consider new table values.

[3] On 21st October new LSPR benchmarks based on z/OS 1.9 have been released.
[4] Previous values based on z/OS 1.6 were 580 and 608. LSPR FAQs for z/OS 1.9 new benchmarks suggest to use 570,176 and 604.
[5] In this case you have to use the Single-Image value.

c) **Workload variability in z10 is much greater than in previous machine generations.**
The reasons are the new design and the HiperDispatch feature described in previous chapters which make a low task switching rates and cache hit ratios more important than in the past to exploit hardware power.

The following graph shows how the difference in capacity between LoIO and DI mixes has increased in z10 compared to z9 machines.

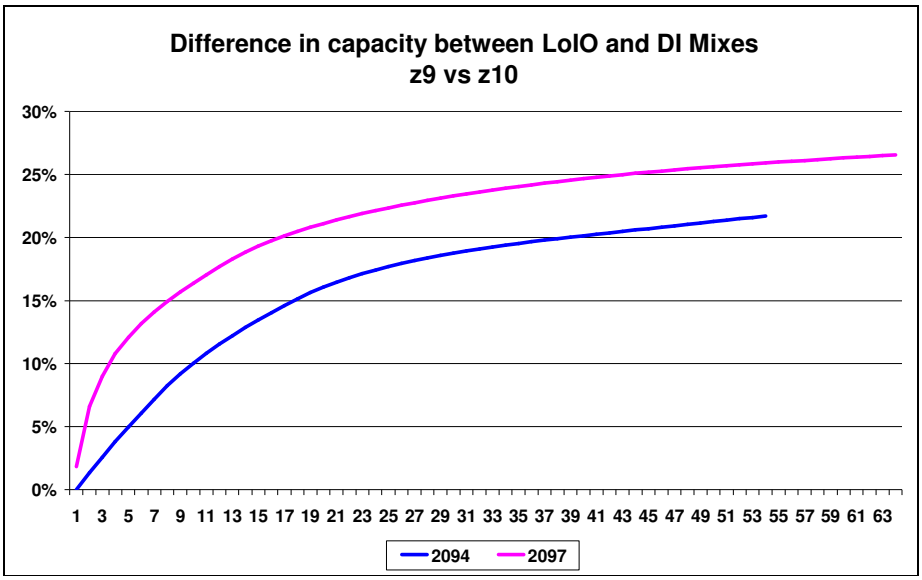

Figure 2

The LoIO-Mix workload is still representing most of customer production systems. You can use the LoIO-Mix workload benchmarks if your system is doing less than 30 DASD I/O per second per used MSU.
Used MSU has to be divided by a coefficient which depends on the hardware where the system is running[6].
You have to use the following values[7]:

- z890, z990 0,91
- z9 0,81
- z10 0,73

However in the last Share and in zPCR documentation IBM is warning about special low DASD I/O rate situations which show higher task switching rates and poorer HSB hit ratios.

This can occur when extensive uses of various Data-in-Memory (DIM) techniques have been employed. DIM has the effect of lowering the actual DASD I/O rate, while the processor's internal activities (task switching and HSB hit ratios) remain essentially the same as though DIM were not

[6] MSU values for the System z10, z9, z990, and z890, recorded in SMF 70 records, are discounted values IBM sets to provide improved software price/performance at each new machine generation.
[7] EPV for z/OS customers can simply look at the IOM index in Resources Monthly Trends which already provides adjusted values.

being use. As a result, some cases can exist where the LoIO-Mix workload should not be the only workload considered for capacity planning purposes[8].

The DI-Mix is intended to represent exactly these situations.
Unfortunately at the moment there is no scientific way to determine when DI-Mix should be used instead of LoIO-Mix[9].
DI-Mix is composed of the OLTP-W workload primitive only. OLTP-W is a web-enabled on-line workload accessing a traditional data base (WebSphere front-end to connect to CICS/DB2). It has been used because it is the most memory intensive workload included in LSPR.

IBM says that the consideration of DI-Mix is particularly important especially when migrating to z10 and customers may have to use it as a new lower bound of capacity. Unfortunately the difference between LoIO and DI mixes is so big that this approach could be extremely expensive for customers.
Based on your knowledge of the workload and on your judgment you have to decide which workload mix to use.

5 Conclusions

Experienced capacity planners know that every time a new generation of machines becomes available the evaluation of their capacity compared to old machines is never a trivial exercise.
The innovative design and the HiperDispatch feature of z10 make the estimation capacity a much more challenging exercise than in the recent past.
The cache design based on a smaller Level 1 cache and on a new intermediate cache called Level 1.5 appears to be particularly critical. This seems to be the main reason for an increased workload variability when running on z10 machines.
Recent IBM studies showed that a new DI-Mix workload could be hidden in an apparent LoIO-Mix.
The difference in capacity between the two mixes, as shown in this paper, is already very big and it will become even bigger on z10.
Unfortunately at the moment there is no scientific way to determine when DI-Mix should be used.
In this paper we provide the most up to date information available to help you take this decision.
Finally, we describe important changes made to LSPR benchmarks and zPCR and provide some advice to allow you to perform more reliable capacity planning studies.

[8] From zPCR help system.
9 New metrics have been added through APAR OA22414 in SMF 23 records but no algorithm has been still developed to test if DI-Mix is appropriate.