Logical processor topology in HiperDispatch mode

Fabio Massimo Ottaviani – EPV Technologies

January 2017

1 Introduction

HiperDispatch's main goal is to maximize the probability for a workload to be re-dispatched to the same group of logical and physical processors, in order to allow it to reuse instructions and data previously loaded in the cache levels closer to the processor.

In order to decide the logical processor topology (mapping to the physical processors) HiperDispatch implements vertical polarization to split the LPAR logical processors in three pools:

- High polarity (high processor share); they will have a target share corresponding to 100% of a physical processor which will be pseudo-dedicated to each of them;
- Medium polarity (medium processor share); they will normally have a target share greater than 0% and less than 100% of a physical processor; these medium logical processors have the remainder of the LPAR's shares after the allocation of the logical processors with the high share; they will compete for the physical processors as before HiperDispatch;
- Low polarity (low processor share); they will receive a target share corresponding to 0% of a physical processor; these are considered discretionary logical processors which are not needed to allow the LPAR to fully utilize the physical processor resource associated with its weight; if there is not unused capacity left from other LPARs in the CEC they will be parked and not used.

By using the information provided in SMF 70 and 113 many consequences of the HiperDispatch activity can be analyzed. However, no information is provided in these records about the logical to physical processors topology.

This information is available in the SMF 99 subtype 14 (SMF 99-14) records and it can be easily collected in an SQL DB by EPV zParser customers.
However the IBM WLM team decided to make processor topology accessible to everyone and provided a free tool: the WLM Topology Report.

The tool provides an Excel spreadsheet that displays lots of interesting information such as:

- the association of logical processors to books, chips, drawers, and nodes,
- the vertical polarization of the processors (high, medium, low),
- the processor type (CPU and zIIP),
- the association to WLM affinity nodes,
- any topology changes.

In this paper, we will show what you have to do in order to install and use the WLM Topology Report. We will also show and discuss two real life examples.

2 Collecting the necessary SMF records

Information about HiperDispatch topology is provided in SMF 99-14. Records are written every 5 minutes or whenever a topology change occurs.

The most common reasons for a topology change are:

- Configuration changes
- Partition weight changes (which can be due to WLM soft capping).

A bit in the SMF99E_VCM_Flag1 field indicates if a topology change has occurred.

Records provide information by partition so to get a complete picture they have to be collected for all the partitions in the machine.

It's important to note that these records are always requested by IBM technical support when they are asked to investigate performance issues so it's highly recommended to collect them on a regular basis.
To collect SMF 99 subtype 14 records you have only to allow it in SMFPRMxx.

3 Downloading and installing the WLM Topology Report

The following pre-requisites apply:

- SMF 99-14 records belonging to a system running in HiperDispatch mode on a z10 or newer machine have to be available;
- MS Excel Version 2007, 2010 or 2013 has to be used.

The topology report tool can be downloaded from the IBM web site by using the following link:
http://www-03.ibm.com/systems/z/os/zos/features/wlm/WLM_Further_Info_Tools.html#Topology

The installation is quite simple; you have just to download and run the SetupTopologyReport.exe program. It will install the tool in the TopoReport folder in Windows Program Files and create the IBM RMF Performance Management shortcuts in the Windows start menu.

The topology report doesn't directly process SMF records. They have to be pre-processed on z/OS in order to produce a CSV file that is the needed input for the tool.

The following picture shows the content of the TopoReport folder.

Nome	Ultima modifica	Tipo	Dimensione
DocuTopo	14/01/2017 15:43	Cartella di file	
HostTopo	14/01/2017 15:43	Cartella di file	
Changes.txt	17/02/2016 16:36	File TXT	1 KB
TopoReport.xlsm	17/02/2016 16:36	Foglio con attivazi...	150 KB
uninst.exe	14/01/2017 15:43	Applicazione	35 KB

Figure 1

In the HostTopo folder, you will find two files:

- a load library, including the S99ERPTD program which converts SMF 99 subtype 14 records to CSV format;
- a JCL library, including the SAMPLE JCL to be used to run the program.

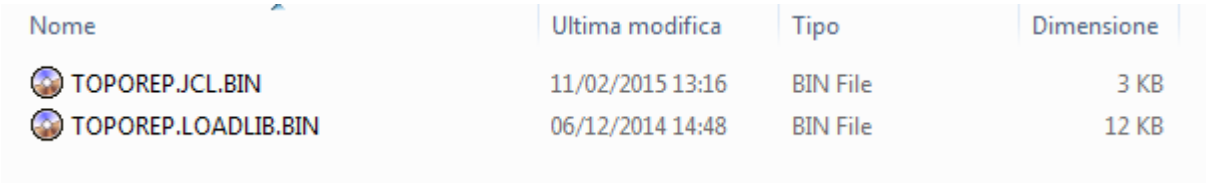


Figure 2

All the instructions to transfer these files to z/OS and use the tool can be obtained by clicking the Topo Report Help shortcut under the IBM RMF Performance Management entry in the Windows start menu.

4 Producing the CSV file

The SAMPLE JCL to be customized and run is provided below.

You have to:

- include your job card;
- set the name of the dataset including SMF 99-14 in SMFDSN;
- set the name of the output dataset including data in CSV format in RPTDSN;
- set the high level qualifier (hlq) for the TOPOREP.LOADLIB.

```
//     SET    SMFDSN=
//     SET    RPTDSN=
//JOBLIB    DD DSN=<hlq>.TOPOREP.LOADLIB,DISP=SHR
//*************************************************
//DECLARE   EXEC PGM=IEFBR14
//INPDS     DD DSN=&SMFDSN.,
//          DISP=SHR
//OUTDS     DD DSN=&RPTDSN.,
//          UNIT=SYSDA,
//          DISP=(MOD,DELETE),SPACE=(CYL,1)
//*************************************************
//READSMF   EXEC PGM=S99ERPTD
//SMFDATA   DD DISP=SHR,DSN=*.DECLARE.INPDS
//SYSPRINT DD SYSOUT=*
//SYSUDUMP DD SYSOUT=*
//REPORT    DD DISP=(NEW,CATLG),
//          SPACE=(CYL,(50,50),RLSE),
//          UNIT=SYSDA,
//          DSN=*.DECLARE.OUTDS
//USRPARMS DD *
  ALLINTV
/*
```

If you want to load only the first record and the records where a topology change occurred, you have to comment out the ALLINTV parameter.

The resulting file has to be transferred in ASCII mode to the Windows system where the topology report tool is installed.

5 Understanding the report – Case 1

To produce the report you have to open the TopoReport.xlsm spreadsheet. This is what you get.

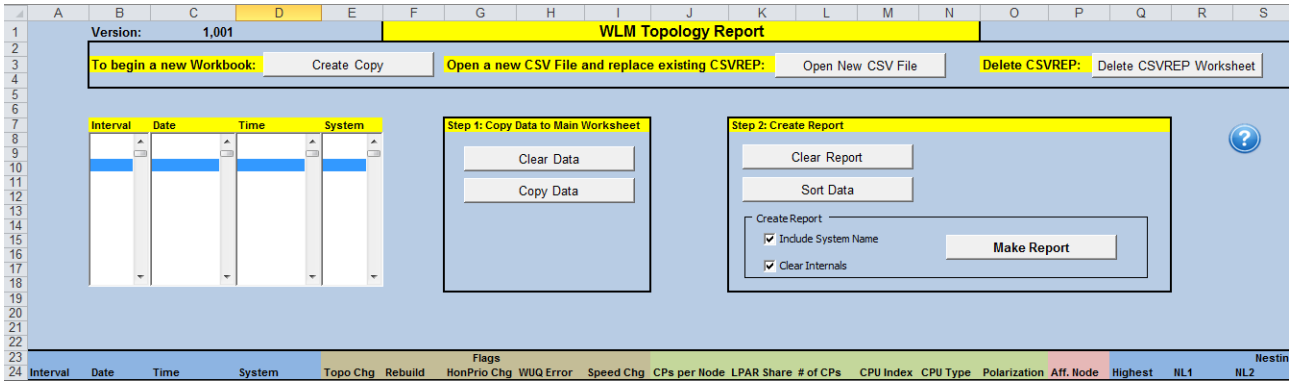


Figure 3

It's a good practice to create a new file to keep the original workbook clean by using the Create Copy button.

To load the CSV file, prepared as described in chapter 4, you have to click the Open New CSV File button.

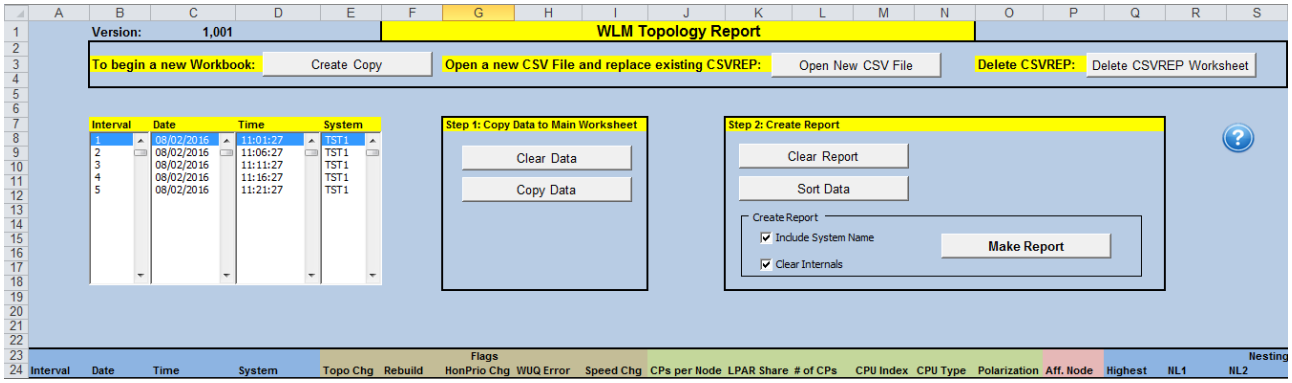


Figure 4

The window in the left section shows the records found. In this example, we collected only 5 records covering the time interval between 11:01:27 and 11:21:27 on February the 8th 2016.
All the records belong to the TST1 system. In this system, all the SMF records are synchronized to 00. You can note that SMF 99-14 records are not synchronized instead.

Once you have selected the desired interval you have to click the Copy Data button to get the information provided in the lower section of the report.

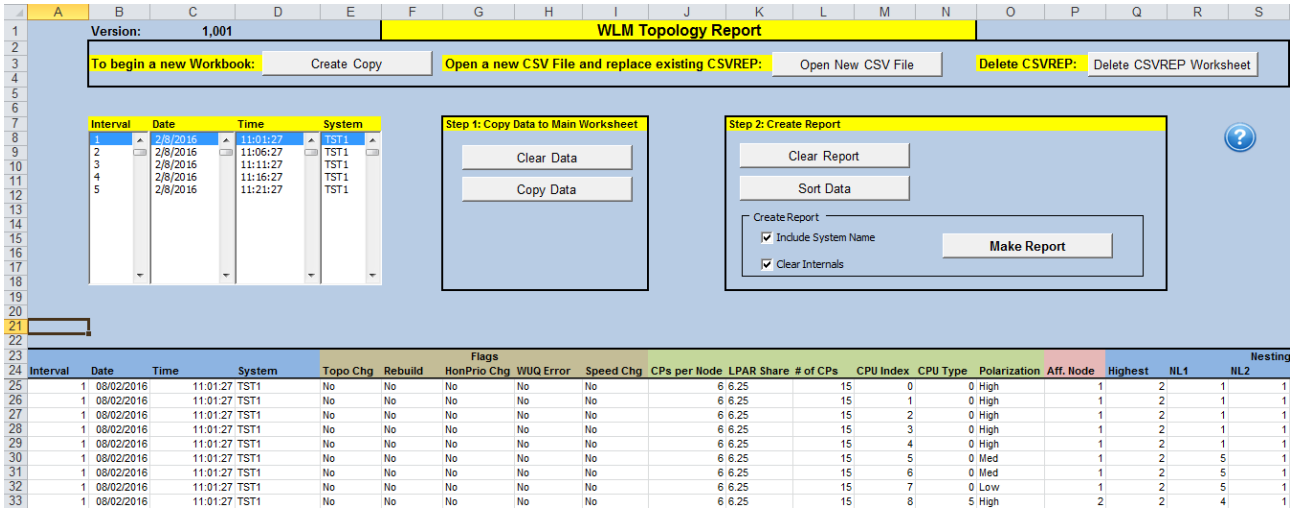


Figure 5

The Flags columns tell you if specific events happened in the selected interval:

- Topo Chg, topology has changed;
- Rebuild, affinity nodes have been rebuilt;
- HonPrio Chg, the honour priority parameter has been changed;
- WUQ Error, an error occurred on a dispatcher queue;
- Speed Chg, the processor speed changed.

The most relevant information provided in the next columns are the logical processor type (0=CPU, 5=zIIP), its polarization and the WLM affinity node they are assigned to.

Scrolling to the right, in the Nesting Levels columns you will get information about the processor cache architecture of the machine.
TST1 runs on a zEC12 so there are only two nesting levels: NL1 (Chip) and NL2 (Book).
For each logical processor, you know which Chip and Book it is dispatched onto. At the moment, no information about the physical processor is provided.
You may see that the first 5 logical processors are CPUs with High polarization and they are all assigned to Chip 1 and Book 1.

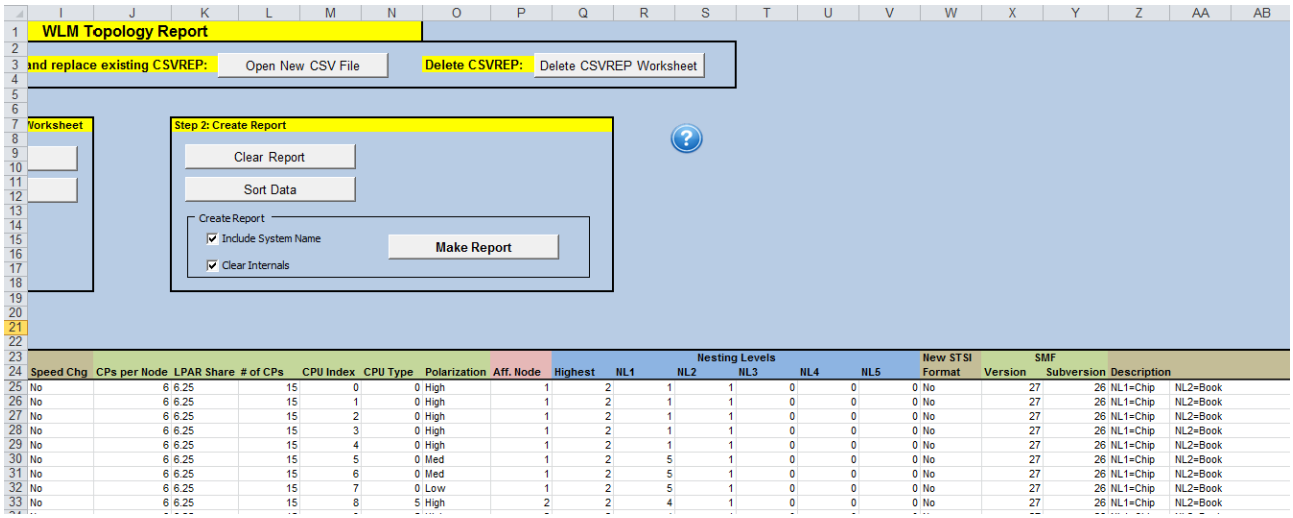


Figure 6

By clicking the Make Report button, you will get a clearer picture of the logical to physical processor topology.

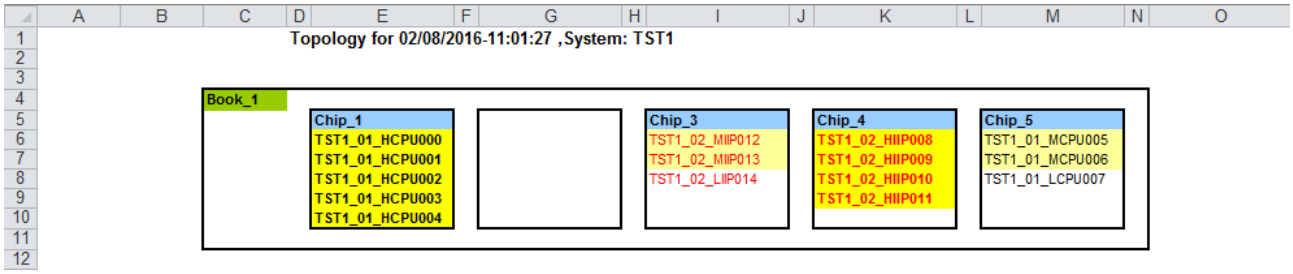


Figure 7

"The logical processors are depicted with the following information: SSSS_NN_Vtttnnn which means:

- SSSS = SMF system id
- NN = WLM affinity node number
- V = polarization = {H,M,L}
- ttt = processor type ={CPU,IIP,AAP}
- nn = processor number

High processors are also depicted in yellow, medium processors in light yellow and zIIPs with red color.[1]"

You can note that:

- all the TST1 logical processors are dispatched to Book 1;
- WLM affinity node 1 includes all the CPUs: the 5 CPUs with High polarization are dispatched to Chip 1 while the other CPUs (2 Medium and 1 Low) are dispatched to Chip 5;
- WLM affinity node 2 includes all the zIIPs: the 4 zIIPs with High polarization are dispatched to Chip 4 while the other zIIPs (2 Medium and 1 Low) are dispatched to Chip 3.

6 Understanding the report – Case 2

In this second example, we refer to systems running on a big z13 machine.

Figure 7 shows the topology for the SYS1 system. You can note that both the CPUs and the zIIPs are split on 2 Drawers and 3 Nodes (even if all the High are in Drawer 3 and most of them in Drawer 3 - Node 1).

[1] From the WLM Topology Report help.

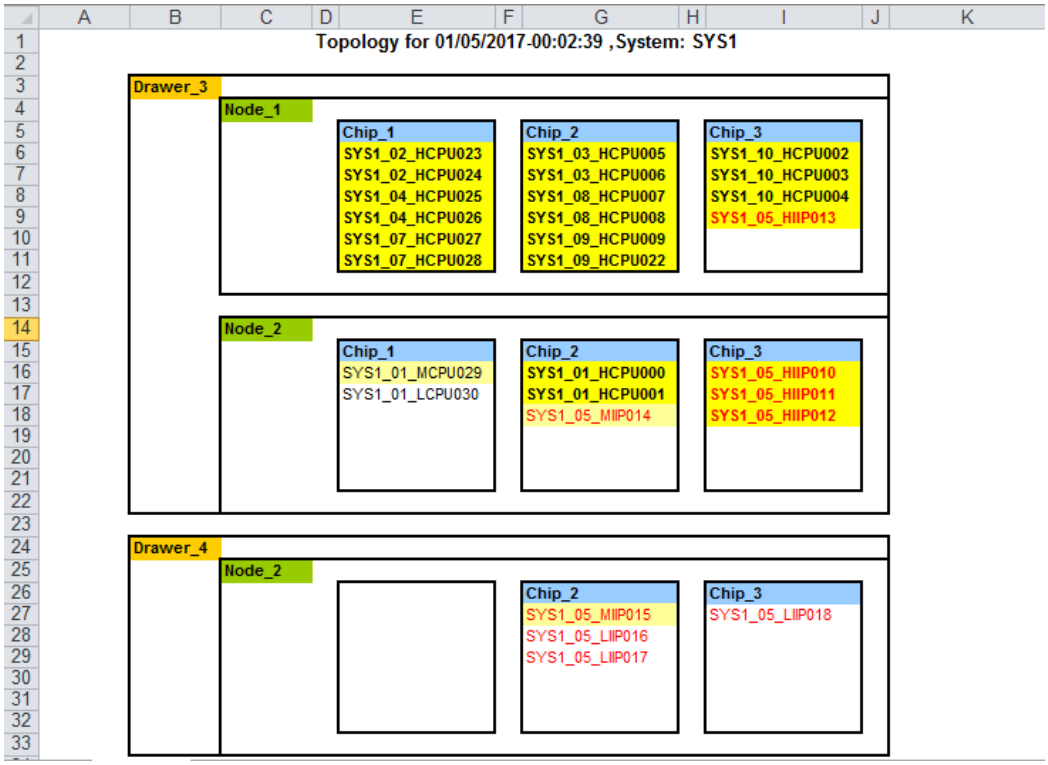


Figure 7

The tool doesn't allow you to select more than one system at a time so you have to do more reports to get a complete picture.

Figure 8 shows the topology for the SYS2 system. You can note that all the CPUs and zIIPs are assigned to Drawer 2. CPUs are split between the two Nodes while most of the zIIPs are in Drawer 2 - Node 1.

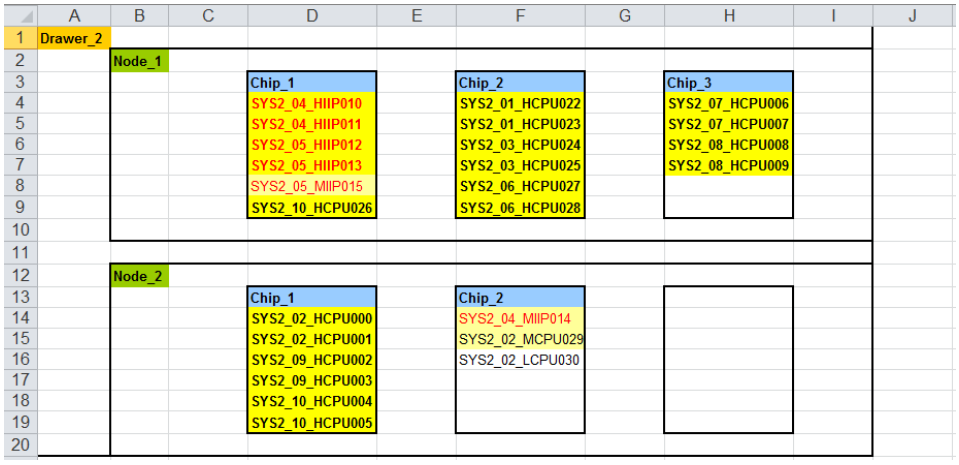
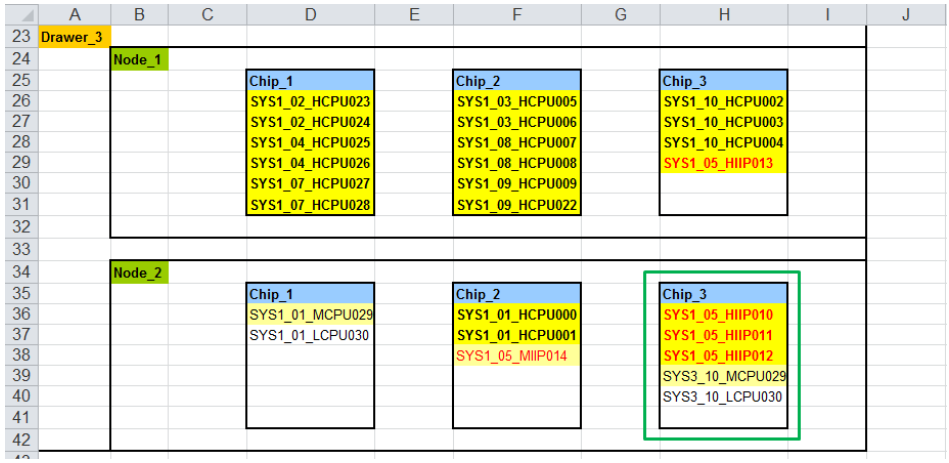


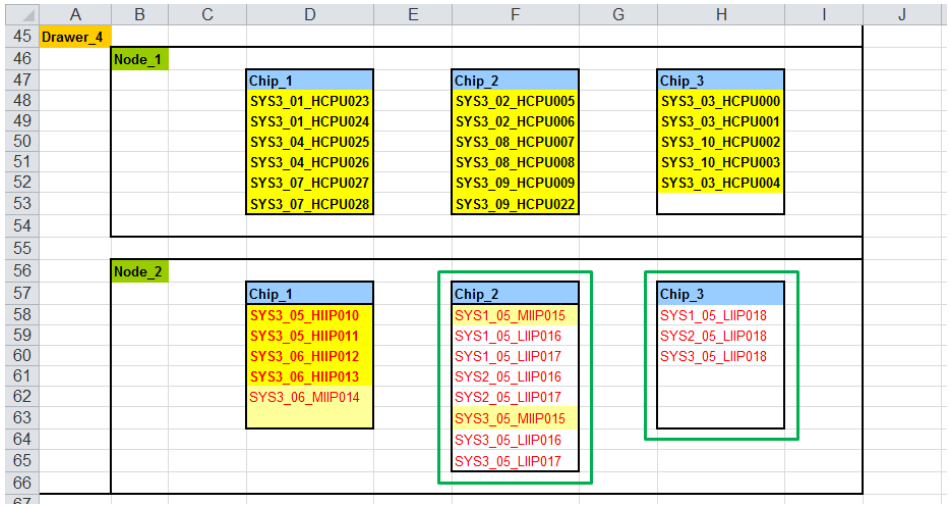
Figure 8

We also produced the SYS3 report and we discovered that it uses the same drawers, nodes and, in some cases, the same chips as SYS1. So we manually integrated SYS1 and SYS3 reports and highlighted the shared chips with a green border.



	A	B	C	D	E	F	G	H	I	J
23	Drawer_3									
24		Node_1								
25				Chip_1		Chip_2		Chip_3		
26				SYS1_02_HCPU023		SYS1_03_HCPU005		SYS1_10_HCPU002		
27				SYS1_02_HCPU024		SYS1_03_HCPU006		SYS1_10_HCPU003		
28				SYS1_04_HCPU025		SYS1_08_HCPU007		SYS1_10_HCPU004		
29				SYS1_04_HCPU026		SYS1_08_HCPU008		SYS1_05_HIIP013		
30				SYS1_07_HCPU027		SYS1_09_HCPU009				
31				SYS1_07_HCPU028		SYS1_09_HCPU022				
32										
33										
34		Node_2								
35				Chip_1		Chip_2		Chip_3		
36				SYS1_01_MCPU029		SYS1_01_HCPU000		SYS1_05_HIIP010		
37				SYS1_01_LCPU030		SYS1_01_HCPU001		SYS1_05_HIIP011		
38						SYS1_05_MIIP014		SYS1_05_HIIP012		
39								SYS3_10_MCPU029		
40								SYS3_10_LCPU030		
41										
42										

Figure 9



	A	B	C	D	E	F	G	H	I	J
45	Drawer_4									
46		Node_1								
47				Chip_1		Chip_2		Chip_3		
48				SYS3_01_HCPU023		SYS3_02_HCPU005		SYS3_03_HCPU000		
49				SYS3_01_HCPU024		SYS3_02_HCPU006		SYS3_03_HCPU001		
50				SYS3_04_HCPU025		SYS3_08_HCPU007		SYS3_10_HCPU002		
51				SYS3_04_HCPU026		SYS3_08_HCPU008		SYS3_10_HCPU003		
52				SYS3_07_HCPU027		SYS3_09_HCPU009		SYS3_03_HCPU004		
53				SYS3_07_HCPU028		SYS3_09_HCPU022				
54										
55										
56		Node_2								
57				Chip_1		Chip_2		Chip_3		
58				SYS3_05_HIIP010		SYS1_05_MIIP015		SYS1_05_LIIP018		
59				SYS3_05_HIIP011		SYS1_05_LIIP016		SYS2_05_LIIP018		
60				SYS3_06_HIIP012		SYS1_05_LIIP017		SYS3_05_LIIP018		
61				SYS3_06_HIIP013		SYS2_05_LIIP016				
62				SYS3_06_MIIP014		SYS2_05_LIIP017				
63						SYS3_05_MIIP015				
64						SYS3_05_LIIP016				
65						SYS3_05_LIIP017				
66										

Figure 10

More systems run on this machine. To get the complete picture we should create a report for each system and integrate them manually.

7 Conclusions

The tool can be very useful to understand the processor placement and how it changes when topology and other changes occur.
Most relevant current limitations are:

- SMF records are not synchronized with other SMF and RMF records;
- No information about logical to a specific physical processor is provided;
- Only one system at a time can be selected.