

Al-driven Memory Exploration in RTL Architect Efficient Physical-aware PPA Enhancement

Shu Rong Lee MediaTek



Agenda

- Motivation
- Reinforcement Learning Memory exploration Methodology (RLMM)
- RLMM Experimental Results
- Summary

3

Motivation – Physical Awareness

Considering physical implementation at the RTL design stage

- RTL designer decides the type of memories and creates a behavioral model for simulation and verification
 - Insufficient for synthesis
- Memory configuration has a big impact on PPA and needs to be physically-aware
 - Existing solutions consider only each individual memory, not physical implementation with all memory wrappers
 - The best configuration for each memory wrapper based on weight does not often correspond to the best for physical implementation
 - The number of combinations of memory configurations for physical implementation is huge – it is impossible to explore all of them physically



MEDIATEK

GRC overflow: 14.83% Total power: 27.4 mW

Memory config 2

GRC overflow: 17.35% Total power: 26.7 mW







Motivation – Feedforward Guidance



4

Reduce project schedule with an automated flow before synthesis

- Manual selection needs iterative work to check physical PPA
- Leverage RTL Architect's quick runtime to do early design space exploration



5

RTL Architect Memory Exploration Methodology

Al-based Reinforcement Learning

Challenges

- The number of combinations of memory configurations for physical implementation is huge – it is impossible to explore all of them physically
- Reinforcement Learning (RL) reduces the iterations to the first best result
 - Reinforcement Learning (RL) is the science of decision-making -- it is about learning the optimal behavior in an environment to obtain the maximum reward

REWARD

NVIRONMEN

ACTION

OBSERVATION

Key Features / Advantage

- Native Synopsys Memory Compiler support
- NDM support to feedforward to Fusion Compiler
- Memory splitting, with control logic inserted
- Powerful filtering capabilities and AI memory selection controls



MEDIATEK

snug

RLMM Key Steps



Identify and Generate

RTL is read with behavioral memory wrappers and the user identifies the required memory instances for exploration

Memories are selected automatically, with support for user-weighting for Area / Performance / Power trade-offs



Select and Instantiate

Top memory candidate solutions are selected (with user control over the number of runs)

Memories are automatically instantiated, with required control logic (where needed)



Explore and Analyze

Exploration Runs Launched in Parallel



Runs can be analyzed / compared

Recipes			Tin	ning	9		Power					Congestion								
QORsum	1	Flows 📰 Met	rics	Info			_													
		Flow -	Timing				Power					Congestion								
QOR Summary			WNS -	TNS A		NVE -	Total		Leakage -		Switching	CellsinCA	OverflowTotal	OverlfowMax						
Power Details		util0.60_ratio1_1_m9	-0.4	7 -3	75.54	2439	1		23	25	513	26	2 33611							
Concession Datail	2	util0.70_ratio1_1_m9	-0.5	0 4	09.84	2374	1	112	2	24	519	206	5 66341							
congestion betan	١.,	util0.80_ratio1_1_m9	-0.5	5 -4	53.82	2579	1	111	2	14	519	1486	1 109423							
Physical	•	util0.85_ratio1_1_m9	-0,4	3 -3	26.34	2098	1	111	2	20	519	1959	4 131245							
Coll Count		util0.60_ratio1_2_m9	-0.5	2 4	09.81	2373	1	112	23	24	519	27	7 34510							
Cencoun		util0.70_ratio1_2_m9	-0,4	6 -3	50.84	2192	1	112	2	25	520	261	2 67633							
Cell Area		util0.80_ratio1_2_m9	-0.5	3 -4	49.89	2527	1	111	2	16	519	1245	1 101786							
Multibit		util0.85_ratio1_2_m9	-0.4	3 -3	33.58	1977	1	111	2	21	519	2423	5 145311							
Clockastina		util0.60_ratio2_3_m9	-0,4	6 -3	35.20	2231	1	109	2	29	488	34	35584							
energyming		util0.70_ratio2_3_m9	-0,5	3 -4	35.38	2527	1	111	2	18	518	189	0 64227							
Timing	٠	util0.80_ratio2_3_m9	-0.3	9 -2	94,12	1933	1	112	22	25	519	1540	9 114389		1					
Logical DDDs		util0.85_ratio2_3_m9	-0.4	3 -3	37.28	2094	1	111	2	19	519	2403	7 144260		1					
Logical DRCs		util0.60_ratio3_2_m9	-0,5	4 4	80.97	3052	1		2	19	519	72	0 35539							
Scenario Details		util0.70_ratio3_2_m9	-0.5	2 -4	10.04	2466	1	110	2	17	505	333	2 68371		1					
Internal/IO Details		util0.80_ratio3_2_m9	-0.5	51 448	48.95	2633	1	110	2	211	519	932	8 92833		1;					
Path Group Details	6	util0.85_ratio3_2_m9	-0,5	3 -4	27.32	2493	1	111	2	14	523	2072	0 126662		3					
· ····································		util0.60_ratio2_1_m9	-0,4	9 -4	00,21	2324	1.11		2	22	516	27	7 33632							

6

RLMM Static and Dynamic Selection







<u>Static</u>: choose top *n* memories from library characterization based on weighting

Collect data from all memories in reference libs

Derive module interface and instance sections of memory configuration

Find **top** *n* **memories** for each wrapper, update memory config with the result of filtered memories

Auto configuration file generation Output characterized configuration data

Dynamic: perform reinforcement-learning physically-aware selection

Apply reinforcement learning (RL) to find the best memory configuration

RLMM Dynamic Workflow

- мерилтек snug
- Based on config file, hook up memory connections and generate the corresponding control logic
- With reinforcement learning, the exploration search space is effectively reduced

Memory configuration file generated automatically from static selection



Reinforcement Learning memory configuration



An intelligent agent takes actions in a dynamic environment to maximize the cumulative reward

Memory candidates for the best physical PPA

Result #5: {u_mdvlpsys-mdvlp_dvfs_top-u_sram_dm0 TGELSCRIJMANSB2KX32B256M8SW1A 1x1} {u_mdvlpsys-mdvlp_dvfs_top-u_sram_pm1 TGEFSCRIJMANSB4KX26B128M8SW0A 1x1} {u_mdvlpsys-mdvlp_dvfs_top-u_sram_dm1 TGEFSCRIJMANSB2KX32B128M8SW1A 1x1} {u_mdvlpsys-mdvlp_dvfs_top-u_sram_pm0 TGEFSSCRIJYMANSA1KX26B128M4SW0A 1x1

PPA report of selected memory configuration

Summary Table Comparator

				Timiı	ng	Instance Count								
Run	WNS	TNS	NVE	Trans Cost	Trans Violation Num	Cap Violation Num	Standard	Register	Repeaters	ICG's				
<u>MEM1_4</u>	0.04	0.00	0	0.00	0	0	1709	26	147	1				
<u>MEM1_3</u>	0.04	0.00	0	0.00	0	0	1651	19	140	1				



Explore with Parallel Space Exploration

Leverage PSE for parameter sweeping



- □ PSE explores memory configurations and placement
- □ Find the best PPA among multiple objectives overall consideration of **congestion/power/timing/area**

9

Parallel Space Exploration PPA Summary

• PSE collects PPA Summary for each memory configuration in HTML format

Desig	Design : /remote/testcases/TC100/032020/3184245/top/kan.ndm														Henre durige timing power companies														
Total	Total 4 experiments, 4 finished														QoR vanmary	Path Summary & Doign Summary Scenaria (unc. 1970)							Semantia func.es0p95v125c +						
Summary Table Comparator															Scenarile free schekfy:125z	WNS	138	NVE 61	10.05	Total _0.34	regioneg J0.34	in-r-reg	regi>est	in-rout 0.00	Number of Nots New With Max Trave Violations	5406			
		Timing Power Congestion									Inst			fear and physical de contract		-34.94 64		TNS NVF	-3.90	-3.90	0.00	0.00	0.00	Nets with Men Cap Vadations	33				
Run		WNS	TNS	NVE	Total	Leakage	Switching	Image	#Cells in Congested Areas	Over	rflow Max	#GR	C's has max overflow	Standard	Register	Re												Nets Will Violeton	30
util90_aspec	<u>a1-2</u> -	-0.58	-15.08	64	4.61e-03	2.93e-03	6.68e-04		25	44	4	34 (0.44%)	1	4413	148			QOR Details								Senaria June	c.as0p86v125c +	Path Group Scenario	Al • func.ss0p05v125c •
<u>util60_aspec</u>	<u>a1-2</u> -	-0.61	-15.72	64	4.20e-03	2.80e-03	4.53e-04	1998 (BM-2)	0	8	3	4 (0.04%)	2	4422	148	•		Path Group REGIN REGULT ck	42.46 0.49 -0.57	-5.53 0.00 -34.57	s N1	t 1.3 0.1 1.2	Path 0 5	Longth	Critical Clas L10 L10 L10	ik Period	26 B D4	Legic Levels 2 15 to 17 132 15 to 20 54 21 to 23 14	amber of Paths \$6.00 27.00 7.00
util90_aspec	<u>d2-1</u>	-0.59	-15.11	64	4.42e-03	2.79e-03	6.37e-04		673	758	3	681 (9.10%)	3	4396	148			Hierarchy Summary	Top 5 Scene	de funcio	0p95v125c •	Cate	pwy 10	•					
<u>ptil60_aspes</u>	<u>d2-1</u>	-0.64	-16.45	64	4.61e-03	2.93e-03	6.68e-04		35	131	3	126 (1.11%)	1	4431	148	4		Hierarchies with largest	WNS Module	1	W	8	ierarchies w	ith largest NVE Module		1	NVE	Hierarchies with largest TNS Module	1 178

QORsum	😑 🗏 Run	Runs ≣≣ Metrics 🛄 Info 🔅 Settings																								
	🗙 Run 🔺	Setu	ıp							Logical DR	Logical DRCs			st					Power				Routability	Runtime		Parameters
QOR Summary		WNS	S 🔺 TNS	-	VVE 4	r2rWNS	r2	rTNS 🚽	r2rNVE	TranDRCs	 CapDR0 	Cs 🔺	Util	🔺 Sto	CellArea 🔺	StdCells	HBufs		TotalPwr 🔺	LeakPwr 🔺	Gated% 🔺	Bits/Flop 🔺	Overflow%	TotalRun	StageRun	memory_cell 🔺
Performance	1 MEM1_1		0.110	0.0		0	0.000	0	.0	0	0	0) 5	59.0	203	7 61	72	0	1488200	2120	0 10.9	9 1.00	0.35	0	.13 0.1	3 MEM1_1
App Option Details	2 MEM1_2		0.130	0.0		0	0.000	0	.0	0	0	0) 5	51.3	202	7 61	39	0	1465400	2040	<mark>0</mark> 10.9	9 1.00	0.36	0	.14 0.1	4 MEM1_2
Timing / LDBC	3 MEM1_3		0.120	0.0		0	0.000	0	.0	0	0	0) 4	18.7	202	3 61	73	0	1470800	2080	0 10.9	9 1.00	0.31	0	.12 0.1	2 MEM1_3
Timing / LDRC	4 MEM1_4		0.100	0.0		0	0.000	0	.0	0	0	0) 6	5.8	202	8 61	19	0	1472700	2270	0 10.9	9 1.00	0.94	0	.11 0.1	1 MEM1_4
Timing Summary	5 MEM1_5		0.100	0.0		0	0.000	0	.0	0	0	0) 5	57.7	204	4 61	58	0	1487800	2480	0 10.9	9 1.00	0.56	0	.11 0.1	1 MEM1_5
Path Group Details	6 MEM1_6		0.080	0.0		0	0.000	0	.0	0	0	0	8	86.5	202	. <mark>1</mark> 61	73	0	4207000	2300	0 12.	5 1.00	1.39	0	.17 0.1	7 MEM1_6
Logical DRCs																										

QoR Summary for the top 6 candidates – the User can specify the numbers of rankings

RLMM Experimental Results

- Validated in FC with AI-suggested memories
 - **Design 1** (4 wrappers) : Focus on power reduction
 - 5.8% power reduction with updated mem configuration and re-placement
 - **Design 2** (54 wrappers) : Focus on power reduction
 - 2.8% power reduction with updated mem configuration and re-placement
 - **Design 3** (135 wrappers) : Focus on congestion reduction
 - GRC down from 1.26% to 0.58%, reduced 51% track overflow with updated mem configuration and replacement
- With RTL Architect's quick runtime, exploration TAT is reduced from months to hours

Power improved without area/timing/GRC overhead

Runtime is dependent on design size, irrelevant to number of wrappers

- Memory selection during the RTL design stage should consider physical implementation → RLMM provides AI-driven memory exploration for physicallyaware PPA optimization
- RLMM facilitates exploration flow with automation instead of manual iterative loops → Memory selection TAT can be reduced from months to hours
- **RLMM** testing produced 3%-6% power improvement and 51% track overflow reduction, and the results were validated with Fusion Compiler synthesis
- **RLMM** helps with AI-based smart exploration and physical placement with quick optimization runtime at the RTL development stage

THANK YOU

Our Technology, Your Innovation™

٠