ART-3D: Analytical 3D Placement with Reinforced Parameter Tuning for 3D ICs.

Gauthaman Murali[1], Sandra Maria Shaji [1], Anthony Agnesina [1], Guojie Luo [2], Sung Kyu Lim[1]

Georgia Institute of Technology[1], Atlanta, USA

Peking University[2], China

Funded by: SRC Semiconductor Research Corporation, DARPA, NRF National Research Foundation of Korea

3D Physical Design Methodology

- 3D ICs offer considerable wirelength and performance, power, and area (PPA) improvements over 2D ICs
- Several pseudo 3D flows have been proposed for physical design of 3D ICs
  - State-of-the-art 3D flows: Shrunk2D[1], Compact2D + Pin3D[2], and Snap3D[3]
- These approaches pose 3D placement as a 2D problem
  - An initial 2D placement is partitioned to obtain 3D placement -> leads to degradation of 3D nets
  - 3D degradation compensated by using commercial routing and optimization tools
- What happens if we use true 3D placement approaches with the commercial routing and optimization tools?

[1] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Shrunk-2-D: A Physical Design Methodology to Build Commercial-Quality Monolithic 3-D ICs," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 36, no. 10, pp. 1716–1724, 2017.
[2] S. Surya, K. Pentapati, K. Chang, V. Gerousis, R. Sengupta, and S. Lim, "Pin-3D: A Physical Synthesis and Post-Layout Optimization Flow for Heterogeneous Monolithic 3D ICs," in 2020 IEEE/ACM International Conference On Computer Aided Design, 2020.
[3] P. Vanna-Iampikul, C. Shao, Y.-C. Lu, S. Pentapati, and S. K. Lim, "Snap-3D: A Constrained Placement-Driven Physical Design Methodology for Face-to-Face-Bonded 3D ICs," in ACM International Symposium on Physical Design, ser. ISPD '21, 2021, p. 39–46.

Academic 3D placement works

- Force-directed 3D placer
  - Force3D from GTCAD[1]
- Analytical 3D placers
  - Non-linear 3D (NL3D) placer from UCLA[2] -> used in our work
  - Electrostatics-based placer (ePlace3D) from UCSD[3]
  - TSV-aware 3D placer from NTU[4]

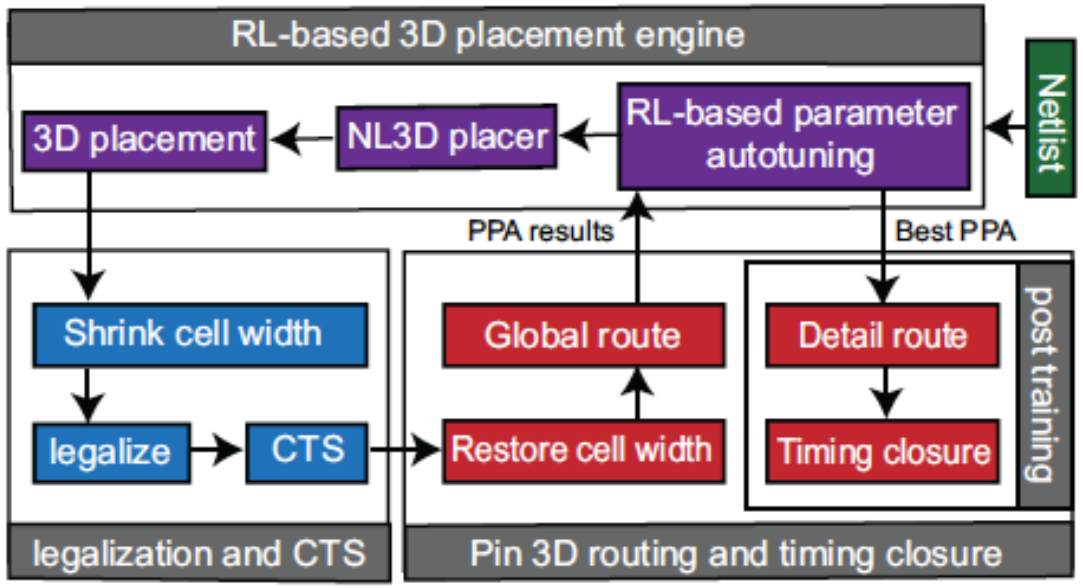
[1] D. H. Kim, K. Athikulwongse, and S. K. Lim, "Study of Through-Silicon-Via Impact on the 3-D Stacked IC Layout," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 21, no. 5, pp. 862–874, 2013.
[2] G. Luo, Y. Shi, and J. Cong, "An Analytical Placement Framework for 3-D ICs and Its Extension on Thermal Awareness," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 32, no. 4, pp. 510–523, 2013.
[3] J. Lu, H. Zhuang, I. Kang, P. Chen, and C.-K. Cheng, "EPlace-3D: Electrostatics Based Placement for 3D-ICs," in Proceedings of the 2016 on International Symposium on Physical Design, ser. ISPD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 11–18.
[4] M. Hsu, Y. Chang, and V. Balabanov, "TSV-aware analytical placement for 3D IC designs," in 2011 48th ACM/EDAC/IEEE Design Automation Conference (DAC), 2011, pp. 664–669.

State-of-the-art 3D flows vs Our work

	Shrunk2D	Compact2D + Pin3D	Snap3D	Our work
Key idea	Cell and wire shrinking	Placement compaction and pin projection	Tier-row snapping	**True 3D placement and cell width shrinking**
Die stacking	Separate dies	3D metal stack	3D metal stack	3D metal stack
Strength	First pseudo 3D flow	Better buffering/sizing	Better power	Best performance, PDP, and EDP
Weakness	Shrinking causes DRC issues	Die-by-die legalization & optimization	Modified pin locations	-
Placement	Commercial 2D + tier partitioning	Commercial 2D + tier partitioning	Tier partitioning + commercial 2D	**True 3D placer tuned with RL**
Legalization	Die-by-die	Die-by-die	Both tiers together	Both tiers together
Signal routing	Die-by-die	3D	3D	3D
Clock tree synthesis	Commercial 2D + tier partitioning	Commercial 2D + tier partitioning	3D based on commercial 2D	3D based on commercial 2D
Post routing optimization	Not supported	Enhanced die-by-die	Not performed	Enhanced die-by-die

ART-3D: Analytical 3D Placement with Reinforced Parameter Tuning for 3D ICs.

- Three training stages
  - **RL-based 3D placement**: Placement parameters of NL3D optimized using RL
  - Commercial tool-based placement optimization and CTS
  - **Pin3D**: Commercial tool-based 3D global routing
- Post-training
  - Detail routing and post-route optimization on best result



Setting up our 3D P&R flow

- We perform 2 tier 3D designs
  - Commercial 2D P&R tools have limitation on number of metal layers that can be used
- True 3D placement engine supports bookshelf format netlists
  - Standard netlists are in Verilog format
  - Verilog -> Design exchange format (DEF) + Library exchange format (LEF) -> Bookshelf
- 3D tech files for commercial router
  - 3D LEF files: Contain physical information of 3D routing stack
  - 3D LIB files: Contain timing information of standard cells belonging to both tiers

True 3D Placement

- **Placement engine**: An analytical 3D placement framework[1] developed by Luo *et al.*, UCLA
- 3D placement objective function:

$$minimize\ OBJ(x_i, y_i, z_i) = \sum_{e \in \mathrm{E}} (1 + \gamma_e)\big(WL(e) + \alpha_{MIV}.MIV(e)\big), \ -(1)$$

where, x_i, y_i, z_i are the x, y, and tier locations of a cell i, E is the set of nets connecting cell i, and WL(e) is the half perimeter wirelength (HPWL), γ_e is the wire weight, α_{MIV} is the via weight, and MIV(e) is the number of vias on net e **ϵ** E.

- Solved using non-linear programming (NLP) – hence, we call it **non-linear 3D** placer
- **Issue:** The placement quality highly depends on how the placer parameters are configured

[1] G. Luo, Y. Shi, and J. Cong, "An Analytical Placement Framework for 3-D ICs and Its Extension on Thermal Awareness," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 32, no. 4, pp. 510–523, 2013

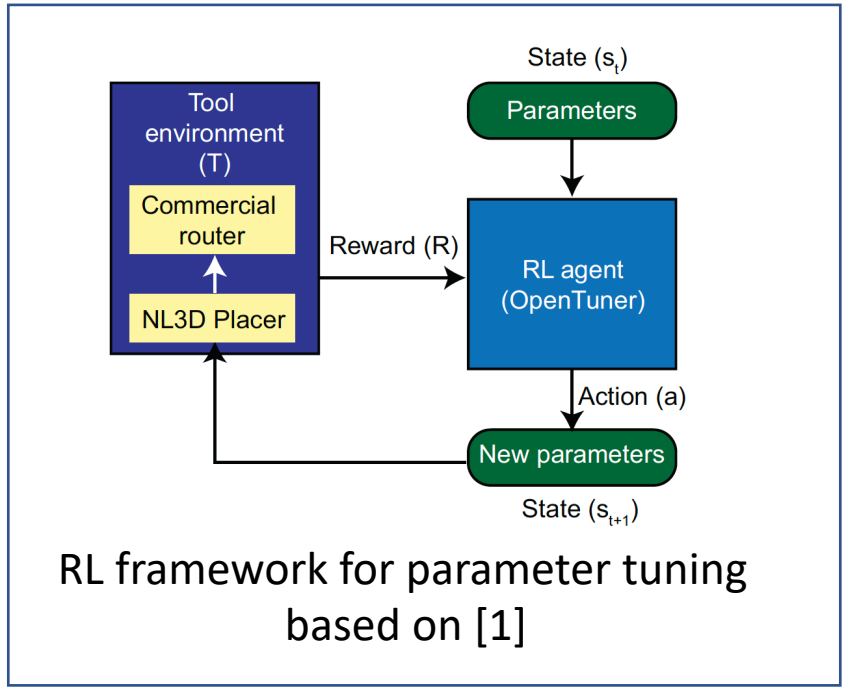
Placement Parameters

Parameter	Description	Type	Value
Grids	Placement bins in X & Y directions	Integer	[50,150]
Clustering depth	Levels of clusters during initial placement	Integer	[1,5]
Cluster ratio	Ratio of clusters at current and previous clustering level	Float	[0.1,0.3]
Wire weight (γ_e)	Weight of HPWL in placer objective function	Float	[1,100]
Via weight (α_{MIV})	Weight of #MIV in placer objective function	Float	[1,10]
Detail place overlap ratio	Max. cell overlap % allowed to terminate detail placement	Float	[0.1,0.25]
Target density	Max. density % allowed in each placement bin	Float	[0.5, 1.0]
ε	Helmholtz bin density smoothing parameter	Float	[0.5, 2.5]
ϒ	Parameter to accurately depict WL using LogSum exp in NLP	Float	[0.5, 2.5]
Congestion-driven placement	Enable congestion driven placement	Boolean	[T,F]
Add dummy cells	Add dummy blocks to prevent routing congestion	Boolean	[T,F]

Placement Parameter Tuning using Reinforcement Learning

- **State**: Set of placement parameters for a given netlist
- **RL agent**: OpenTuner based RL agent. Tuned placement parameters intelligently to obtain better PPA
- **Tool environment**: Parameter tuned true 3D placer + Commercial global router. Used to evaluate the quality of a given parameter configuration
- **Reward**: $\frac{1}{4\tanh(WL)} + \frac{1}{\tanh(|WNS|)} + \frac{1}{\tanh(P)}$,

where, WL is wirelength in m, WNS is the design's worst negative slack in nanoseconds, and P is the total design power in watts.



RL framework for parameter tuning based on [1]

[1] A. Agnesina, K. Chang, and S. K. Lim, "VLSI Placement Parameter Optimization using Deep Reinforcement Learning," in 2020 IEEE/ACM International Conference On Computer Aided Design (ICCAD), 2020, pp. 1–9.

RL Agent for Parameter Autotuning

- Parameters tuned using multi-armed bandit (MAB) scheme using a set of search techniques (T).

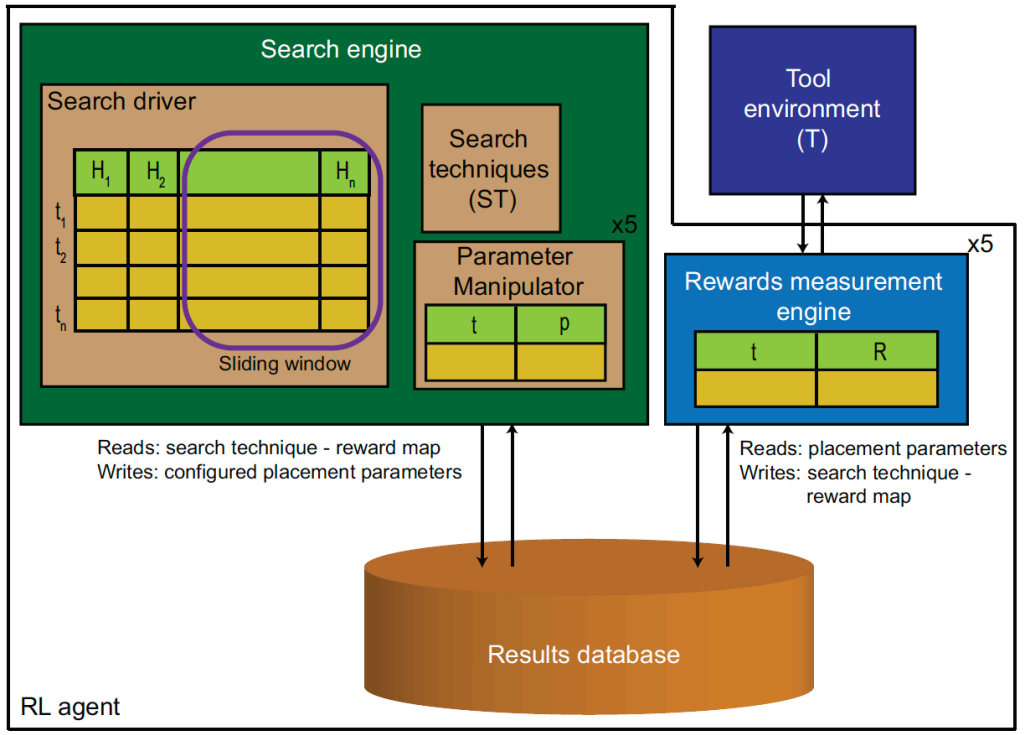
Goal of RL Framework

For a given netlist, find a parameter setting $\mathbf{p} \in \rho$ such that the reward $\mathbf{R}$ is maximized, where ρ is the set of all parameter combinations.

- MAB scheme called the area under the curve credit assignment (AUC) bandit meta technique is used in our work
- 20 parameter search iterations run during training (5 P&R runs in each iteration)
- Search technique chosen based on equation:

$$t = \arg\max\left(ST\left(AUC_t + C\sqrt{\frac{|H|}{H_t}}\right)\right),$$

where, |H| is the length of the sliding history window, AUC_t is the credit assignment term quantifying the performance of the technique t in the sliding window, and H_t is the number of times the technique t is used in the sliding window.



RL agent for parameter tuning

Training the RL agent

- We train our RL agent using
  - cell-dominant benchmarks (AES_128, Netcard),
  - net-dominant benchmarks (LDPC, ECG),
  - hard-macro based benchmarks (RocketCore, Cortex A7).
- **Trained parameters**:
  - AES_128: Focus on cell spreading to avoid placement and routing congestion
  - LDPC: Focus on optimizing wirelength by avoiding routing congestion
  - RocketCore: Focus on avoiding placement and routing congestion around large hard-macros

Trained parameters	28nm AES_128	28nm LDPC	28nm RocketCore
Grids	144	76	99
Clustering depth	1	4	2
Cluster ratio	0.29	0.15	0.25
γ_e	96.77	97.17	80.71
α_{MIV}	4.75	4.55	6.47
Overlap ratio	0.2	0.1	0.25
Target density	0.55	0.63	0.97
ε	1.93	2.29	1.64
ϒ	1.5	1.78	0.83
Congestion-driven	0	1	0
Add dummy cells	1	1	0

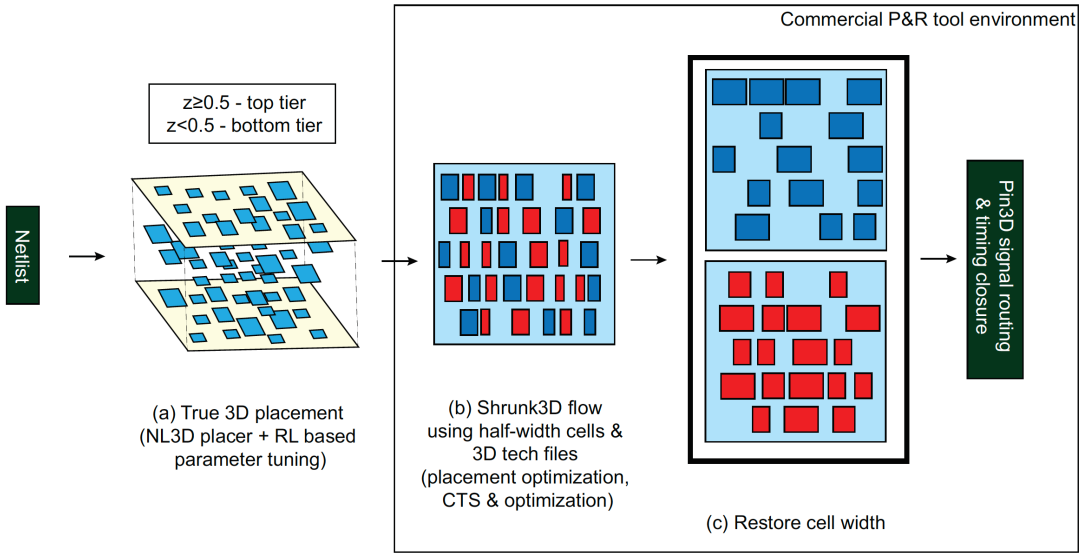
Making RL tuning realistic

Metric	Full P&R	Partial P&R (preCTS)	Partial P&R (globalRoute)
Runtime (mins)	2405	829	1432
WNS (ps)	-30	-78	-45
Power (mW)	182.1	171.8	176.6
PDP (pJ)	126.8	127.8	125.2

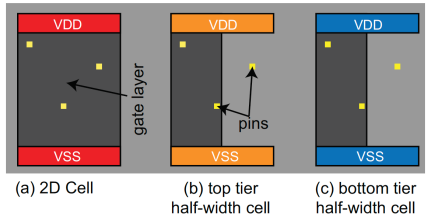
- We tried performing RL tuning on 28nm LDPC using PPA results obtained from different stages of P&R flow
  - Full P&R
  - Partial P&R up to pre-CTS
  - Partial P&R up to global route
- Tuning based on the results obtained from global routed design offers a good trade-off between runtime and PPA

Commercial Router optimized for ART-3D

- Commercial 2D P&R tool used to perform optimization and routing
- 2D tools do not support placement density > 100%
- Cell width halved to place all cells
- 3D LEF (physical information) and LIB (timing information) files used to perform placement optimization and CTS
- Expand cells, perform final routing and timing closure using Pin3D[1] router.



Commercial Router

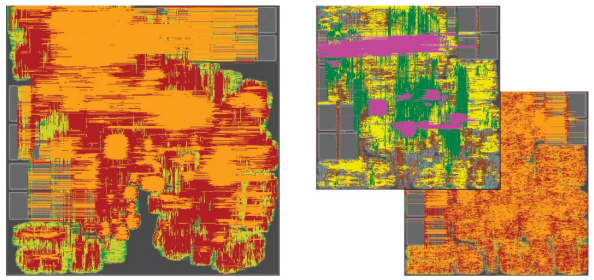


Half-width cells

[1] S. Surya, K. Pentapati, K. Chang, V. Gerousis, R. Sengupta, and S. Lim, “Pin-3D: A Physical Synthesis and Post-Layout Optimization Flow for Heterogeneous Monolithic 3D ICs,” in 2020 IEEE/ACM International Conference On Computer Aided Design, 2020.

Analyzing our 3D flow

- Some of the circuits used for **training**: LDPC, RocketCore, CA7
- Circuits used for **inferencing**: VGA_LCD, CortexA53
- We use a commercial 28nm PDK with 70nm MIVs and 16nm PDK with 40nm MIVs. We compare our flow against 2D, 3D run with manually tuned NL-3D placer and two state-of-the-art 3D flows: Pin3D and Snap3D
- We perform following analyses to evaluate our 3D flow
  - PPA analysis
  - Critical path analysis
  - Clock metric analysis
  - Memory net analysis



(a) 2D (b) Our work

Final 28nm GDSII layouts of CortexA53

ART-3D vs State-of-the-art 3D flows

	28nm Benchmarks (Training)					
	LDPC [1]			RocketCore [2]		
	2D	Pin3D	A3D	2D	Snap3D	A3D
Target Freq. (GHz)	1.5			1.0		
Footprint (mm^2)	0.11	0.05	0.05	0.31	0.15	0.15
Total WL (m)	2.29	1.43	1.45	1.87	1.46	1.50
Eff. Freq. (GHz)	1.43	1.36	1.41	0.95	0.93	**0.97**
Total Power (mW)	256.4	181.2	**176.6**	151.4	142.55	**142.3**
PDP (pJ)	179.5	133.2	**125.2**	158.9	154.0	**146.6**
Runtime (mins)			1332			2345

[1] S. Surya, K. Pentapati, K. Chang, V. Gerousis, R. Sengupta, and S. Lim, “Pin-3D: A Physical Synthesis and Post-Layout Optimization Flow for Heterogeneous Monolithic 3D ICs,” in 2020 IEEE/ACM International Conference On Computer Aided Design, 2020.
[2] P. Vanna-Iampikul, C. Shao, Y.-C. Lu, S. Pentapati, and S. K. Lim, “Snap-3D: A Constrained Placement-Driven Physical Design Methodology for Face-to-Face-Bonded 3D ICs,” in ACM International Symposium on Physical Design, ser. ISPD ’21, 2021, p. 39–46.

PPA Comparison ART-3D (A3D) vs 2D, 3D with manually tuned NL-3D placer (NL3D)

	16nm Benchmarks									28nm Benchmarks					
	Training			Inferencing						Training			Inferencing		
	LDPC			VGA_LCD			Cortex A53			Cortex A7			Cortex A53		
	2D	NL3D	A3D	2D	NL3D	A3D	2D	NL3D	A3D	2D	NL3D	A3D	2D	NL3D	A3D
Target Freq.	3 GHz			8 GHz			1 (no units)			1 (no units)			1 (no units)		
Footprint	1.00	0.50	0.50	1.00	0.50	0.50	1.00	0.50	0.50	1.00	0.50	0.50	1.00	0.50	0.50
Total WL	1.00	0.82	0.77	1.00	0.89	0.85	1.00	0.85	0.81	1.00	1.03	0.84	1.00	0.84	0.78
Eff. Freq.	1.00	0.95	1.00	1.00	0.92	1.00	1.00	0.95	**1.05**	1.00	0.96	**1.03**	1.00	1.13	**1.25**
Total Power	1.00	0.83	**0.79**	1.00	0.97	**0.94**	1.00	0.95	**0.94**	1.00	0.91	**0.90**	1.00	0.73	**0.72**
PDP	1.00	0.87	**0.79**	1.00	1.05	**0.95**	1.00	1.11	**0.90**	1.00	0.95	**0.87**	1.00	0.63	**0.57**
Runtime	1.00	1.19	4.89	1.00	2.54	2.00	1.00	3.12	2.90	1.00	1.02	12.69	1.00	0.59	0.53

*Results normalized w.r.t 2D

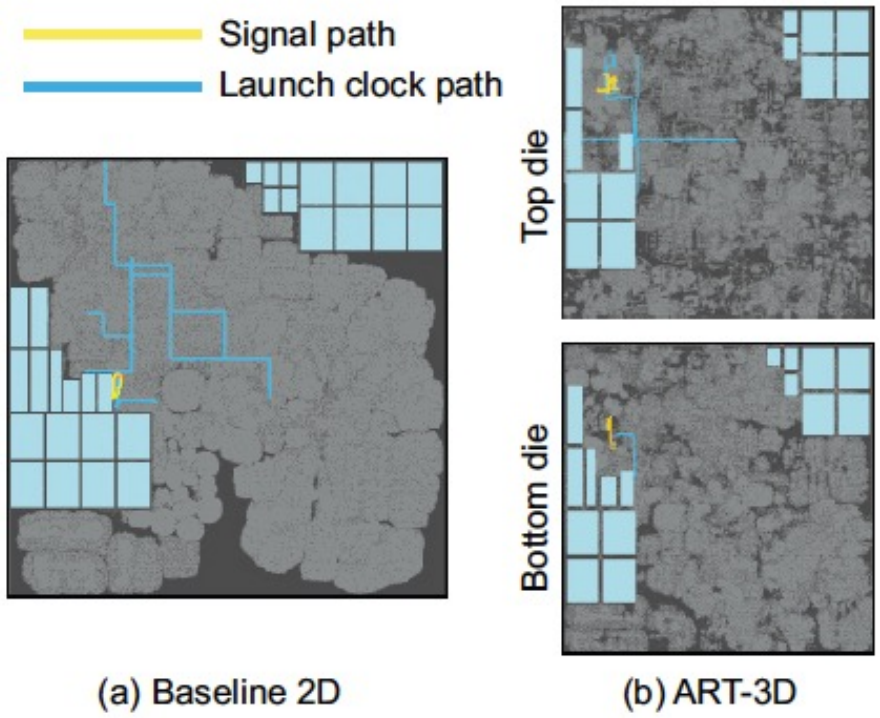
3D benefits: 28nm vs 16nm

- Gate capacitance dominates wire capacitance in 16nm
  - Multiple fins per standard cell
- 3D design helps reduction of wire capacitance
- So, PPA improvement in 16nm is less than that of 28nm
- However, this is expected to be avoided on advanced nodes below 5nm
  - #Fins per standard cell is less -> Low gate capacitance

	28nm RocketCore		16nm RocketCore	
Cap. Metrics (pF)	**2D**	**A3D**	**2D**	**A3D**
Gate Capacitance	**239**	230	**173**	142
Wire Capacitance	**265**	221	**152**	141

Critical Path analysis of 28nm Cortex A53

	CortexA53*	
Metrics	2D	Our work
Slack	-1.00	-0.13
Critical path delay	1.00	**0.90**
(+) Launch clock delay	1.00	1.30
Signal arrival time	1.00	0.98
Capture clock delay	1.00	**1.62**
(+) Clock period	1.00	1.00
(-) Setup time	1.00	0.26
Required arrival time	1.00	1.17



*Results normalized w.r.t 2D

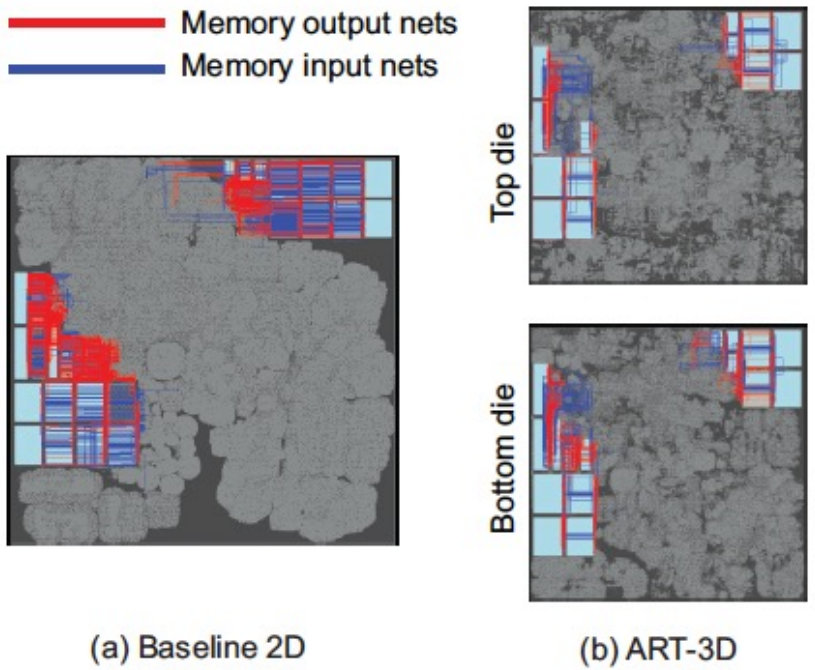
Clock Metric analysis of 28nm Cortex A53

Clock Metrics	2D	Our work
Frequency	1.0	
Clock latency	1.0	0.87
Clock skew	1.0	1.06
Clock WL	1.0	1.04
#Clock buffers	1.0	2.39

Memory net analysis of 28nm Cortex A53

Memory metrics	2D	Our work
Max. memory latency	1.00	**0.75**
Avg. memory latency	1.00	0.53
Max. mem path length	1.00	**0.70**
Avg. mem path length	1.00	0.66
Mem. access power	1.00	0.67

*Results normalized w.r.t 2D



Summary

- Our works consists of a 3D P&R flow with true 3D placement
- The placement quality is improved by tuning the placement parameters using RL
- The effective frequency of circuits designed using our flow is higher than state-of-the-art 3D flows.
- ART-3D also offers better power than 2D and state-of-the-art 3D flows
- These result in highly optimized PDP