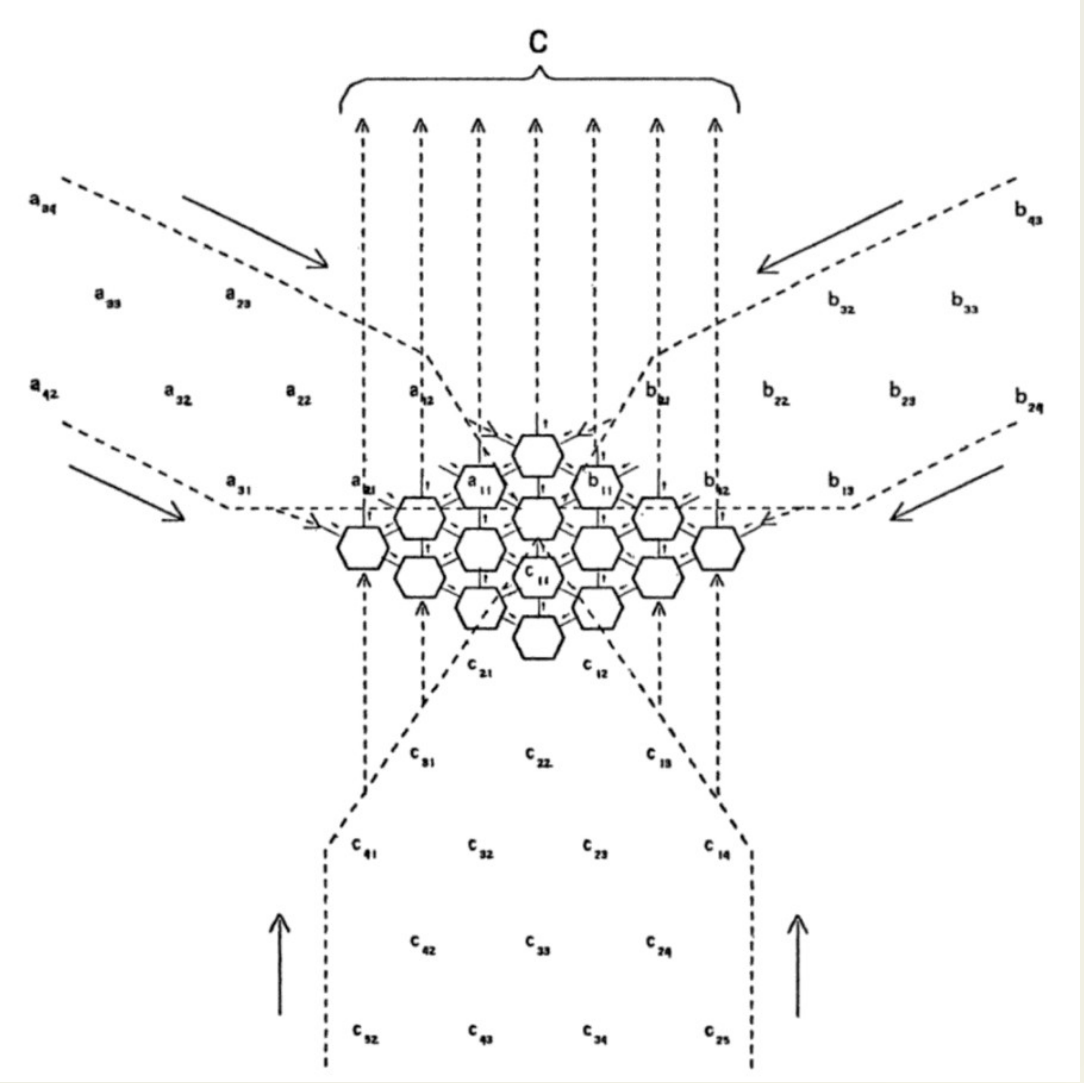
Physical Design for Systolic Array-Based Integrated Circuits

Jiang Hu
Dept of Electrical and Computer Engineering
Texas A&M University

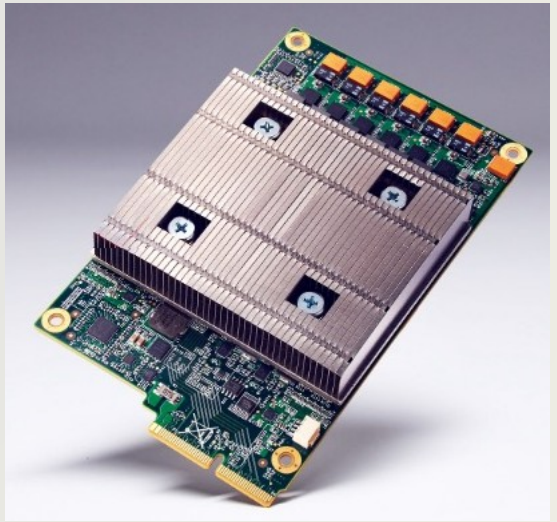
Systolic Arrays

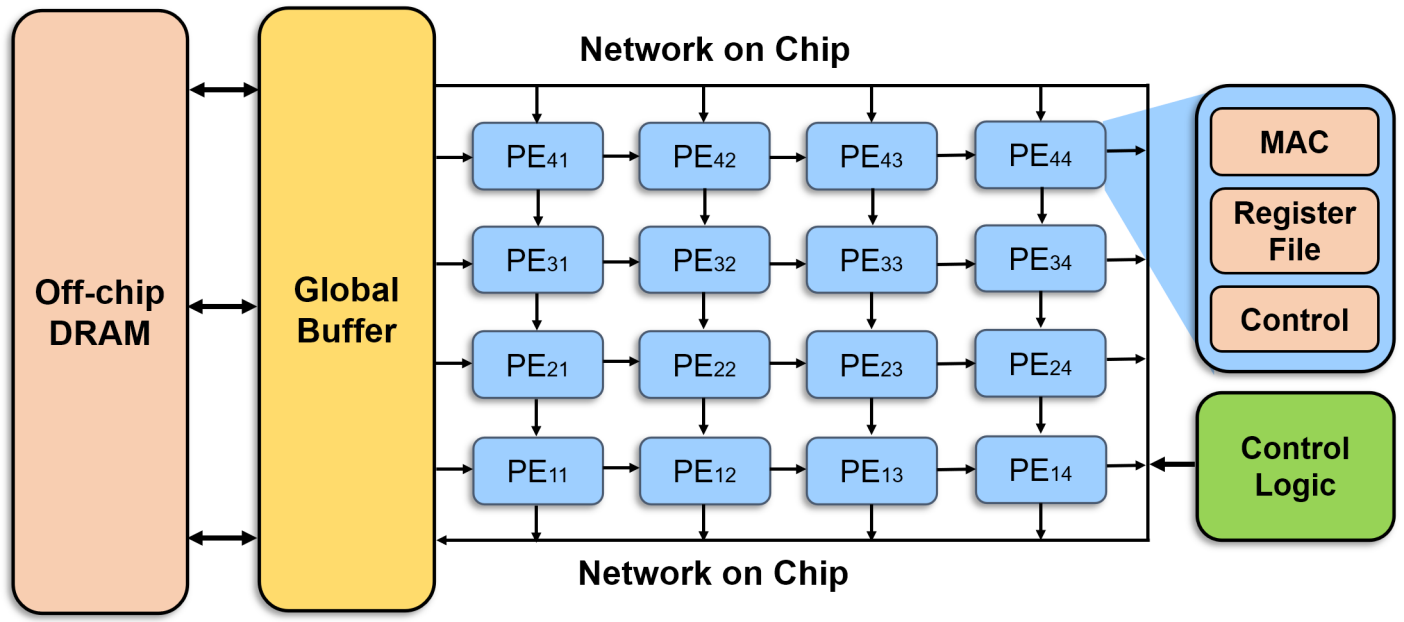
- Kung and Leiserson in late 1970s’
- Regular array of Processing Elements (PEs)
- No global interconnect



Systolic Array Applications

- Signal processing, linear algebra
- ML/AI accelerators: Eyeriss, TPU





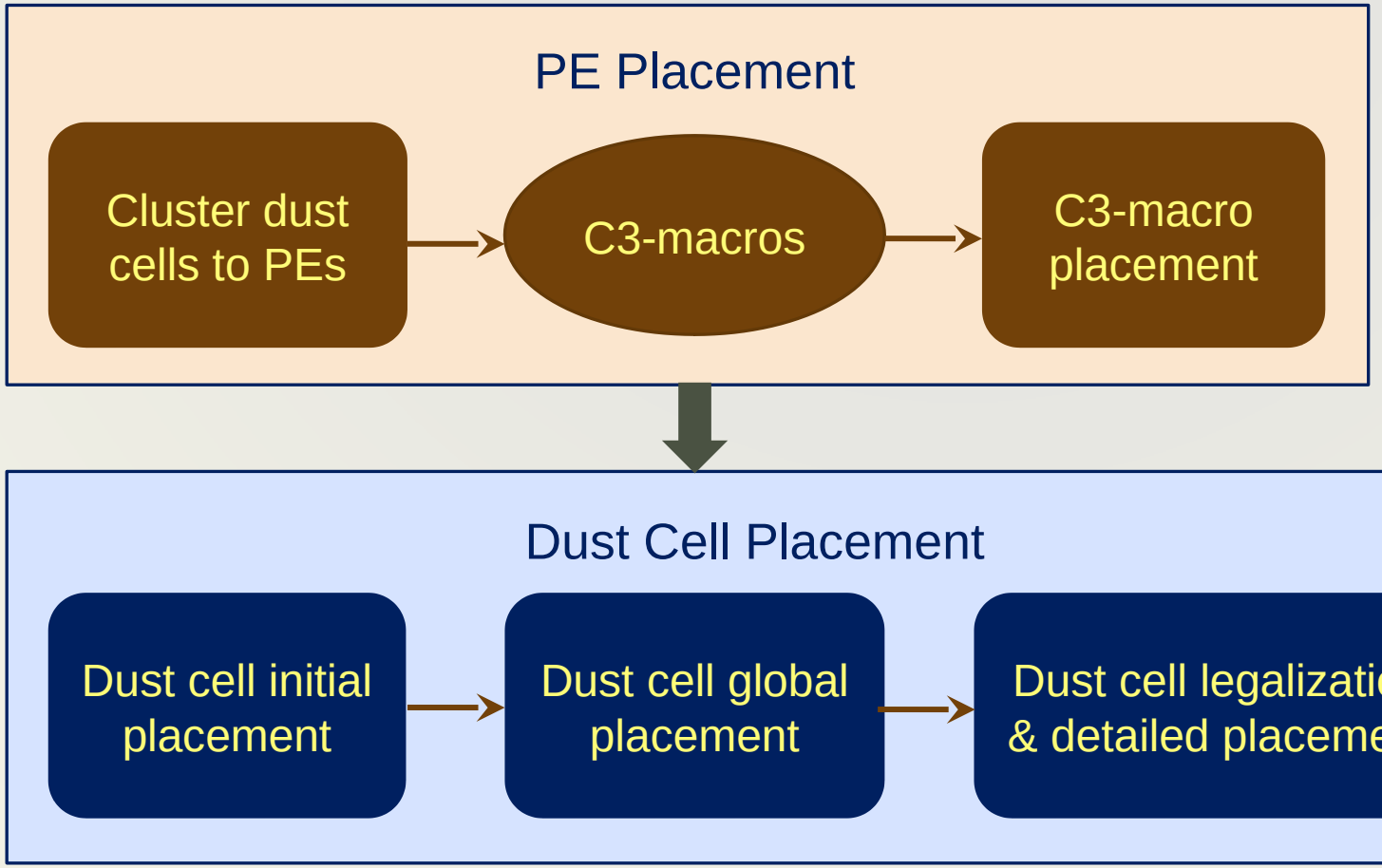
ML/AI Accelerator Research

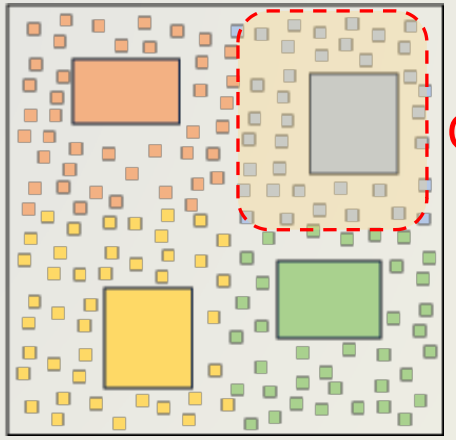
- Architecture design
- Workload mapping/scheduling
- Algorithm-hardware co-design
- High-level synthesis

- Physical design? Dedicated layout tools for ML/AI accelerators?

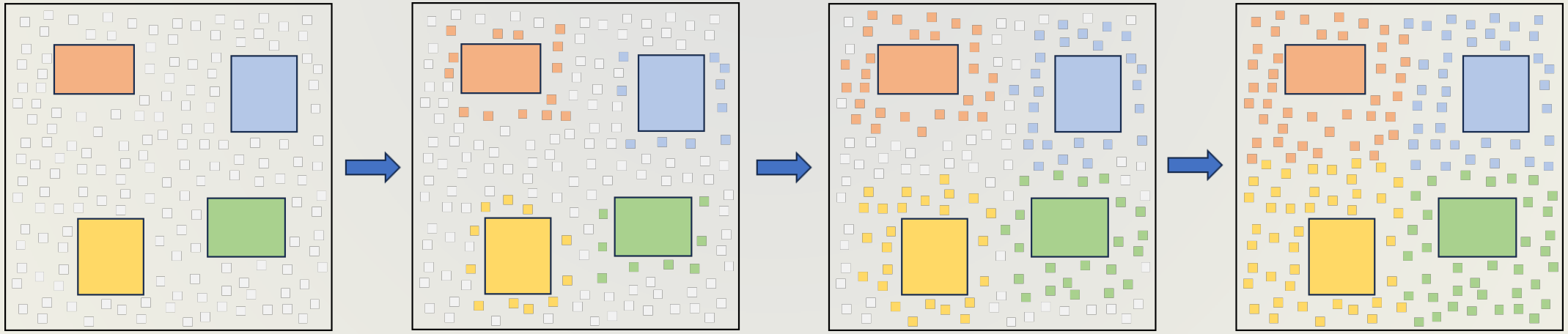
SysMix: Mixed Size Placement for Systolic Arrays

Place macros (PEs) + dust cells





Dust Cell Clustering



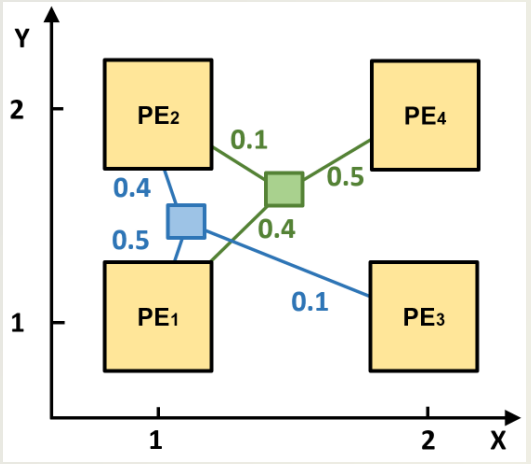
- Diffusion based Bubble-FOS/C algorithm [*H. Meyerhenke, et al.*, “Graph partitioning and disturbed diffusion,” *Parallel Computing* 2009]

C3-Macro Placement

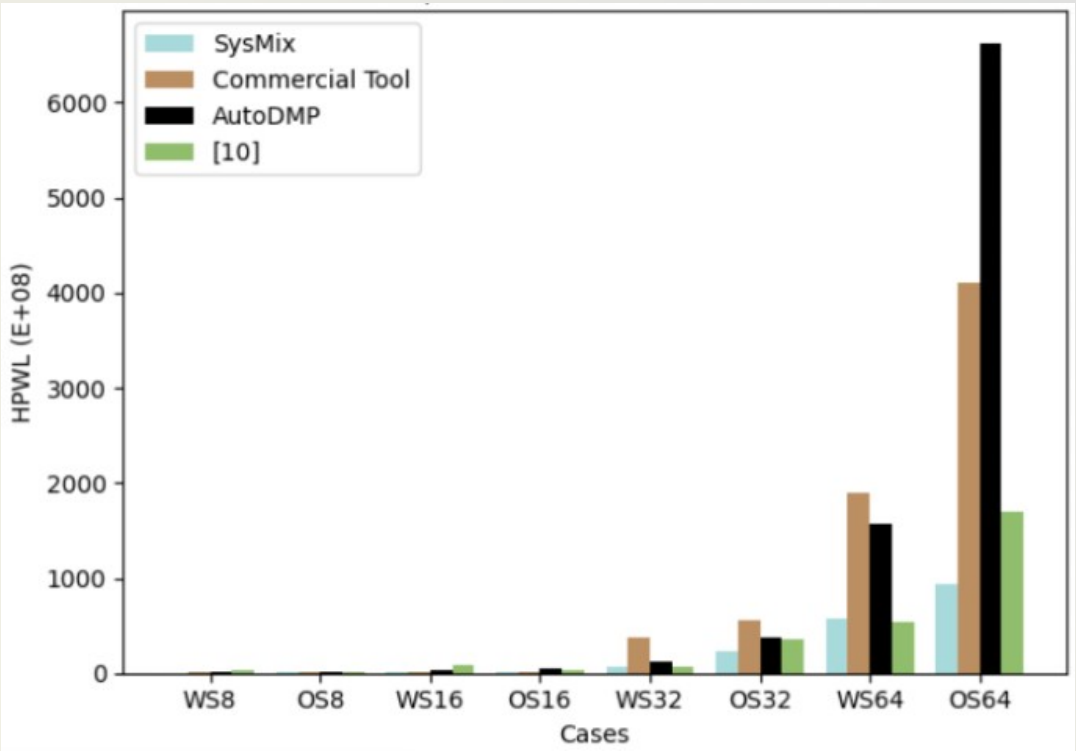
- C3-macro area == PE + clustered cells
- Linear programming to minimize HPWL
- C3-macros' 2D order follows the schematic
- Non-overlapping constraints
- Two iterations LP compact *x* and *y* directions separately

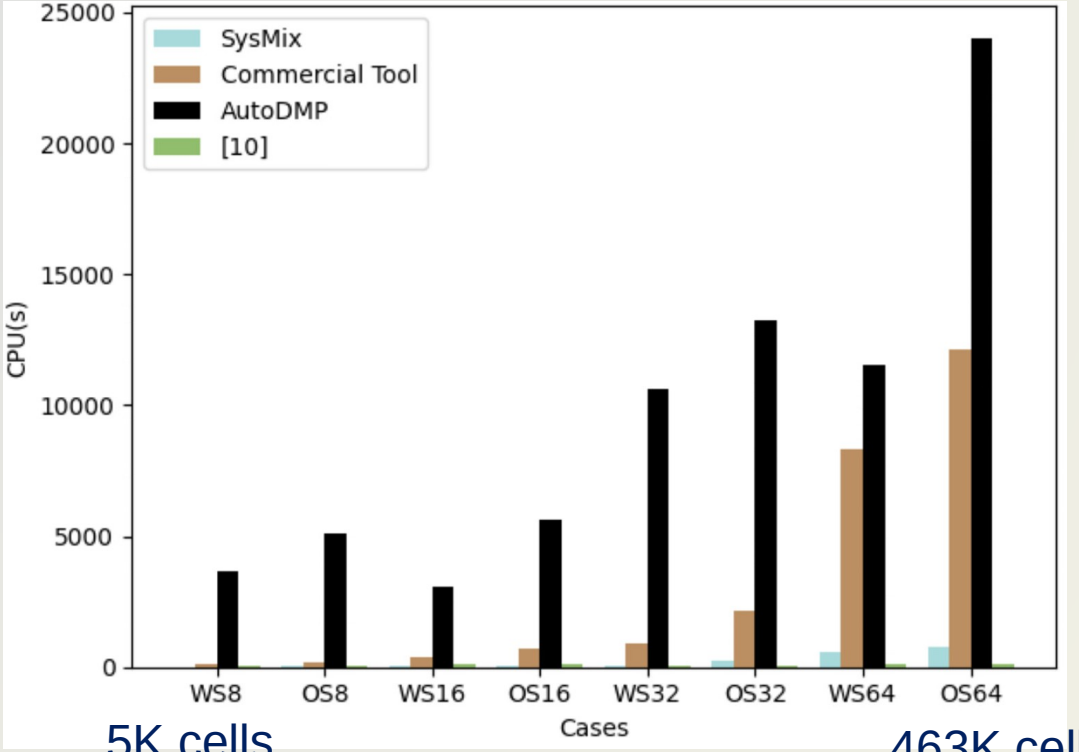
Dust Cell Placement

- PE locations are fixed
- Initial dust cell placement, equilibrium among forces toward associated PEs
- Global placement by RePlace
- Legalization and detailed placement by NTUPlace



Experimental Results





5K cells

463K cells

[10] *Y. Chen, Z. Wen, Y. Liang and Y. Lin*, "Stronger Mixed-Size Placement Backbone Considering Second-Order Information," *ICCAD* 2023

When PEs Are Also Dust Cells

Ours

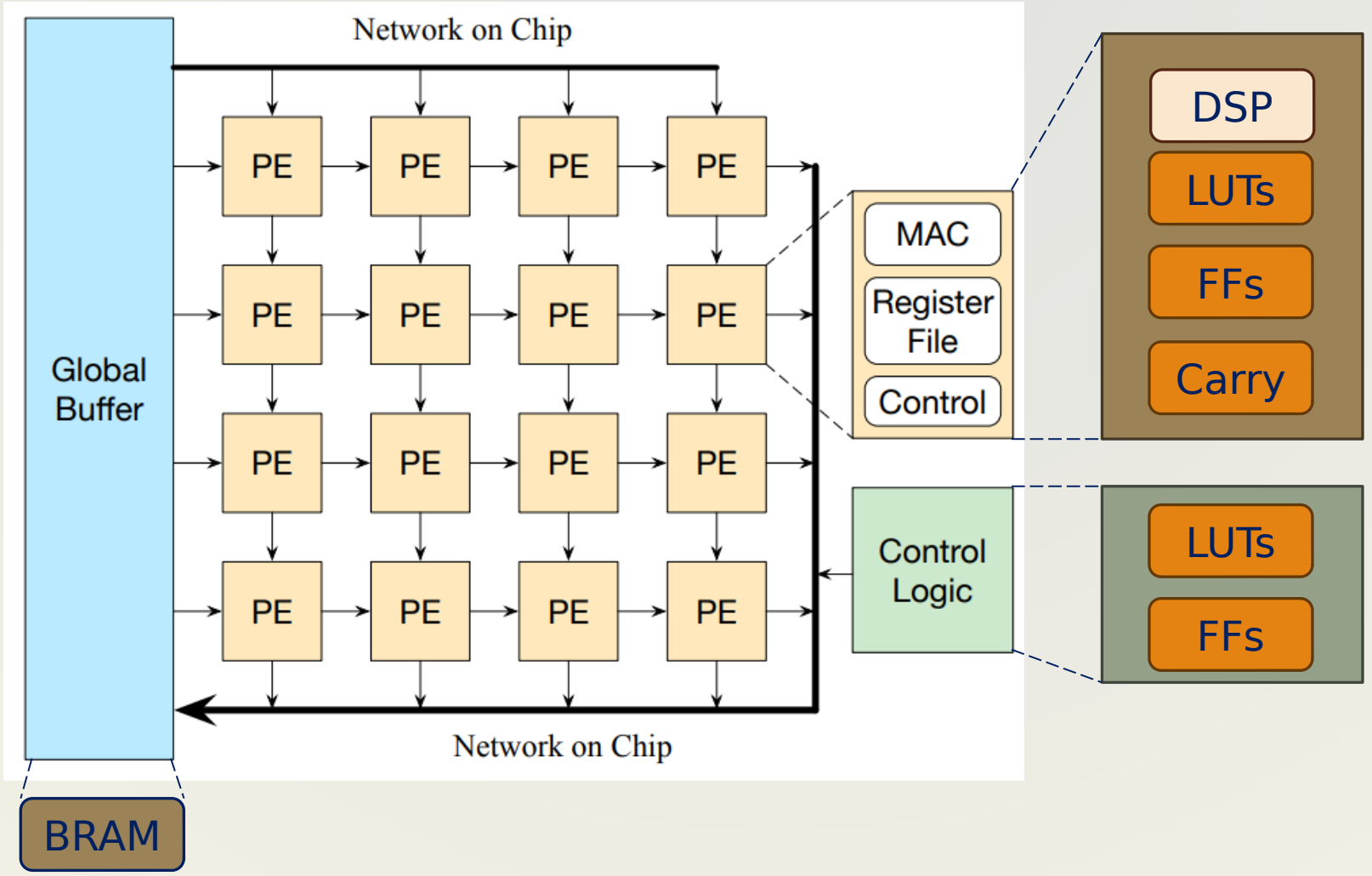
	PASOR		RePlAce		NTUplace3		POLAR		Ref [9]	
Testcases	GR WL	CPU	WL	CPU	WL	CPU	WL	CPU	WL	CPUP+GR
CNNWS16FP8-45 nm	9.3	11	9.6	12	9.8	13	9.9	13	9.8	12
CNNWS32FP8-45 nm	40.9	62	43.9	64	47.8	67	47.4	43	44.2	63
CNNWS64FP8-45 nm	167.8	215	191.3	221	236.2	226	211.8	506	193.8	237
CNNOS16FP8-45 nm	9.1	12	9.3	12	9.5	14	9.4	16	9.3	12
CNNOS32FP8-45 nm	41.4	58	44.6	63	47.5	66	45.1	38	43.9	60
CNNOS64FP8-45 nm	166.4	210	196.4	215	221.1	241	204.9	515	186.7	236
MM16FP8-45 nm	9.2	10	10.8	12	10.1	13	9.5	11	9.5	10
MM32FP8-45 nm	37.6	57	40.6	62	45.9	65	42.4	35	40.6	60
MM64FP8-45 nm	159.6	212	192.6	217	225.0	236	197.3	493	179.1	228
Norm average	1.00	1.00	1.11	1.06	1.18	1.16	1.14	1.11	1.08	1.06

T. Lin, C. Chu, J. R. Shinnerl, I. Bustany, and I. Nedelchev, "POLAR: Placement based on novel rough legalization and refinement." *ICCAD* 2013.

[9] *S. Chou, M.-K. Hsu, and Y.-W. Chang*, "Structure-aware placement for datapath-intensive circuit designs," *DAC* 2012.

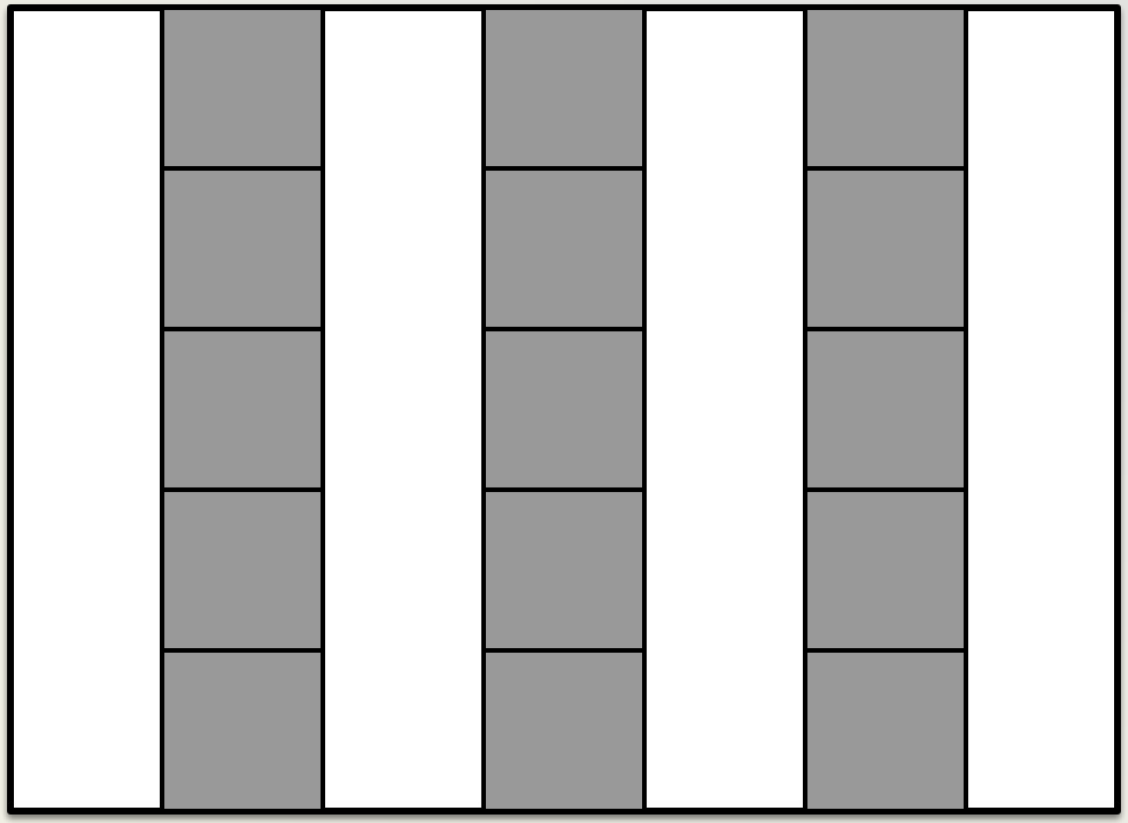
Systolic Array Placement on FPGAs

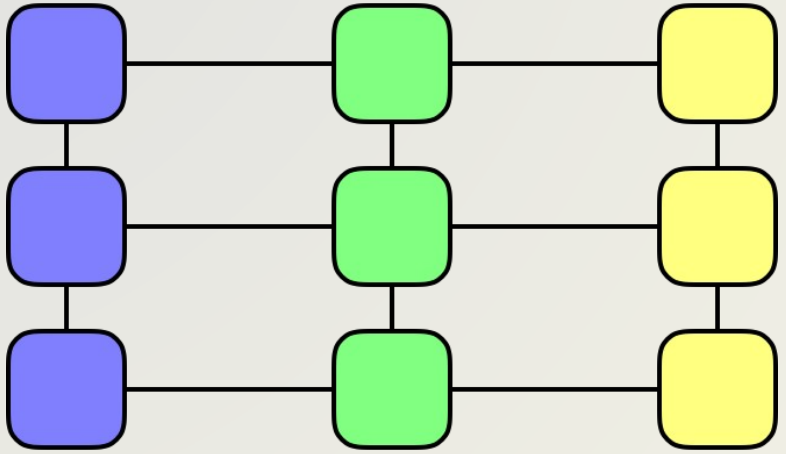
- 1 PE = 1 DSP + ~50 LUTs/FFs/Carry-chain



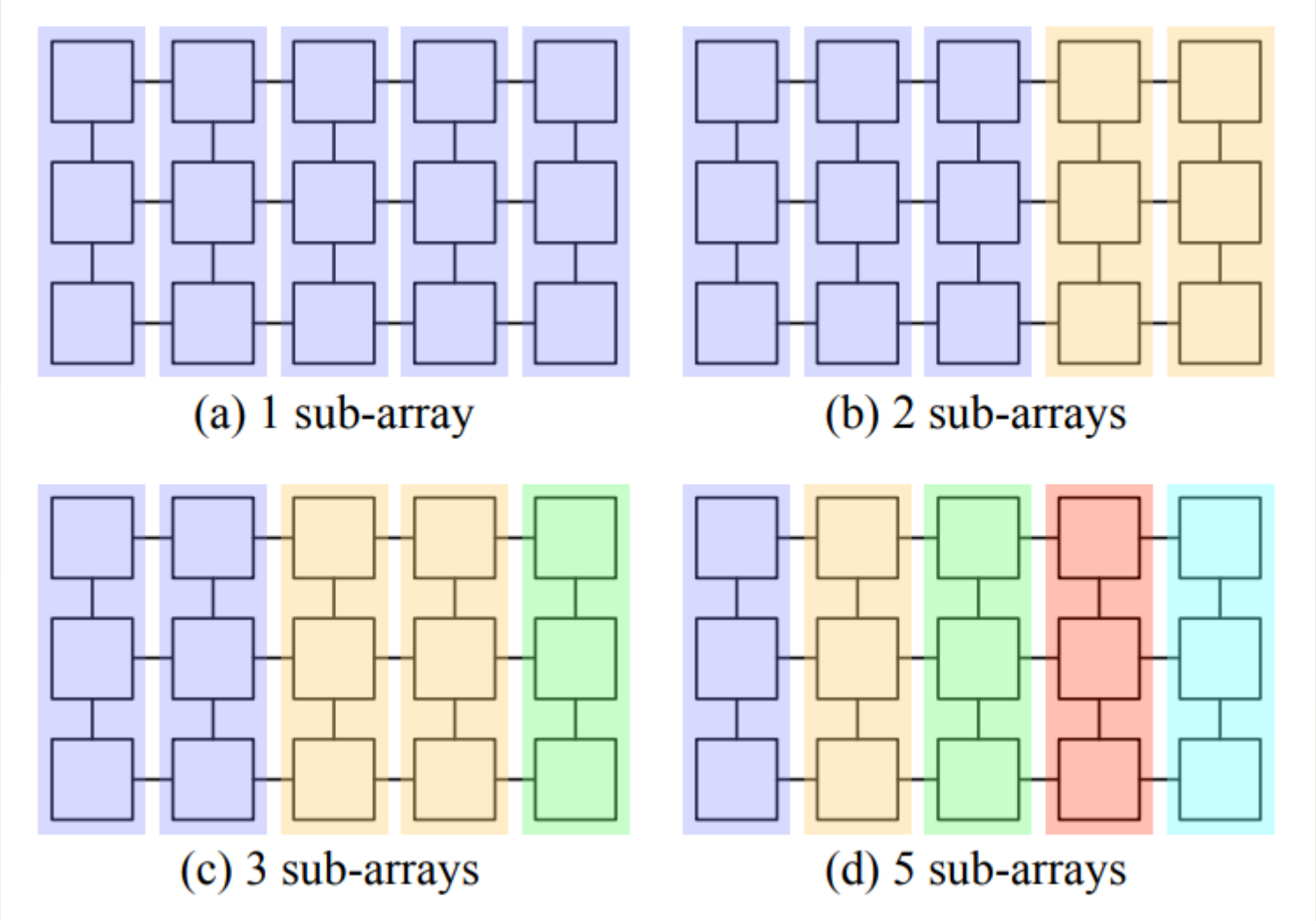
2D Macro Placement Problem

- Mapping a macro block array to DSP columns with minimum HPWL





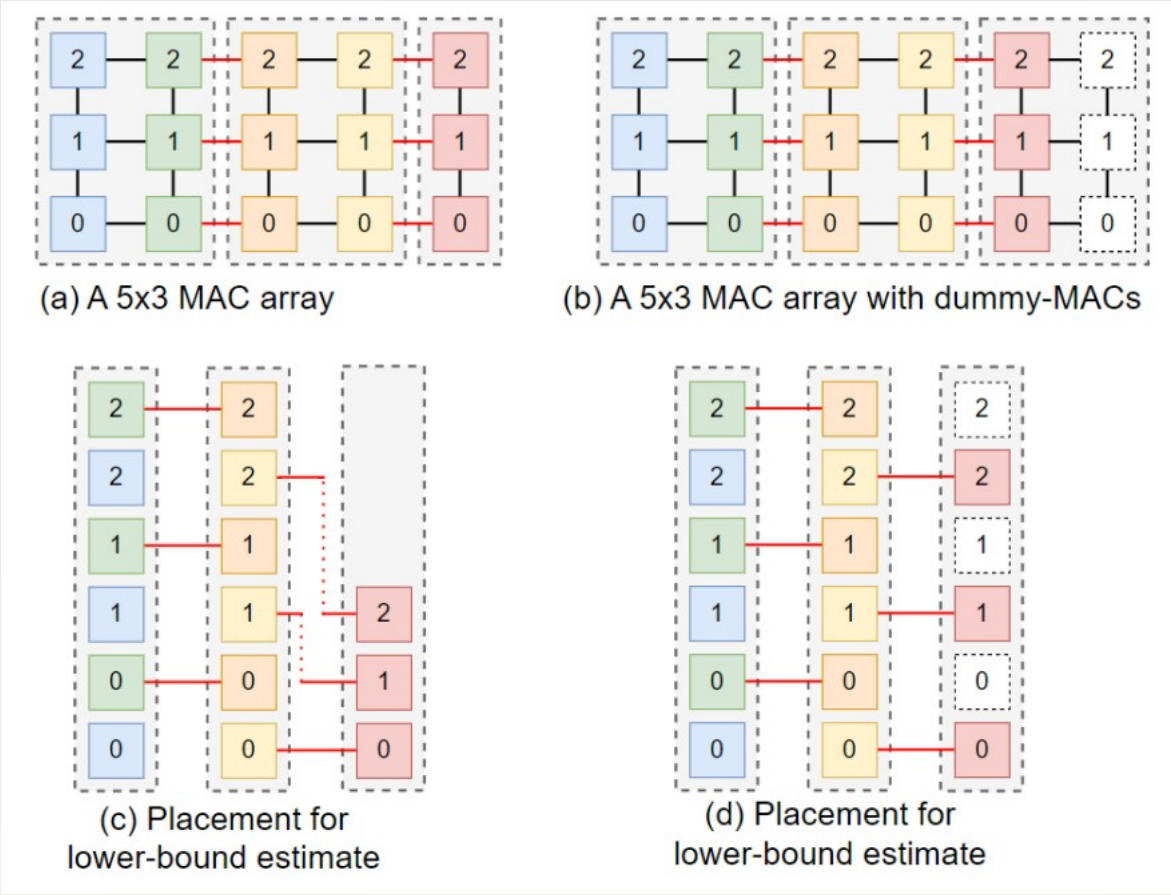
Step 1: Partition Candidate Generation



One sub-array corresponds to one DSP column

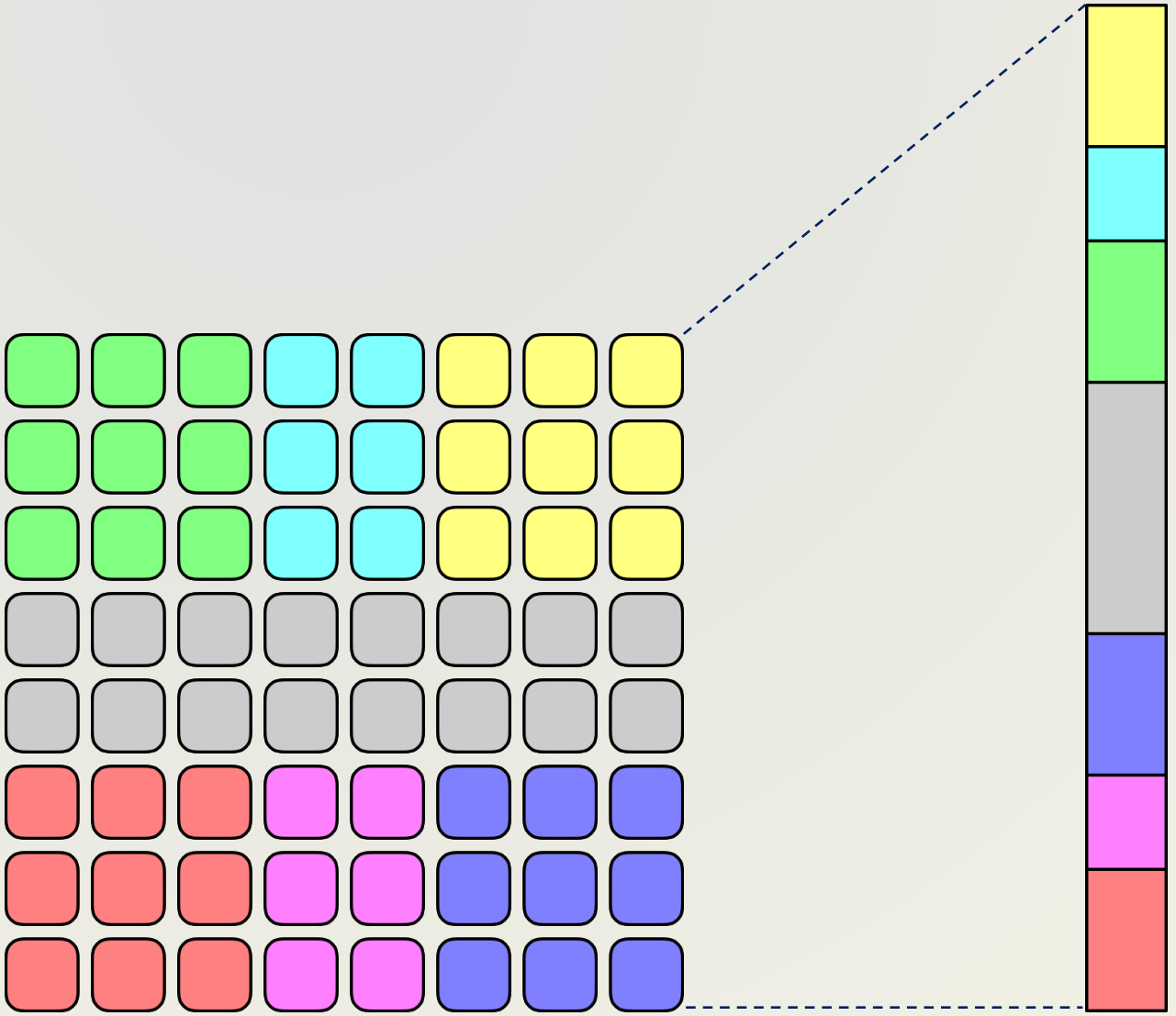
Step 2: Partition Candidate Pruning

- HPWL upper-bound and lower-bound estimated for each candidate
- Candidates are pruned based on the bounds

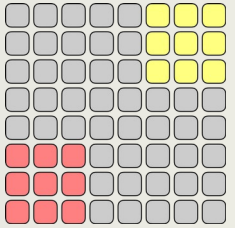


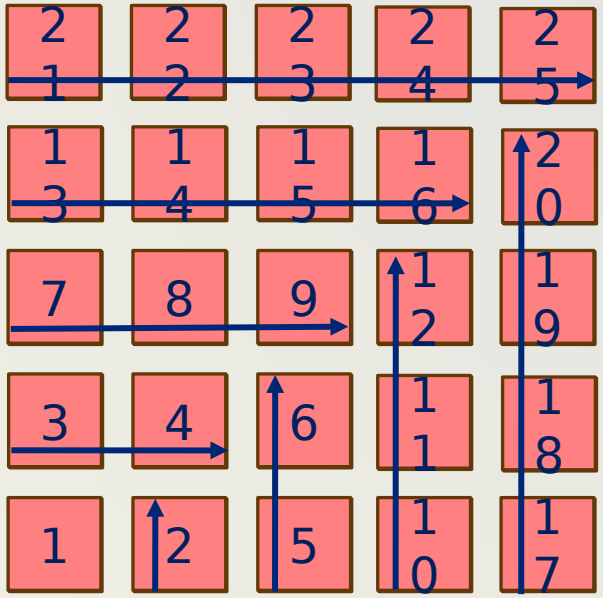
Step 3: Place Remaining Candidates

- Remaining candidates are placed by R-SAD (Region-wise Sweep in Alternating Directions)
- The one with min HPWL is selected as final solution

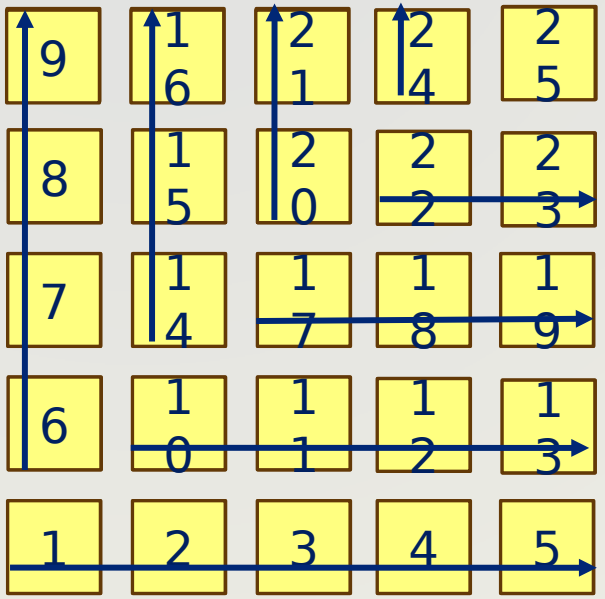


R-SAD: Lower-left and Upper-right Regions





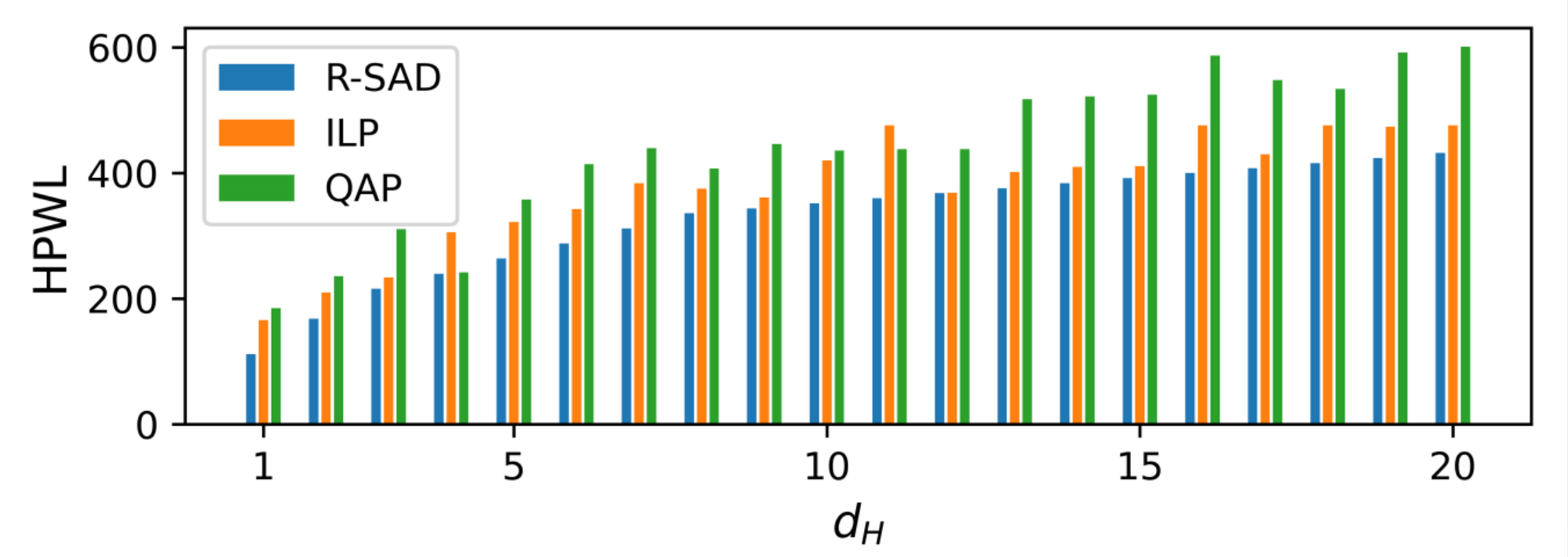
(a) Lower-left



(b) Upper-right

R-SAD vs ILP and QAP (Quadratic Assignment Problem)

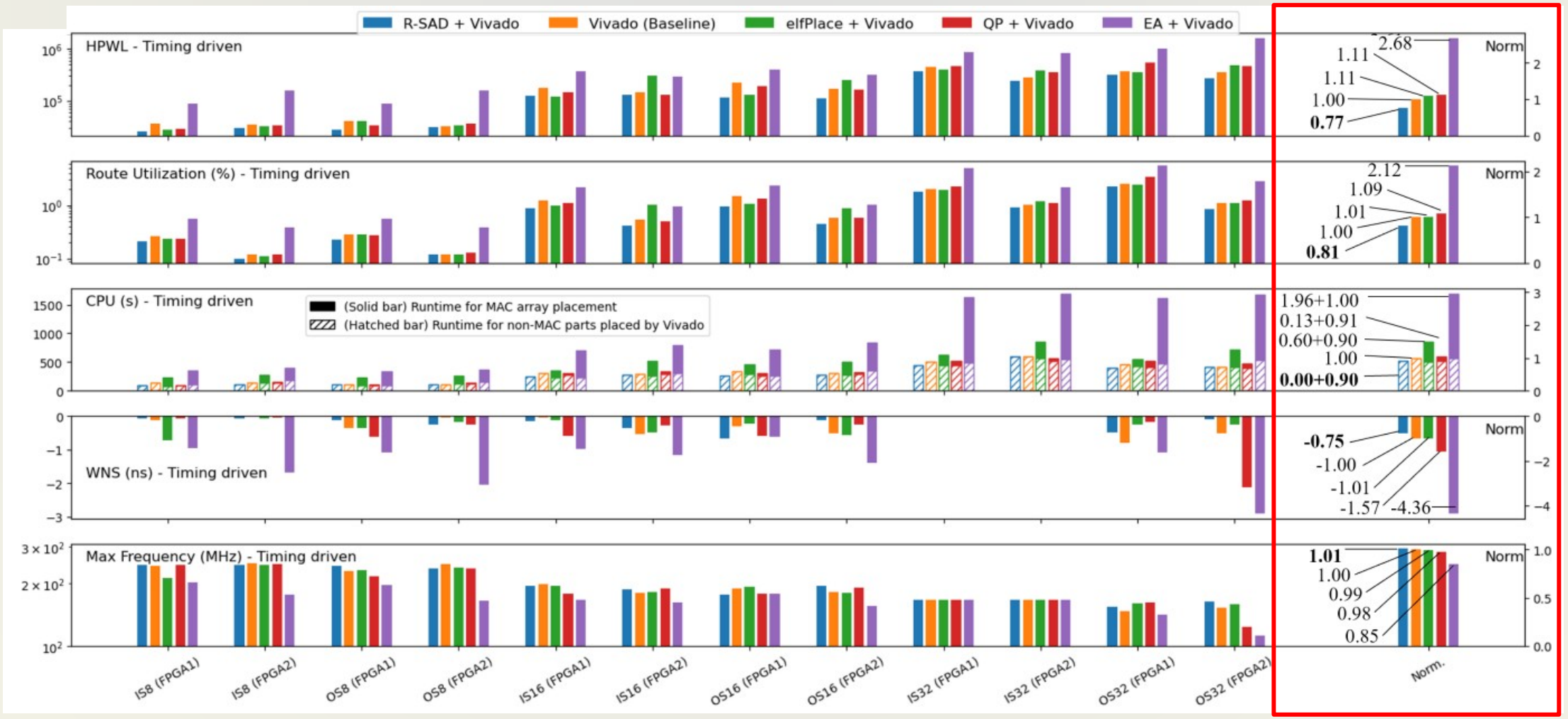
- HPWL of mapping an 8x8 macro array
- : horizontal distance between two adjacent DSP columns



Effectiveness of R-SAD in Different Placers

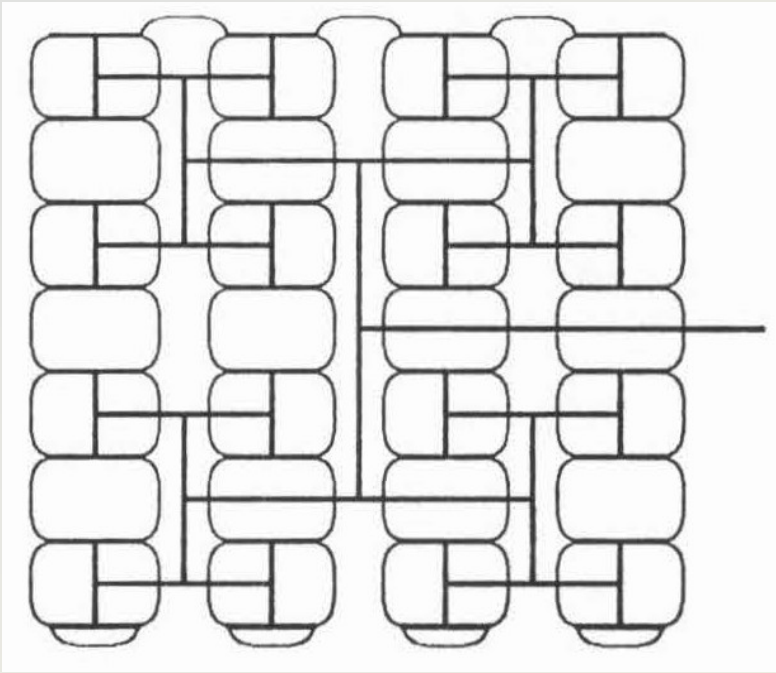
Device	Design	Vivado		R-SAD + Vivado		elfPlace		R-SAD + elfPlace		QP		R-SAD + QP	
		HPWL	CPU	HPWL	CPU	HPWL	CPU	HPWL	CPU	HPWL	CPU	HPWL	CPU
FPGA1	IS8	0.25	47	0.19	45	0.37	146	0.20	141	0.25	9	0.21	8
	OS8	0.25	61	0.21	44	0.29	129	0.28	115	0.30	11	0.25	11
	IS16	1.24	101	0.92	104	0.82	124	0.75	116	1.36	28	0.88	25
	OS16	1.19	110	0.90	102	0.91	171	0.81	113	1.71	39	0.97	38
	IS32	3.91	207	2.36	219	2.73	194	1.94	184	4.94	62	2.15	58
	OS32	3.43	199	2.76	190	2.73	128	2.14	123	4.61	99	3.48	89
	Norm.	1.00	1.00	0.71	0.97	1.00	1.00	0.78	0.89	1.00	1.00	0.60	0.92
FPGA2	IS8	0.26	59	0.23	76	0.21	142	0.20	237	0.25	9	0.21	12
	OS8	0.25	61	0.25	75	0.21	135	0.19	204	0.28	12	0.24	15
	IS16	1.08	120	0.90	102	2.71	255	0.69	182	1.21	42	0.85	37
	OS16	1.18	116	0.89	111	1.72	229	0.78	237	1.49	36	0.93	52
	IS32	2.22	216	1.66	202	3.26	297	1.89	286	3.51	57	2.43	60
	OS32	2.75	222	2.26	200	2.77	305	1.89	256	4.39	76	1.96	72
	Norm.	1.00	1.00	0.80	0.96	1.00	1.00	0.52	1.03	1.00	1.00	0.60	1.07

Comparison for Timing Driven Placement



Other Physical Design Problems for Systolic Arrays

- Clock network synthesis
- Routing
- Design prediction



Conclusions

- Systolic array designs are increasingly popular
- Physical design techniques customized for systolic arrays can significantly outperform general PD techniques

Acknowledgement

- Students: Donghao Fang, Hailiang Hu
- Collaborator: Wuxi Li, AMD
- Funding support: NSF, SRC

Thank You!
Questions?

MLCAD Symposium 2025

- Chaminade resort, Santa Cruz
- September 8-10
- Contest registration by April 20
  - Sponsored by Nvidia
- Paper abstract due May 16
- Industry track abstract due May 23
- Artifact badge