A Comprehensive Study of Monolithic 3D Cell on Cell Design Using Commercial 2D Tool

O. Billoint[1], H. Sarhan[1], I. Rayane[2], M. Vinet[1], P. Batude[1], C. Fenouillet-Beranger[1], O. Rozeau[1], G. Cibrario[1], F. Deprat[1], A. Fustier[1], J-E. Michallet[1], O. Faynot[1], O. Turkyilmaz[1], J-F. Christmann[1], S. Thuries[1], F. Clermidy[1]

[1]Univ. Grenoble Alpes, F-38000 Grenoble, France
CEA, LETI, MINATEC Campus, F-38054 Grenoble, France

[2]Mentor Graphics, 110 rue Blaise Pascal, 38330 Montbonnot-Saint-Martin, France

Email: olivier.billoint@cea.fr, Tel: (+33) 4.38.78.59.31

***Abstract*—In this paper we present a methodology allowing an emulated-3D two tiers physical implementation of any design using 2D commercial tools. Place and Route is achieved through similar steps as required by 2D designs: pre clock tree synthesis (including placement), clock tree synthesis and routing; to which we added a folding step in order to emulate the 3D placement. Routing of both tiers in parallel using inter-tier metal layers is made possible by modifying input files of the tools. Our study covers power supply network on both tiers, forbidden inter-tier via on active placement and inter-tier back end flavors in order to refine quality of results. Benchmark results on two tiers 3D Monolithic integration have been done on several IPs (microcontroller, reconfigurable FFT and LDPC) using as reference ST 28nm FDSOI technology and show the correlation between cell density, routing congestion, wire length, operating frequency and power consumption. To our knowledge, this paper is the first one to evaluate monolithic 3D physical implementation using full 3D Back End description and taking into account power supply distribution on both tiers.**

I. Introduction

3D Monolithic technology, which consists in fabricating MOS transistors above MOS transistors, is becoming more and more a reality [1]. Process technology is progressing each year, various design aspects such as thermal behavior [2], process corner variation between tiers [3], FPGA design exploration [4] or alternative physical implementation solutions [5] have been covered and the need for a tape-out compatible design environment is growing each year. To go one step further in this direction, we have developed a methodology based on a 2D Place and Route commercial tool that allows placing standard cells in an emulated-3D way using a folding technique (reaching an area reduction of 50%) and routing them using a specific extended 3D back end process based on 4 inter-tier metal layers combined with a regular 11 Copper based metal layers. State of the art papers [6][7] are placing both tiers in one run using show cell-shrinking techniques that we avoided in order to get rid of the tricky overlapping and legalization phases that can create a lot of Design Rules Check (DRC) errors, thus slowing the Place and Route tool a lot.

In [7], the inter-tier via parameters are added in a second step outside of the Place and Route flow before timing analysis, making it more difficult to evaluate its true benefits as it wasn't placed by the tool as part of the timing-driven routing strategy. Moreover, power supplies are connected only at the borders of the block [8], leading to an optimistic routing channels evaluation. In this paper we present a methodology that allows emulating 3D placement while offering a full back end routing capability on both tiers at the same time to meet full parasitic-aware timing closure and Power Distribution Network (PDN) setup. The whole methodology is explained below, from input file modification to place and route steps. Evaluation has been done on three growing complexity blocks : a open MSP430 microcontroller, a reconfigurable butterfly FFT and a LDPC. For each block, we have compared several implementations with and without PDN: 2D, 2D with extended back end (BE) implemented as 4 under-tier metal layers (copper based), 2D with extended back end implemented as 4 under-tier metal layers (tungsten based), 3D with 4 inter-tier metal layers (copper based) and finally 3D with 4 inter-tier metal layers (tungsten based). Operating frequency of the blocks ranges from around 1 GHz (open MSP and LDPC) to almost 2 GHz (reconfigurable butterfly FFT), showing consistent results for every case. The paper is organized as follows: section II is an overview of the 3D Monolithic process technology, section III details the file modifications needed to extend the Place and Route tool awareness of the 3D back end, section IV is explaining the emulated 3D methodology and section V shows the results we've obtained on several benchmark blocks.

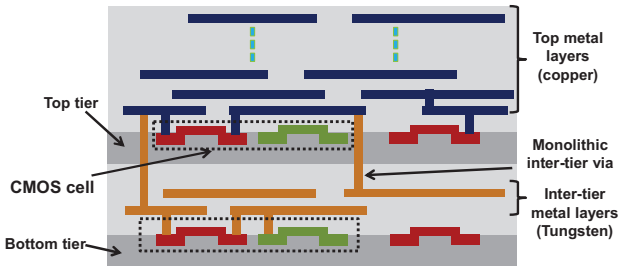

Fig 1. Monolithic 3D cell-on-cell integration technology.

This work was funded thanks to the French national program 'IRT Nanoelec' ANR-10-AIRT-05 and by the ST/IBM/LETI Alliance Program.

II. 3D Monolithic process technology overview

Monolithic 3D integration (M3D) is the stacking process where a top transistor layer is fabricated sequentially with the bottom one. It has been demonstrated that monolithic 3D integration can be achieved at low temperature process for the top tier (<600°C), providing inter-tier via diameter ~50 nm. The main advantages of 3DMI technology compared to TSV are fine-grain partitioning because of smaller inter-tier vias and less parasitics of the vertical interconnects.

Few previous works have studied the design techniques using 3DMI with two design approaches: transistor-level integration (N/P) and gate-level (cell-on-cell) integration. The N/P approach is implemented by splitting the standard cell (digital gate) so that the NMOS transistors are placed on one tier and the PMOS transistors on another tier. The monolithic 3D inter-tier vias are then used as connections inside the cell. On the other hand, the cell-on-cell approach is achieved by implementing the whole cell in one tier, either in the top or the bottom tier [9]. In this case, the inter-tier vias are used to connect top and bottom cells if needed.

The N/P design approach has the advantages of (i) compatibility with 2D design flow (but it requires designing of new 3D transistor-stacked standard cells) (ii) no inter-tier metal layers are needed for routing. On the other hand, cell-on-cell design approach has the advantage of using existing 2D standard cells, with the disadvantage of the incompatibility with the conventional physical implementation tools. Also, the cell-on-cell gate-level approach has other design issues such as the need of routing the bottom tier with the inter-tier metal layers to decrease the global routing congestion which will increase the fabrication costs. In addition, the clock tree balancing between the two tiers in the gate-level approach requires some additional design add-ons.

In this work we are focusing on cell-on-cell integration with 4 inter-tier metal layers. The inter-tier metal layers are implemented in two cases: (i) Copper and (ii) Tungsten, to show its effect on the design power and performances. Figure 1 shows a M3D cell-on-cell integration technology.

III. 3D Back End and standard cells description

First steps of our 3D-emulated physical implementation methodology is to prepare input files for the Place and Route flow. The back end description of standard cells in the bottom tier must be modified to reflect the Monolithic 3D technology effect. Conventional design methodology, whatever the tool, requires several files that are listed below:

- A technology lef file that includes design rules for routing (spacing, width, ...etc), description of vias (size, shape, ...etc) and process stack from bottom to top.
- A process technology file containing resistivity and coupling values between all metal and via layers of the process stack.
- Standard cells lef description files that contain physical descriptions of the cells (dimensions, area, pin placement, layers, ...etc)
- Standard cells lib description files that contain timing and power informations for each cell (rise and fall time against load, power, ...etc)

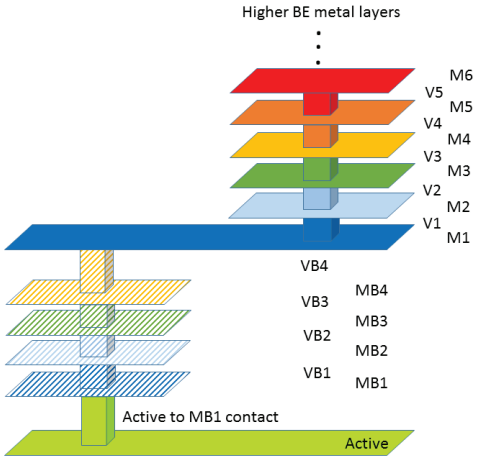

Fig 2. Technology lef 3D BE description.

A. 3D Back End description

For this study, we have been using as a reference the 28nm FDSOI technology from ST in its 11 metal layers configuration to which we've added 4 inter-tier metal layers. In our study, we have implemented two different flavors of the inter-tier metal layers, one based on Copper and another one based on Tungsten, in order to evaluate the process parameter impact on quality of results. In order to make the aforementioned files compatible with a 3D implementation flow, we need to modify them by adding informations about the inter-tier layers that we'll be using.

First of all, we must add the bottom tier process stack to the technology lef file. In our case, it is composed of 4 inter-tier metal layers (named MBx) as shown on figure 2. These metal layers are coded in the same way as the standard back end one, same thing for vias. Secondly, the process technology file is generated by the Place and Route tool starting from the Interconnect technology (ICT) file in which we have added the inter-tier metal layers in the same way the standard back end layers are described. We have generated two versions of these process technology files, one with inter-tier metal layers described using Copper properties (same as standard back end) and another one using worst case Tungsten properties that consist in multiplying the resistivity of metal lines and vias by a factor of 6 that is assumed to be a pessimistic case. State of the art [10] already implemented this methodology but with an « outside of the tool » step to add the top level parasitics for the inter-tier vias before timing analysis. Main advantage of our methodology is to give back end full awareness to the Place and Route tool from bottom to top tier without exception.

B. Standard cells description

Standard cells are composed of Metal 1 for intra-cell routing and Metal 2 for power supply rails. We are creating the lef and lib files for the bottom tier cells, that are exactly similar to the top cells, except for the metal layers in their lef description files. Consequently we'll be copying the files and changing Metal 1 (M1) by Metal Bottom 1 (MB1), via 1 (V1) by via bottom 1 (VB1) and Metal 2 (M2) by Metal Bottom 2 (MB2). No changes are done on wells and active regions as they are not creating DRC errors during Place and Route.

Compared to state of the art regarding cell placement strategy [7], our strategy gives the Place and Route tool full awareness of intra-cell routing and connectivity which is obviously better regarding routing strategy and optimization. One critical point to handle is the fact that inter-tier vias cannot go through active so this means that to be realistic, we have to add an obstruction into the standard cells lef description files for each gate in order to prevent the tool from placing an inter-tier via on a top cell. Vias are assumed to have the same size as regular back end copper vias, i.e. 50nm by 50nm for 28nm technology. Lib description files, containing all timing and power informations for each gate will stay the same, however it's interesting to notice that later on, these bottom tier cells lib files could easily be modified to reflect a process corner different from the top cells one.

IV. Methodology

As 2D commercial tool are handling one and only one cell layer, it is impossible to place cells from different tiers in parallel. However, in case we achieve to place cells in a 3D way, it is possible without increasing computing time (which could be due to placement errors detected by the tool) to route the cells using a user-defined back end as the one we've described previously. This concept will be the main driving argument of the methodology we are presenting below and shown on figure 3 (a).

A. Non optimal 3D placement emulation: folding technique

In order to somehow optimize the first placement and maximize efficiency of the starting point placed and routed database, we are doing a 2D Place and Route of the 2D synthesized netlist by putting the top I/O ports (including control signals like clock) at the middle of the chip in the y direction as shown on figure 3 (b).

Once 2D pre-Clock Tree Synthesis (pre-CTS) and Clock Tree Synthesis (CTS) are done, we are changing cells above the middle I/O ports from top tier to bottom tier. Bottom tier cells are then folded over top tier cells. Same methodology is applied to lateral I/O ports, if they exist, as well with a metal layer change that allows to avoid overlapping with existing ports. Obtained area reduction is then 50%.

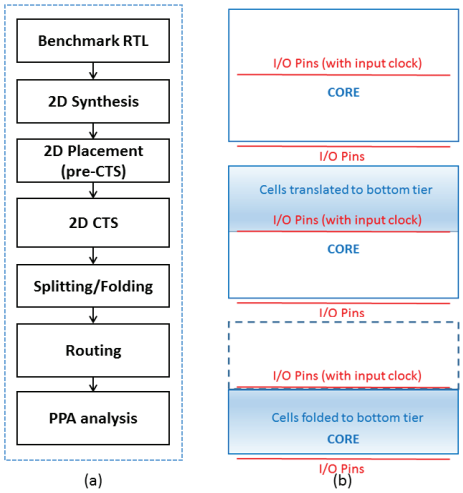

Fig 3. (a) Emulated-3D flow / (b) Folding technique

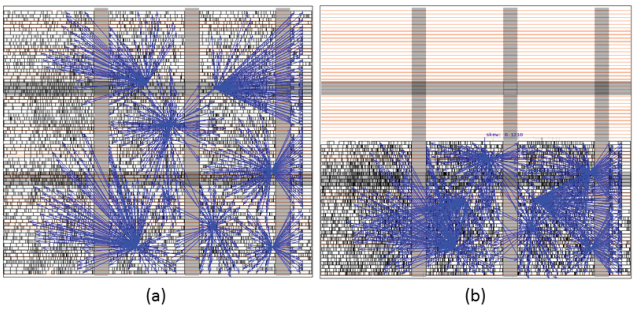

Fig 4. Physical view of clock tree folding from 2D to 3D

B. Clock Tree

The folding technique presented above transforms by construction a 2D clock tree into a 3D clock tree without any errors. Once clock tree cells have been assigned to a tier, routing phase is then relatively straightforward as the tool has full knowledge of the back end and access to all layers for bottom to top tier. Figure 4 shows the effect of the folding technique on the clock tree of the open MSP430.

C. Power Supply Distribution

Usually out of studies scope [8], Power Distribution Network (PDN) creates new constraints for 3D Monolithic physical implementation and needs to be clearly addressed in order not to over-estimate inter-tier routing capabilities and associated performances (operating frequency and power consumption). Two main parameters are impacting power distribution to the bottom tier, the first one is increased resistivity of wiring and vias for inter-tier BE, usually considered to be Tungsten. The second one is the y direction routing-obstruction created by vias used to connect top and bottom PDNs. In order to avoid adding too many inter-tier metal layers, we are assuming that there will not be a PDN for the bottom tier but that power will come from the top tier PDN and be distributed directly by vertical contacts from M2 down to MB2 using overlapping regions as shown in red on figure 5 below. This distribution requires to choose correctly the number of inter-tier vias because of their increased resistivity that may limit maximum current flow between the tiers. Figure 6 shows how connecting top and bottom tier power supply rails creates routing obstructions in the y direction.

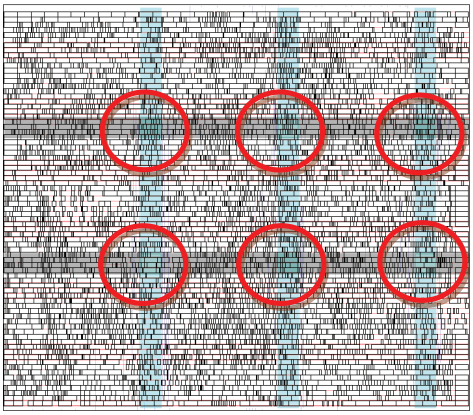

Fig 5. Floorplan of standard PDN with via insertion (in red)

Table I. 2D implementation results with conventional and extended back end description

		Wire Length (mm)			Power (mW)			Congestion			Operating
		top	bottom	**Total**	clock	data	**Total**	X	Y	Vias	Frequency
Open MSP 2D Area= 0.01mm^2 Density=76%	2D	78,8	--	**78,8**	3,5	5,4	**8,9**	0,1	0,14	0,035	**1,09 GHz**
	2D 4BM (Copper)	61,1	22,6	**83,7**	3,5	5,4	**8,9**	0,06	0,08	0,042	**1,09 GHz**
	2D 4BM (Tungsten)	62,5	23,6	**86,1**	3,6	5,5	**9,1**	0,06	0,08	0,043	**0,99 GHz**
FFT 2D Area= 0.033mm^2 Density=77%	2D	265,7	--	**265,7**	15,1	43,4	**58,5**	0,08	0,17	0,033	**1,75 GHz**
	2D 4BM (Copper)	208,4	72,1	**280,5**	15	44,4	**59,4**	0,05	0,1	0,044	**1,70 GHz**
	2D 4BM (Tungsten)	217,3	70,9	**288,2**	15,4	45,2	**60,6**	0,05	0,1	0,044	**1,54 GHz**
LDPC 2D Area= 0.1mm^2 Density=46%	2D	1613,9	--	**1613,9**	23,2	67,3	**90,5**	0,25	0,3	0,041	**0,83 GHz**
	2D 4BM (Copper)	998	591,2	**1589,2**	22,3	62,2	**84,5**	0,15	0,16	0,034	**0,93 GHz**
	2D 4BM (Tungsten)	1061,8	602	**1663,8**	18,1	67,9	**86**	0,15	0,17	0,036	**0,88 GHz**

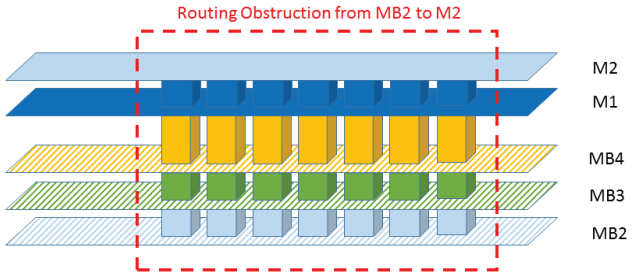

Fig 6. Routing obstructions due to PDN vias starting from MB2 and going to M1 & M2.

V. Results

In order to qualify the ability of the Place and Route tool to handle correctly the extended inter-tier back end, we have first been using Mentor Graphics Olympus to compare 2D physical implementation results of three benchmark blocks (Open MSP430, reconfigurable butterfly FFT and LDPC) with conventional and extended back end, which means four additional Bottom Metal layers (4BM) accessible under the cell layer. We have then applied the emulated-3D physical implementation methodology to compare results with 2D in different configurations like with and without PDN and with two extended back end flavors: one all Copper and the other one based on worst case Tungsten.

A. Comparison of 2D implementations with extended BE

Table I shows a comparison of block area, density, wire length by tier, power consumption for clock tree and data, operating frequency and congestion for the three benchmark blocks. First point is that adding more metal layers for routing lowers routing congestion in both x and y directions but increases via congestion for high density designs (open MSP430 and FFT), making it more critical for the tool to find some inter-cell space to put them, thus increasing wire length, power and decreasing operating frequency. On The contrary, for block with lower density but high routing congestion (LDPC), adding more metal layers for routing reduces congestion (x, y and vias) and allows to slightly increase in the operating frequency while lowering power consumption.

B. Comparison of 2D and 3D implementations

Table II shows the same parameters as Table I but between 2D and 3D, with and without PDN and with two different BE flavor for the four additional inter-tier metal layers. First remark is that using Tungsten BE increases slightly power consumption, this being mainly due to timing constraints applied on higher resistivity wires, leading to a higher number of buffers. Second remark is that the impact of the power distribution network is not negligible as it noticeably changes the congestion values for high density blocks (open MSP430 and FFT), thus impacting power consumption and wire length as it makes it more difficult for the tool to find a short path for point to point connections. That's why compared to PDN implementations, removing the PDN allows to increase the operating frequency of those blocks while reducing slightly power consumption. For lower density block (LDPC), compared to the 2D cases explained above, going 3D increases via congestion which leads to a reduction of the operating frequency.

C. Interpretation and consequences

Performance results of 3D implementation show that PDN and inter-tier BE flavor are impacting performances. Moreover, for high density designs, we clearly see that the intra-cell inter-tier via obstruction creates more constraints for the routing tool as it has to find a free inter-cell space to put inter-tier vias, leading to increased wire length and power consumption. Thus, 3D physical implementation of high density designs may require to save some space on the top tier for via implantation while keeping density as high as possible in order not to lose too much area. One possible solution to fulfill the two aforementioned conditions could be to aim a non 50/50 area ratio, for example 40/60, 40 being the top ratio and 60 the bottom one. With this solution, some more free space would be available at top tier in order to have enough space for inter-tier via implantation as well as for pads integration. Regarding lower density designs (mainly because of congestion issues), adapting the number of inter-tier metal layers could be a possibility to reduce wire length and power consumption while increasing the operating frequency.

Table II. 3D implementation results for with & without PDN cases, using either Copper or Tungsten for inter-tier metal layers

		Wire Length (mm)			Power (mW)			Congestion			Operating
		top	bottom	**Total**	clock	data	**Total**	X	Y	Vias	Frequency
Open MSP 2D Area= 0.01mm^2 Density=76%	2D	78,8	--	**78,8**	3,5	5,4	**8,9**	0,1	0,14	0,035	**1,09 GHz**
	3D 4BM (Copper)	51,1	56,5	**107,6**	3,6	6,3	**9,9**	0,1	0,1	0,041	**1,03 GHz**
	3D 4BM (Copper / no PDN)	33,4	42,9	**76,3**	3,5	5,9	**9,4**	0,07	0,06	0,031	**1,14 GHz**
	3D 4BM (Tungsten)	53,1	62,5	**115,6**	3,8	6,4	**10,2**	0,11	0,1	0,043	**0,9 GHz**
	3D 4BM (Tungsten / no PDN)	34	46,1	**80,1**	3,6	5,9	**9,5**	0,08	0,06	0,032	**1,04 GHz**
FFT 2D Area= 0.033mm^2 Density=77%	2D	265,7	--	**265,7**	15,1	43,4	**58,5**	0,08	0,17	0,033	**1,75 GHz**
	3D 4BM (Copper)	195,8	186,2	**382**	14,9	51,3	**66,2**	0,1	0,12	0,054	**1,45 GHz**
	3D 4BM (Copper / no PDN)	129,2	124,4	**253,6**	14,4	47,2	**61,6**	0,06	0,08	0,034	**1,83 GHz**
	3D 4BM (Tungsten)	208,1	198,3	**406,4**	15,2	52,2	**67,4**	0,1	0,13	0,056	**1,25 GHz**
	3D 4BM (Tungsten / no PDN)	132,5	135	**267,5**	14,5	47,7	**62,2**	0,06	0,08	0,034	**1,69 GHz**
LDPC 2D Area= 0.1mm^2 Density=46%	2D	1613,9	--	**1613,9**	23,2	67,3	**90,5**	0,25	0,3	0,041	**0,83 GHz**
	3D 4BM (Copper)	1140,1	493,7	**1633,8**	21,3	67,6	**88,9**	0,18	0,15	0,047	**0,62 GHz**
	3D 4BM (Copper / no PDN)	1080,4	456,8	**1537,2**	21	63,3	**84,3**	0,16	0,13	0,04	**0,81 GHz**
	3D 4BM (Tungsten)	1212,1	521,7	**1733,8**	18,8	74,5	**93,3**	0,19	0,17	0,051	**0,58 GHz**
	3D 4BM (Tungsten / no PDN)	1125,1	471,1	**1596,2**	18	69,7	**87,7**	0,17	0,14	0,042	**0,72 GHz**

VI. Conclusion

We have demonstrated an emulated-3D physical implementation flow using a folding technique that goes through the same Place and Route steps as required by a 2D design with the addition of one extra step that is splitting/folding. Cells are first placed in a 2D design footprint before clock tree synthesis, then 50% area reduction is done using the splitting/folding technique, finally inter-tier routing is done in one run and connects correctly all pins from both tiers while meeting timing closure. This "one tool only" methodology should be applicable to any design including those with memories and allows to easily estimate the impact of different options on 3D physical implementation. In order to fully evaluate the benefit of Monolithic 3D technology (especially the possibility to noticeably increase the operating frequency while reducing wire length and power consumption in a 50% reduced footprint), optimization of cell placement between tiers is mandatory as it is the only possibility to reach the optimal cell to tier configuration. From the results point of view, emulated-3D physical implementation show that power supply distribution, even if it's not a showstopper is one major point to be addressed in studies as it lowers routing resources thus performances. As well, depending on cell density, inter-tier via placement without going through the transistor active region can create via placement congestion, thus impacting negatively wire length and power consumption. Finally, inter-tier back end flavor has some impact on power consumption which may be controlled and managed by the Place & Route tool if tier to tier cell placement optimization becomes feasible.

VII. References

[1] P. Batude et al., "3D sequential integration opportunities and technology optimization", Interconnect Technology Conference / Advanced Metallization Conference (IITC/AMC), 2014 IEEE International

[2] S. Samal, K. Samadi, P. Kamal, Y. Du, and S.K. Lim, "Full Chip Impact Study of Power Delivery Network Designs in Monolithic 3D ICs", IEEE International Conference on Computer-Aided Design, 2014.

[3] S. Panth, K. Samadi, Y. Du and S. Lim, "Power-Performances Study of Block-Level Monolithic 3D ICs Considering Inter-Tier Variations", Design Automation Conference (DAC), 2014 51st ACM/EDAC/IEEE , vol., no., pp.1,6, 1-5 June 2014.

[4] O. Turkyilmaz, G. Cibrario, O. Rozeau, P. Batude, F. Clermidy, "3D FPGA using high-density interconnect Monolithic Integration", DATE 2014.

[5] H. Sarhan, S. Thuries, O. Billoint, and F. Clermidy, "3DCoB: A new design approach for Monolithic 3D Integrated Circuits", Asia and South Pacific Design Automation Conference (ASP-DAC), IEEE, 2014.

[6] S. Panth, K. Samadi, Y. Du, and S.K. Lim, "Design and CAD Methodologies for Low Power Gate-level Monolithic 3D ICs", IEEE International Symposium on Low Power Electronics and Design, 2014.

[7] S. Bobba, et al., "CELONCEL: Effective design technique for 3-D monolithic integration targeting high performance integrated circuits." Proceedings of the 16th Asia and South Pacific Design Automation Conference. IEEE Press, 2011.

[8] Y-J. Lee and S.K. Lim, "Ultra High Density Logic Designs using Monolithic 3D Integration", IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Vol. 32, No. 12, pp. 1892-1905, 2013.

[9] C. Liu and S. Lim, "A Design Tradeoff with Monolithic 3D Integration", Quality Electronic Design (ISQED), 13th International Symposium on, IEEE, 2012.

[10] S. Panth, K. Samadi, Y. Du, S.K. Lim, "Placement-Driven Partitioning for Congestion Mitigation in Monolithic 3D IC Designs", ACM International Symposium on Physical Design, 2014.