3D Logarithmic Interconnect: Stacking Multiple L1 Memory Dies Over Multi-Core Clusters

Erfan Azarkhish, Igor Loi, and Luca Benini
DEI, University of Bologna
Bologna, Italy
erfan.azarkhish@unibo.it, igor.loi@unibo.it, and luca.benini@unibo.it

7th International Symposium on Networks-on-Chip
April 2013

Outline

Objectives:

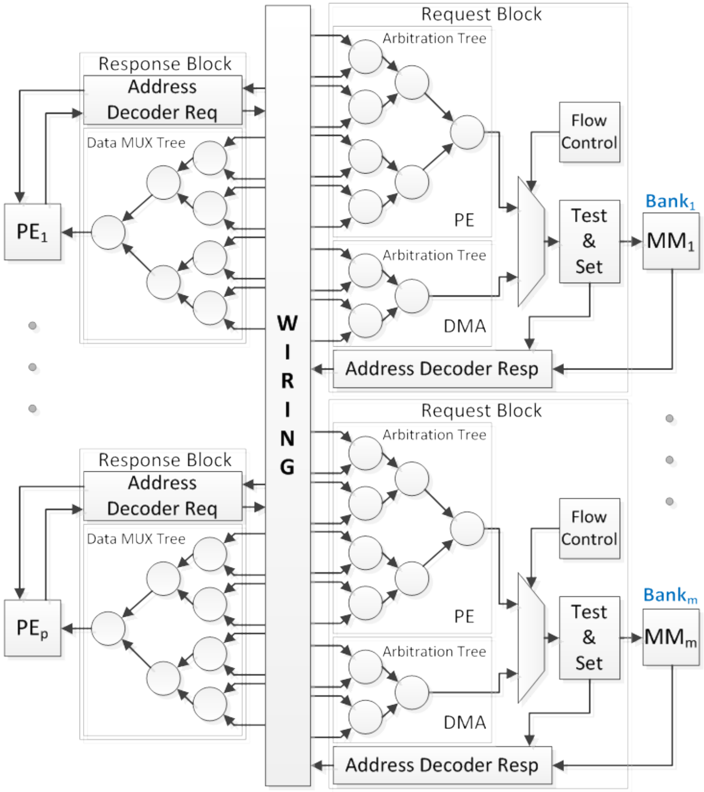
- We propose two synthesizable 3D network architectures: C-LIN and D-LIN, which can be integrated with 3D Stacking technology to provide access to tightly coupled shared memory banks stacked over multi-core clusters.
- We devise a modular design strategy which allows users to stack multiple memory dies and create different height stacks with identical dies, without the need for different masks for dies at different levels in the stack.
- Two Through Silicon Via (TSV) technologies are used:
  - Micro-bumps and Cu-Cu Direct Bonding
  - With consideration of the ESD protection circuits.

1. Overview of 2D Logarithmic Interconnect
2. 3D Design Alternatives

   C-LIN and D-LIN
3. 3D Integration Issues
4. Experimental Results
5. Conclusions

Logarithmic Interconnect (LIN)

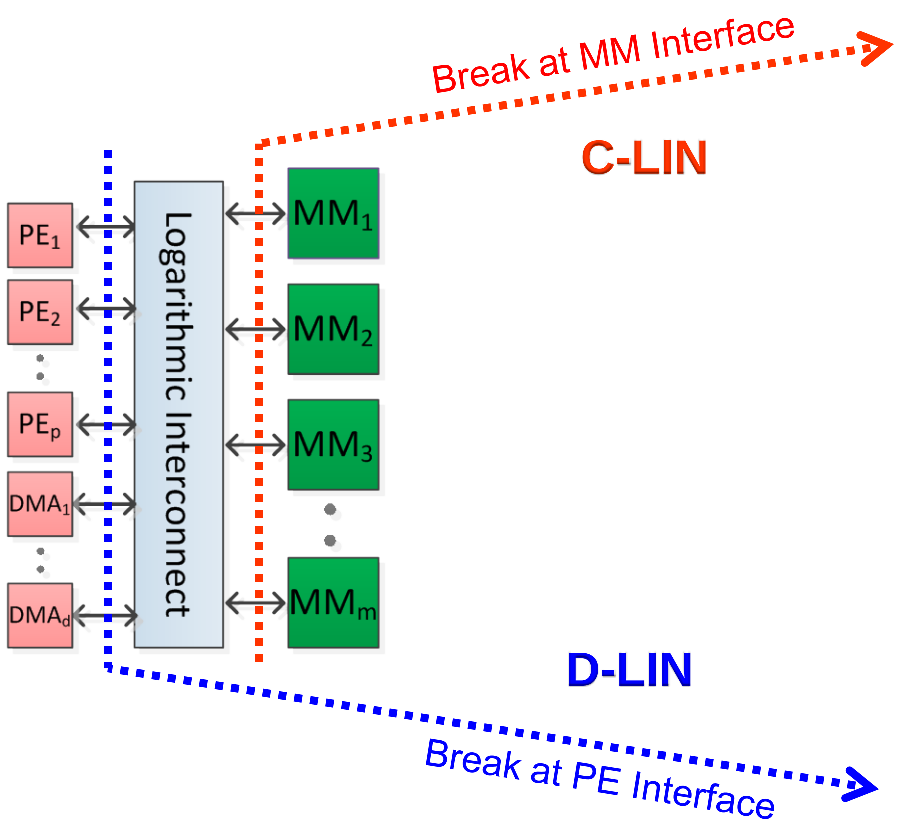
- A low-latency and flexible crossbar that connects multiple processing elements (PEs) to multiple SRAM memory modules (MMs).

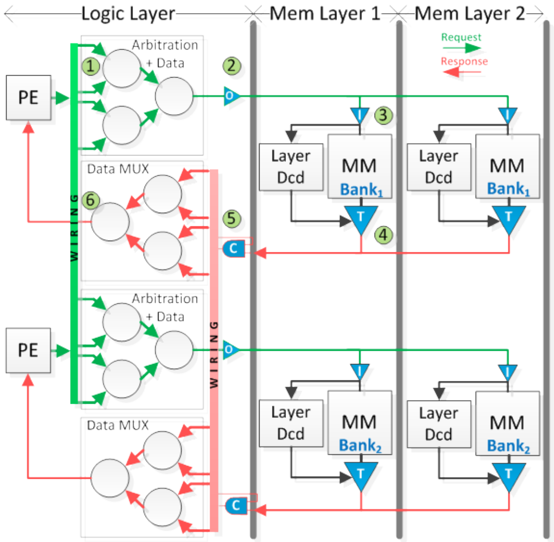


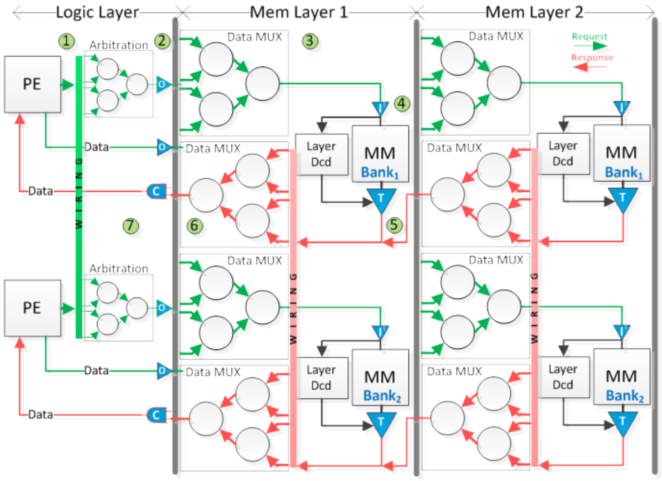
Parameter	Definition
P	Number of processor channels
D	Number of DMA channels
M	Number of memory modules
S	Size of memory modules (KB)
W	Width of data bus (b)
Arbitration Method	Pseudo LRG / Pseudo Round-Robin
Interleaving Method	Word Level (WLI) / Bank Level (BLI)

+ Synthesizable RTL Description
+ Configurable Parameters

3D Design Choices







C-LIN vs. D-LIN

C-LIN and D-LIN:

- Overcome the 2D limitation by automatically splitting the design into:
  - one logic layer
  - several memory layers with identical layouts, stacked over each other
- All parameters are automatically configured during the boot procedure.

  ➔ reduction in the chip cost and design effort

C-LIN:

- Logic and memory elements are completely separated

  ➔ different technologies and optimizations may be utilized for design of the logic and memory dies.
- Memory layers in C-LIN can be designed as simple, small, and inexpensive as possible.

D-LIN:

- Reduction in number of TSVs (Banking Factor > 1)

3D Integration Issues

- **ESD Protection**

→ TSV Micro-buffers

- **Identical Layouts for Memory Dies**

→ Boot-time Configuration

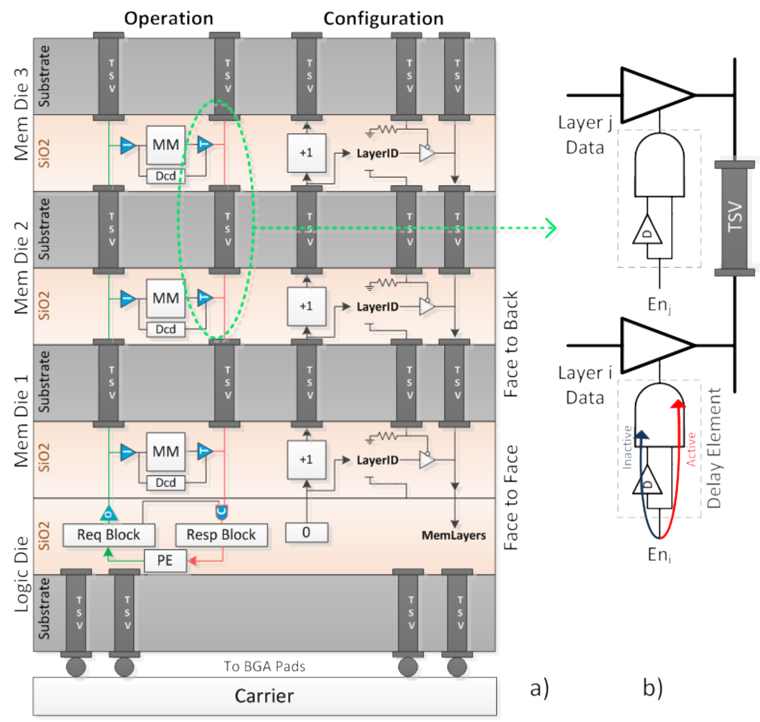
- **Process/Voltage/Thermal Variations**

→ High Current Glitches

→ Glitch Removal Delay Element

→ Clock Skew among memory layers

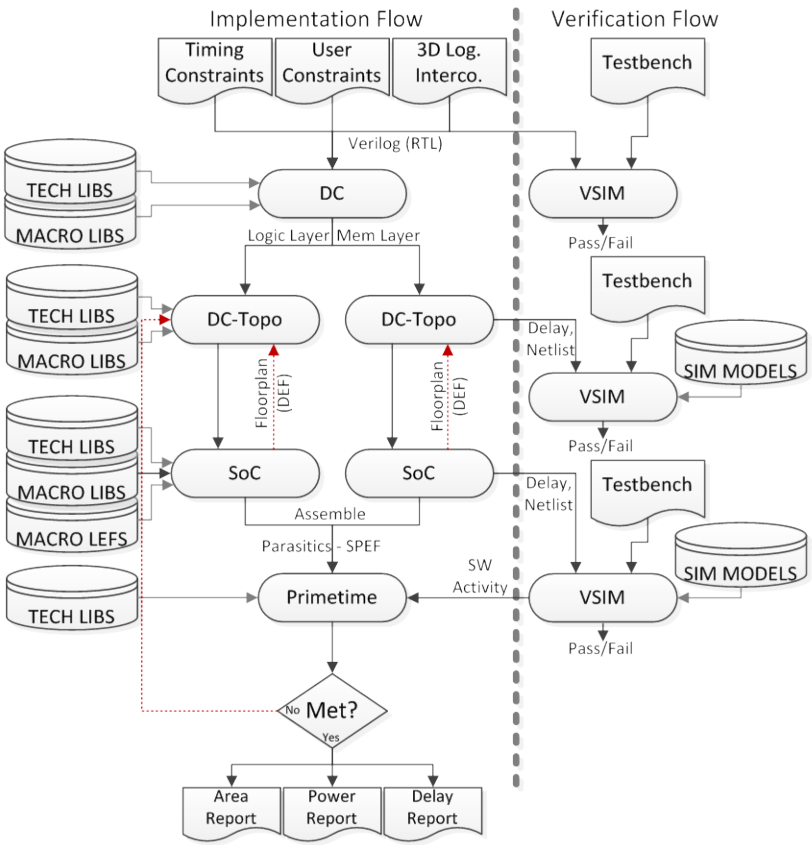
→ Handled in Clock Tree Synthesis Phase



Physical Design Flow

Hierarchical Design Flow is utilized

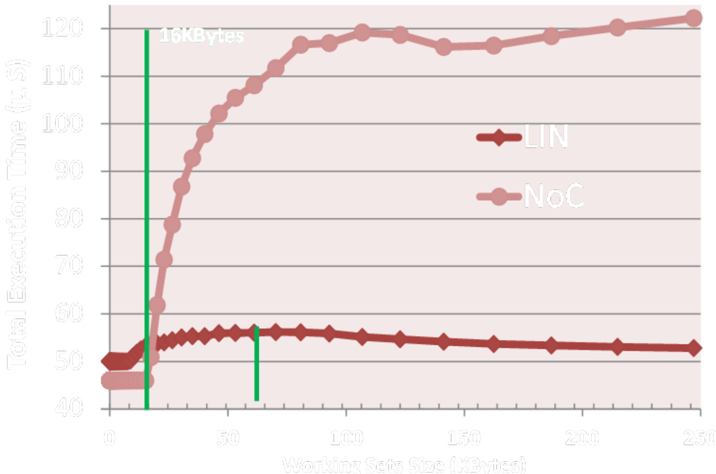
- **Technology Library:**
  - STM CMOS-28nm Low Power
- **Front-end Synthesis:**
  - Synopsys Design Compiler Graphical (2011.09)
- **Place and Route:**
  - Cadence SoC Encounter Digital Implementation (10.1)
- **Sign-off Tasks:**
  - Primetime (2011.09)



Experimental Results

- **Baseline Configuration:**
  - 16 STMicroelectronics xP70 ASIP RISC PEs
  - On-chip TCDM with 32 memory banks
- **2D Design:**
  - Memory Bank from 8KB to 64KB
- **3D Designs:**
  - Fixed Memory Bank Size of 8KB
  - Number of Stacked Memory Dies: 8
- **Models:**
  - Micro-bumps: (40m x 50m pitch) [5]
  - Cu-Cu Direct Bonding (10m x 10m pitch) [5]
  - TSV Capacitive Load: 30fF [6]

Comparison of LIN with Other Topologies

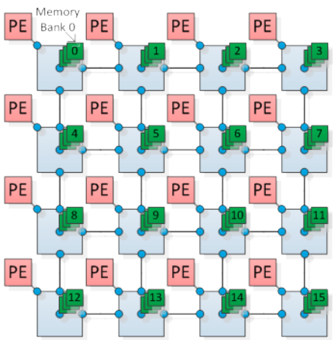


Performance comparison between NoC and LIN (Random Test-cases)

Interconnect	Cardinality	Area (mm²)
LIN	(p=16,m=32)	0.09
NoC-3.6GHz [1]	4x4	0.29
MIRA (3DM) [2]	4x4	0.40
MIRA (3DM-E) [2]	4x4	0.40
NoC-5.1GHz [3]	4x4	1.02

Comparison of post-layout area between LIN and NoCs

NoC 4x4
vs.
LIN 16x32

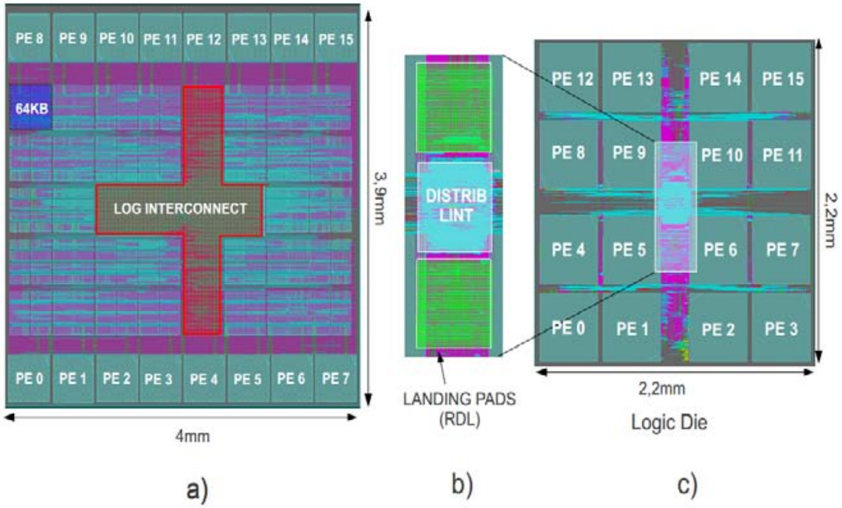


Benchmarks		FAST	CT	SIFT
LIN	Execution Time (ms)	5.59	79.89	4464.07
	AMAT (ns)	6.54	6.47	6.46
NoC	Execution Time (ms)	8.21	106.92	4943.43
	AMAT (ns)	7.53	7.09	6.57
Bus	Execution Time (ms)	46.40	730.51	30799.99
	AMAT (ns)	81.30	82.40	81.90

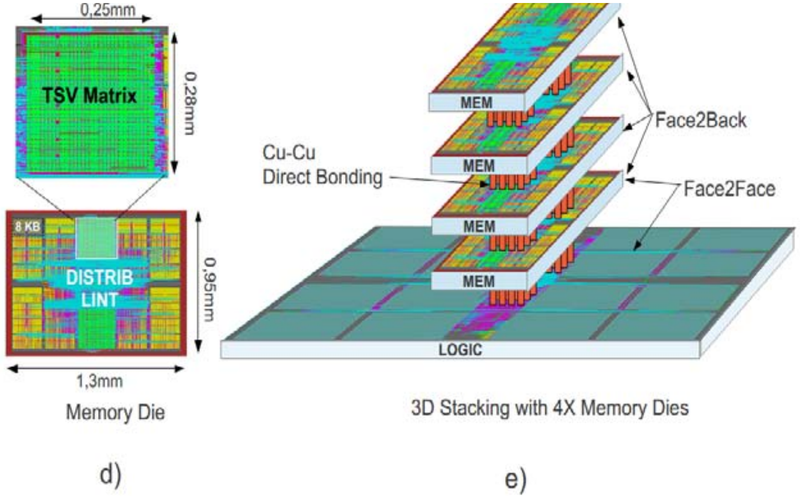
Performance comparison between LIN, NoC, and Bus (benchmarks [7])

Physical Implementation

a) 2-LIN with 2MB SRAM
b) Details of the landing pads (RDL) in D-LIN
c) Logic Die of D-LIN with Cu-Cu Direct bonding



d) Memory die of D-LIN with details of the TSV Matrix
e) 3D Stacking with 4 stacked memory dies



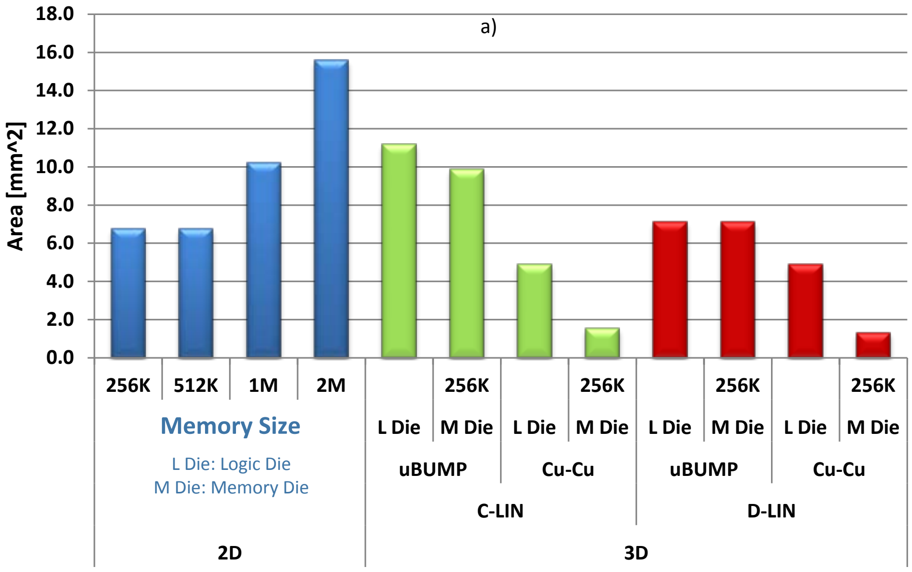
Number of TSVs:

C-LIN: 2688

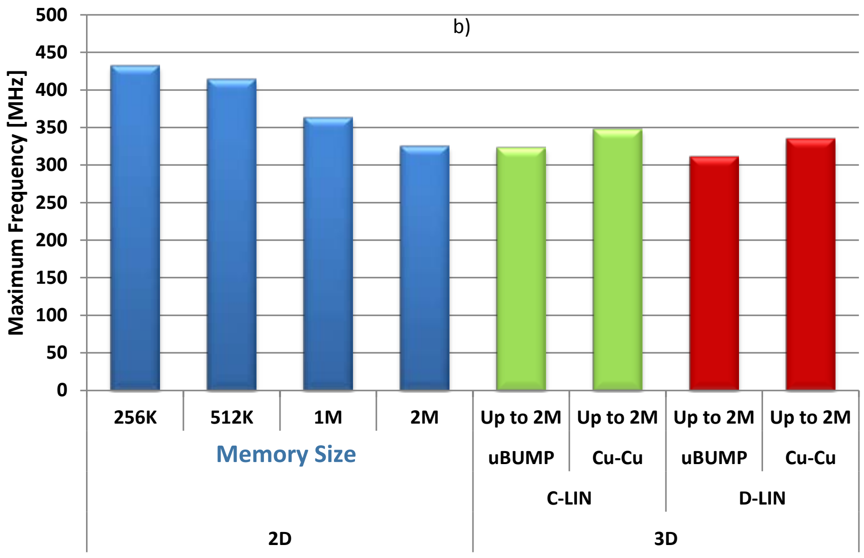
D-LIN: 1424

➔ 47% Reduction

Comparison of Design Alternatives



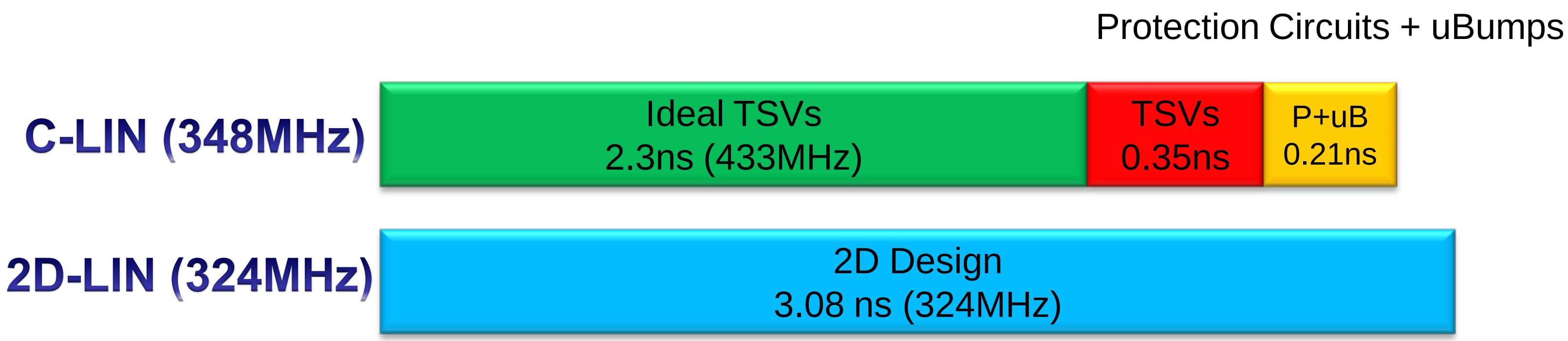
Comparison of silicon area (mm^2) between 2D and 3D alternatives



Comparison of Maximum Achievable Frequency (MHz) between 2D and 3D alternatives

Discussion

- C-LIN and D-LIN improve the performance over 2D-LIN with the same memory size by small factors of 6.7% and 3.7% respectively



Conclusions

- In processor-to-L1-memory context, LIN outperforms both traditional network on chips (NoC) and simple time-division multiplexing buses.
- For large 2D designs the main problems are routing congestion, signal integrity, and the mask cost.
- Our proposed 3D designs offer better scalability with a similar performance, however, in terms of delay, the 3D designs are not so competitive with the 2D planar design, unless we go towards larger 2D chips.
- Even though the current TSVs are still not much better in terms of speed than global on-chip wires, they can provide more freedom in heterogeneous integration of dies with cost-optimized technologies, since they are definitely much better than traditional off-chip links.

References

1. A. Kumar et al. ”A 4.6Tbits/s 3.6GHz Single-Cycle NoC Router with a Novel witch Allocator in 65nm CMOS,” Proc. Int’l Conf. Computer Design (ICCD ’07), p. 63-70, Oct. 2007.
2. D. Park et al. ”MIRA: A Multi-layer On Chip Interconnect Router Architecture,” in Proceedings of the 35th Annual International Symposium on Computer Architecture (ISCA), June 2008.
3. S. Vangal et al. ”A 5.1GHz 0.34mm2 Router for Network-on-Chip Applications,” VLSI Circuits, 2007 IEEE Symposium on ,pp.42-43, 14-16.
4. G. Beanato et al., ”3D-LIN: A Configurable Low-Latency Interconnect for Multi-Core Clusters with 3D Stacked L1 Memory,” In Proceedings of the 2012 IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC’12), Santa Cruz, USA, October 2012.
5. E. J. Marinissen et al. ”Wafer Probing on Fine Wafer Probing on Fine-Pitch Pitch Micro Bumps for 2.5D and 3D SICs”, Int. Report IMEC http://www.swtest.org/swtw_library/2011proc/PDF/S04_03 Marinissen_SWTW2001.pdf
6. G. Van der Plas et al. ”Design Issues and Considerations for Low-Cost 3-D TSV IC Technology,” Solid-State Circuits, IEEE Journal of, Volume : PP , Issue:99 On page(s): 1 15 ISSN : 0018-9200
7. SW/HW extensions for heterogeneous multicore platforms (vIrtical) [Online]. Available. http://www.virtical.eu/
8. L. Benini et al. ”MPARM: Exploring the Multiprocessor SoC Design Space with SystemC,” VLSI Signal Processing 41(2): 169-182 (2005)
9. K. Kang, L. Benini, and G. D. Micheli, ”A High-throughput and Low-Latency Interconnection Network for Multi-Core Clusters with 3-D Stacked L2 Tightly-Coupled Data Memory”, In Proceedings of the 2012 IFIP/IEEE Inter-national Conference on Very Large Scale Integration (VLSI-SoC’12), Santa Cruz, USA, October 2012.

Thank you!

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA