ASP-DAC 2025
30th Asia and South Pacific Design Automation Conference

The Survey of 2.5D Integrated Architecture: An EDA Perspective

Shixin Chen[1], Hengyuan Zhang[2], Zichao Ling[2],
Jianwang Zhai[2], Bei Yu[1]

[1]The Chinese University of Hong Kong
[2]Beijing University of Posts and Telecommunications

Jan. 21, 2025

Outline

1. Why 2.5D?
2. EDA for 2.5D Architecture Design
3. Partition and Interconnection
4. EDA for 2.5D IC Physical Design
5. Conclusion

Why 2.5D?

LLM and Computation Demand

Fig 1. LLM is widely used and in high demand.

- **Huge amount of parameters**: GPT-3 has 175 billion parameters.
- **Inference computation**: Inference requires dozens of high-performance GPUs.
- **Training process**: Thousands of NVIDIA V100 GPUs are needed for training.
- **How can we improve computational capabilities?**

- More advanced tech-nodes: The commercial benefit is decreased after 28 nm.
- More computation resource: Wafer-scale chip is expensive
- More advanced architecture: **3D IC or 2.5D IC**

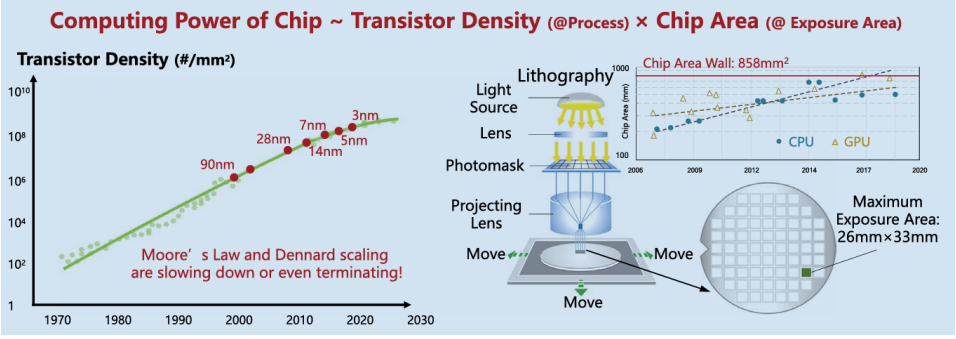


Fig 2. Moore's law and the chip area wall.[1]

[1]**Yang Hu, Xinhan Lin, et al. (2024). "Wafer-Scale Computing: Advancements, Challenges, and Future Perspectives [Feature]".** In: *IEEE Circuits and Systems Magazine* **24.1, pp. 52–81.**

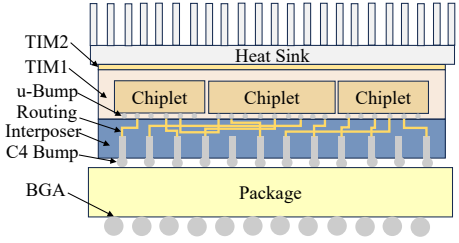


Fig 3. 2.5D package

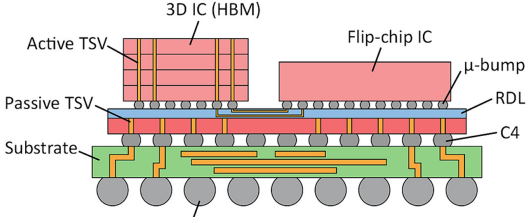


Fig 4. 2.5D and 3D hybrid package

- **Manufacturing Complexity:**
  2.5D: interposer for multiple chips;
  3D: advanced techniques like TSVs.
- **Thermal Management:**
  2.5D: heat dissipation on flat layout;
  3D: overheating from stacked chips (e.g., multiple DRAM layers).
- **Manufacturing Yield:**
  2.5D: defects in chip don't affect others;
  3D: defect in any chip can lead to total package failure.

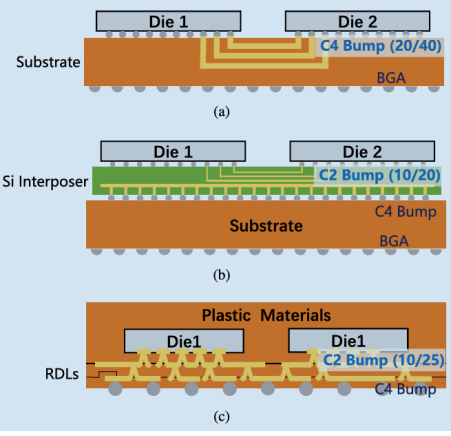


Fig 5. (a) substrate-based, (b) silicon-based, and (c) RDL-based packages (bump:μm)

EDA tools are essential for IC design, while the tools for 2.5D are still in development.

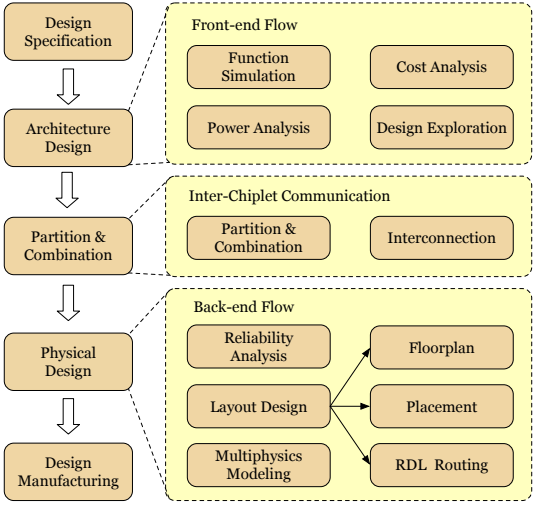


Fig 6. The EDA flow of the chiplet-based architecture.

EDA for 2.5D Architecture Design

The adopted simulator from NoC

Most chiplet simulators are based on simulation frameworks designed for NoC like Booksim[2], Noxim[3], and Sniper[4].

The following characteristics require more attention:

- Consider accurate latency of chiplet interactions
- Model heterogeneous system with different tech nodes
- Support various communication protocols

[2] **Nan Jiang, Daniel U Becker, et al. (2013). "A detailed and flexible cycle-accurate network-on-chip simulator".** In: *Proc. ISPASS*, pp. 86–96.

[3] **Vincenzo Catania, Andrea Mineo, Salvatore Monteleone, et al. (2016). "Cycle-accurate network on chip simulation with noxim".** In: 27.1, pp. 1–25.

[4] **Trevor E Carlson, Wim Heirman, et al. (2011). "Sniper: Exploring the level of abstraction for scalable and accurate parallel multi-core simulation".** In: *Proc. SC*, pp. 1–12.

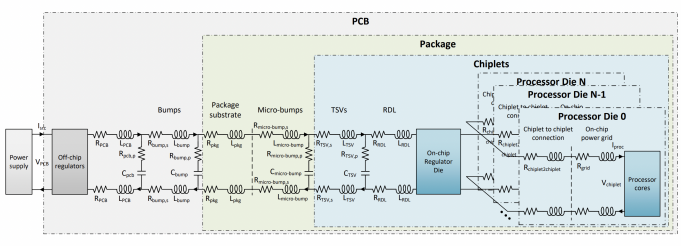


Fig 7. The model of a hybrid power deliver network for 2.5-D chiplet-based multicore systems

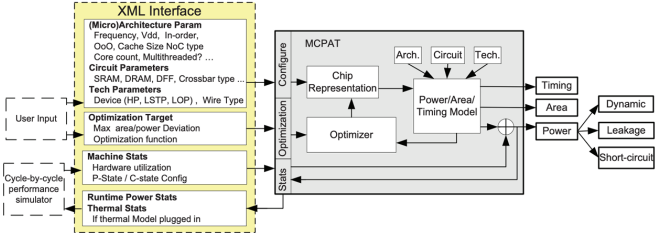


Fig 8. The mcpat-based modeling from simulation results

Cost Model

- Chiplet Actuary[5]: presents a quantitative cost model tailored for multi-chip systems
- These models will take into account as many factors as possible, such as materials, area, yield, know-good-die, all stages in manufacturing.

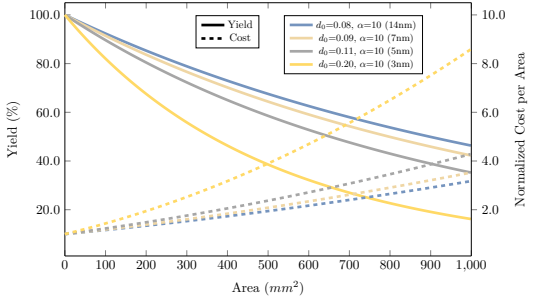
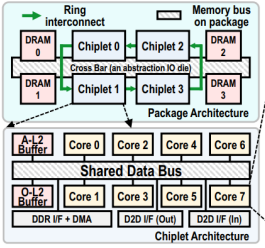


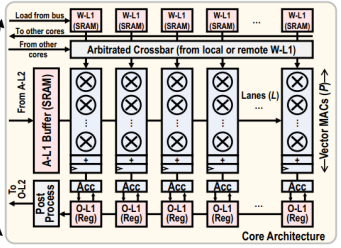
Fig 9. The cost, yield, and chip area trends with different technology nodes.

[5]**Yinxiao Feng and Kaisheng Ma (2022). "Chiplet actuary: a quantitative cost model and multi-chiplet architecture exploration".** In: *Proc. DAC*, pp. 121–126.

DSE framework

- RapidChiplet[6]: chiplet-based multicore architecture
- NN-Baton[7]: chiplet-based DNN accelerator design space exploration





Computation	Storage	Interconnect
Package-Level: mapping along HO & CO dimension		
N_p Chiplet parallelism	DRAM	Crossbar (DRAM) and ring bus (chiplet)
Chiplet-Level: mapping along HO & WO & CO dimension		
N_c Core parallelism	A-L2, O-L2	Shared data bus (multicast / unicast)
Core-Level: mapping along CI & CO dimension		
$P \times L$ MAC parallelism	A-L1, O-L1, W-L1	Array with multiple lanes (broadcast)

Fig 10. The NN-Baton chiplet exploration framework

[6] **Patrick Iff, Benigna Bruggmann, Maciej Besta, et al. (2023). "RapidChiplet: A Toolchain for Rapid Design Space Exploration of Chiplet Architectures".** In: *arXiv preprint*.

[7] **Zhanhong Tan et al. (2021). "NN-Baton: DNN Workload Orchestration and Chiplet Granularity Exploration for Multichip Accelerators".** In: *Proc. ISCA*, pp. 1013–1026.

Partition and Interconnection

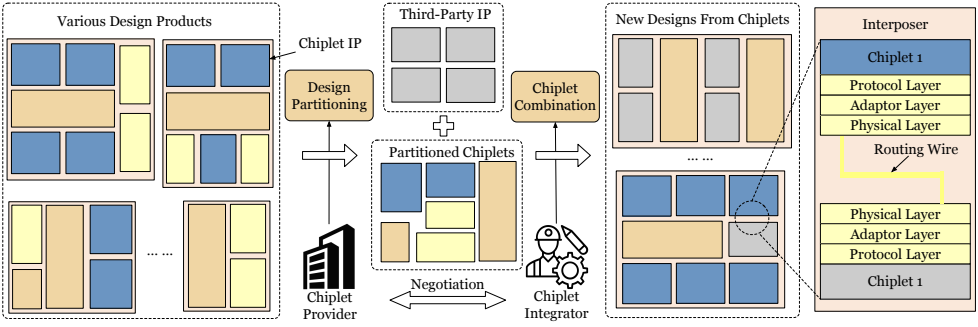


Fig 11. Illustration of partitioning and combining in chiplet-based architecture.

- Chipletizer[8]: Employs multi-layer partitioning and simulated annealing to enhance core reuse and reduce costs.

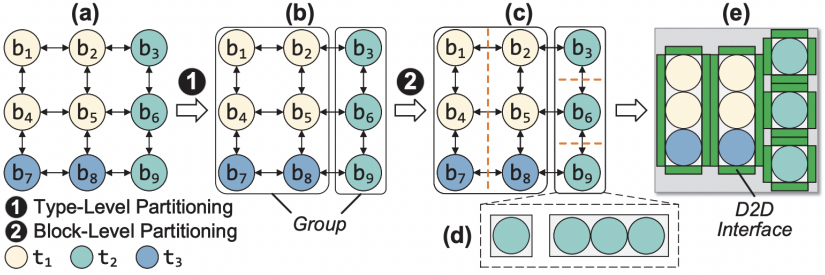


Fig 12. Two-level hierarchical partitioning

[8]**Fuping Li, Ying Wang, Yujie Wang, et al. (2024). "Chipletizer: Repartitioning SoCs for Cost-Effective Chiplet Integration".** In: *Proc. ASPDAC*, pp. 58–64.

Topology optimization

Kite[9]: Design a router-based network to improve data bandwidth and decrease data deadlock.

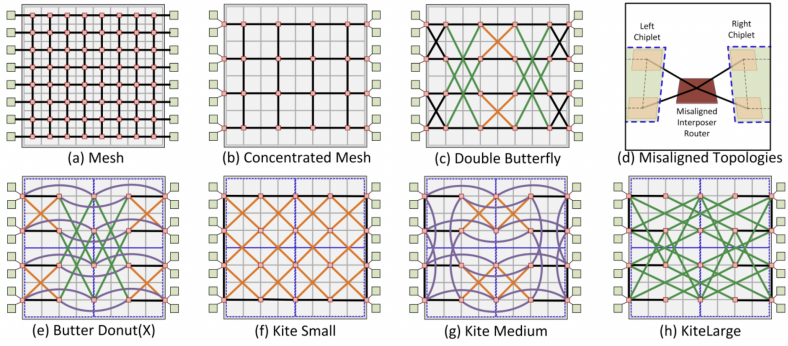


Fig 13. The various combination methods for multi-chiplet systems

[9] **Srikant Bharadwaj et al. (2020). "Kite: A family of heterogeneous interposer topologies enabled via accurate interconnect modeling".** In: *Proc. DAC*. IEEE, pp. 1–6.

- Interface protocols:[10]: heterogeneous (parallel and serial) interface to enable complex data flow.
- Optical-based interconnection[11]: decrease latency and improve flexibility.

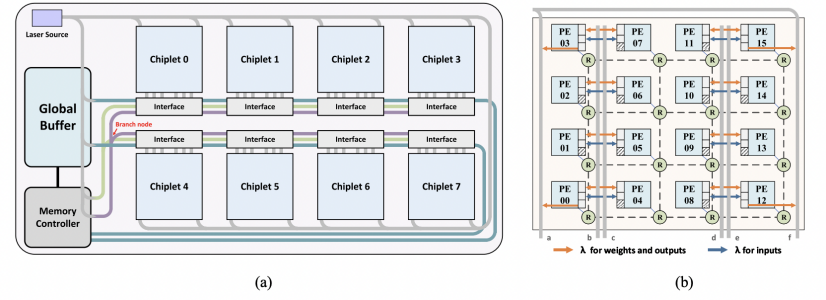


Fig 14. An eight-chiplet DNN accelerator with the proposed optical interface

[10] **Tianqi Wang et al. (2022). "Application defined on-chip networks for heterogeneous chiplets: An implementation perspective".** In: *Proc. HPCA*, pp. 1198–1210.

[11] **Guanglong Li and Yaoyao Ye (2024). "HPPI: A High-Performance Photonic Interconnect Design for Chiplet-Based DNN Accelerators".** In: *IEEE TCAD* 43.3, pp. 812–825.

EDA for 2.5D IC Physical Design

The methods to do floorplan&placemnet

- Heuristic methods[12]
- Mathematical analytic optimization[13]
- Machine learning approaches: RL-based[14]

[12] **Hong-Wen Chiou, Jia-Hao Jiang, et al. (2023). "Chiplet placement for 2.5 D IC with sequence pair based tree and thermal consideration".** In: *Proc. ASPDAC*, pp. 7–12.

[13] **Shixin Chen, Shanyi Li, Zhen Zhuang, et al. (2024). "Floorplet: Performance-Aware Floorplan Framework for Chiplet Integration".** In: *IEEE TCAD* 43.6, pp. 1638–1649.

[14] **Yuanyuan Duan, Xingchen Liu, et al. (2024). "RLPlanner: Reinforcement Learning based Floorplanning for Chiplets with Fast Thermal Analysis".** In: *arXiv preprint*.

Objective-driven floorplan:

- Performance-aware Floorplan[15]
- Thermal-aware Floorplan[16]
- Warpage-aware Floorplanning[17]

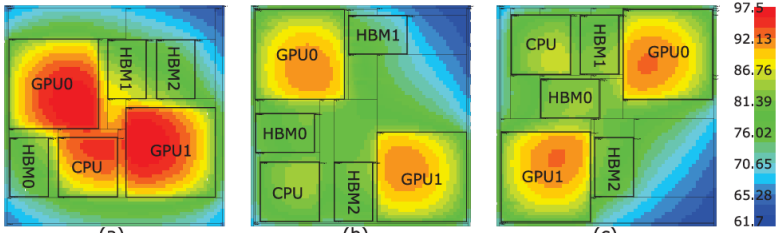


Fig 15. The different placement strategies will influence the thermal dissipation.

[15] **Shixin Chen, Shanyi Li, Zhen Zhuang, et al. (2024). "Floorplet: Performance-Aware Floorplan Framework for Chiplet Integration".** In: *IEEE TCAD* 43.6, pp. 1638–1649.

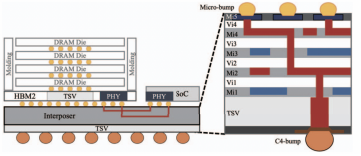


Fig 16. A cross-section view of the interposer layers that consist of five metal layers

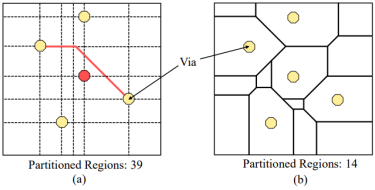


Fig 17. Vias can be placed at arbitrary locations with any angle[18]

[18] **Min-Hsuan Chung, Je-Wei Chuang, and Yao-Wen Chang (2023). "Any-angle routing for redistribution layers in 2.5 D IC packages".** In: *Proc. DAC*, pp. 1–6.

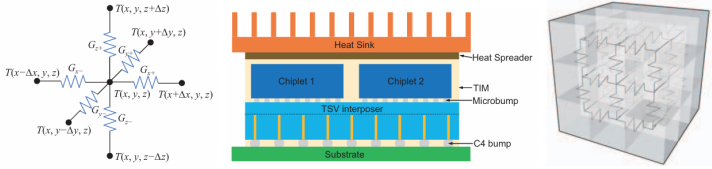


Fig 19. Thermal resistance circuit for a thermal cell and the thermal resistance network

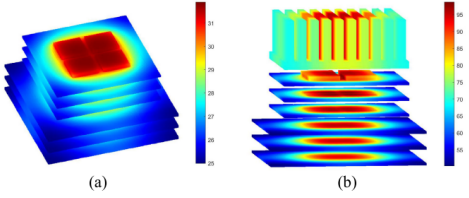


Fig 20. The thermal field simulation of 3D IC

Conclusion

Challenges and Opportunities

Challenges:

- The existing tools are still immature and lack systematic support.
- The academic tools require more customization for 2.5D architecture to improve accuracy and efficiency.
- Emerging technologies and architectures are advancing rapidly, while EDA tools are being left behind.

Opportunities:

- Utilizing machine learning algorithms to optimize the design workflow.
- There is a shortage of point-tools, which provides many startup opportunities and will bring commercial benefits.
- 2.5D architectures will demonstrate even greater potential in high-performance computing with efficient EDA tools.

THANK YOU!