RioSet: A Design-Centric Open-Source Dataset for Enhanced EDA Applications in Machine Learning

Yifei Zhu, Zhenxuan Luan, Guohua Yin, Xinze and Zhangxi Tan*

RISC-V International open-source Laboratory, TsingHua University

OVERVIEW

RioSet, a pioneering open-source Electronic Design Automation (EDA) dataset, is tailored specifically for various Machine Learning (ML) tasks. Compatible with OpenEDA tools, RioSet abstracts the EDA system into a universal set of features, organized into multiple ML-specific views. Each view encapsulates elements and features vital for training, enabling a structured and focused approach to optimize system performance.

PROBLEM FORMULATION

- The hardening process can be formalized as optimizing DB feature sets **F** through operations $\mathbf{s_1}$ to $\mathbf{s_N}$:

$$s_1(F_1) = F_{s1}, \ldots, s_N(F_N) = F_{sN} = F_{final} \quad (1)$$

- Optimization methods $\mathbf{o_i}$ aim to minimize cost functions $\mathbf{f_i}$ for each task:

$$\arg\min_{o_i} f_i[o_i(F_i)] \quad (2)$$

- Finding suitable views for a task $\mathbf{f_i}$ is solving Eq.2's dual:

$$\arg\min_{F_i} f_i[o_i(F_i)] \quad (3)$$

FEATURE EXTRACTION FROM OPEN DB

Key contributions:

- ❖ 1. Predefined optimization views;
- ❖ 2. HDL classification using LLMs;
- ❖ 3. Full compatibility with open-source EDA tools.

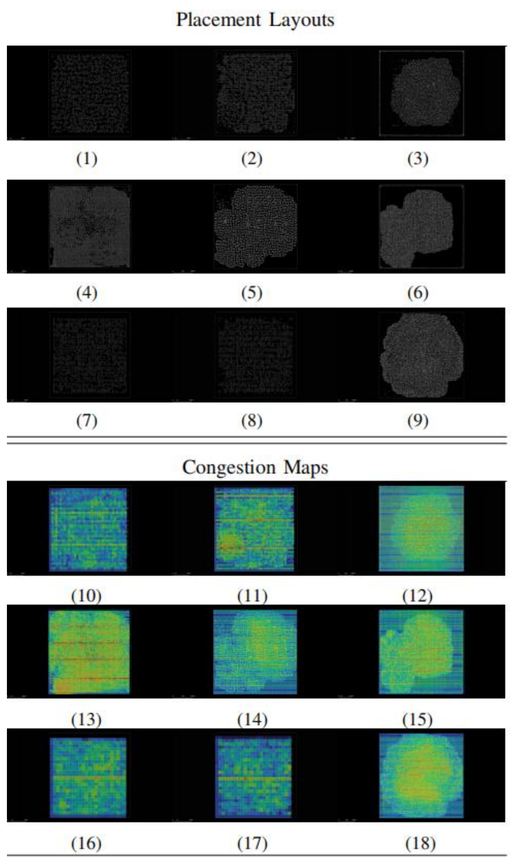

```
"CLASS": "function_module",
"PDK": "sky130A",
"CELL": "fd_sc_hd",
"AREA": 128718,
"Hierachy": 0,
"MACRO": 0,
"STRATEGY::AREA0":{
    "GATE_COUNTS": 12229,
    "SYN_AREA": 128718
},
"CLOCK_PERIOD::11.35":{
    "SYN_SLACK_MAX": "5.65",
    "START": "Y87[1]",
    "END": "_21258_",
    "PATH_GROUP": "clk",
    "INPUT_DELAY": 2.27,
    "DEPTH": 9,
    "STATUS": "MET",
    "SYN_SLACK_MIN": "0.17"
},
"CLOCK_PERIOD::3":{
    "SYN_SLACK_MAX": "-0.76",
    "START": "_21366_",
    "END": "_21380_",
    "PATH_GROUP": "inout",
    "INPUT_DELAY": 0,
    "DEPTH": 12,
    "STATUS": "VIOLATED",
    "SYN_SLACK_MIN": "0.17"
}
```

(a) logic synthesis

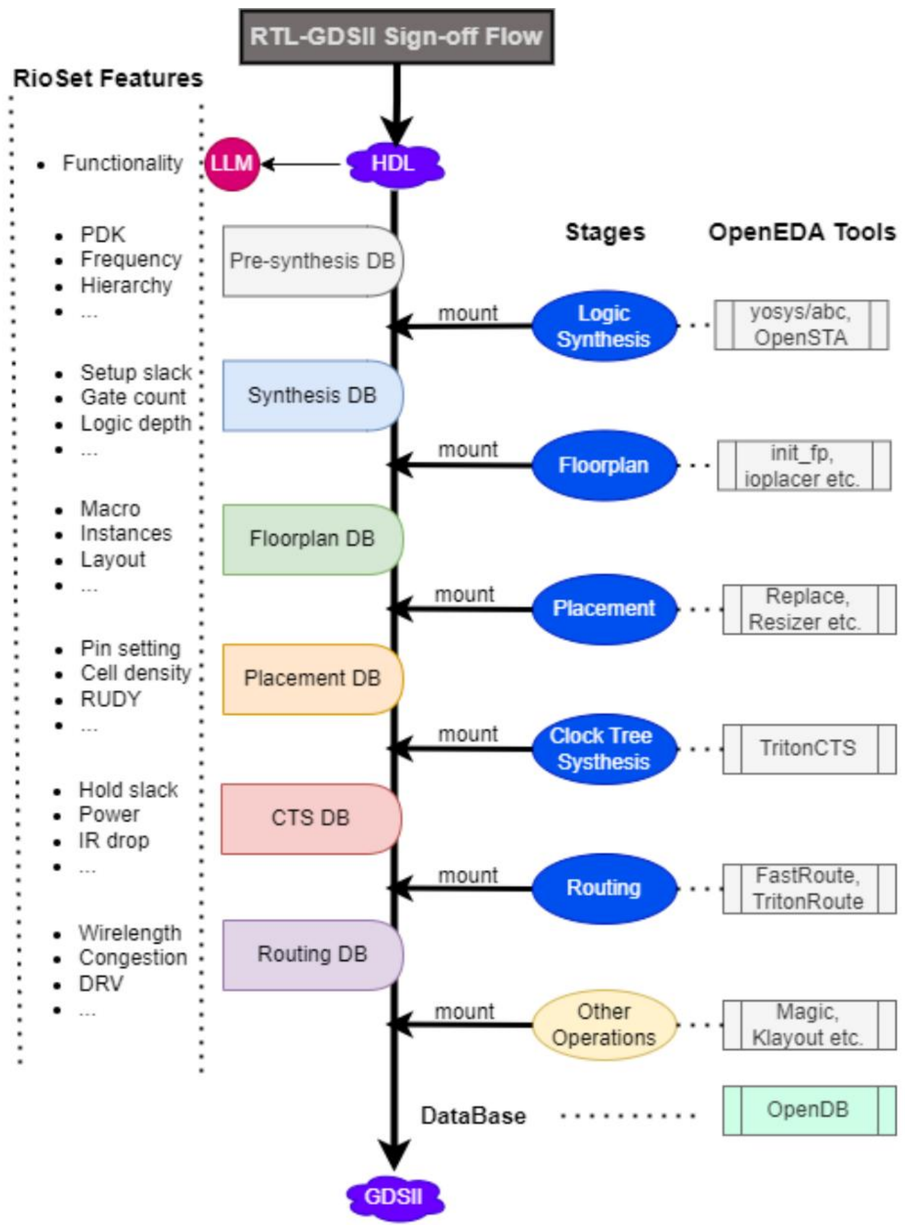
```
"NET": 14145,
"AREA": 143772,
"VIAS": 110173,
"DIE": [0,0,700,700],
"WIRELENGTH":{
    "li1": 0,
    "met1": 342190,
    "met2": 354733,
    "met3": 123665,
    "met4": 108018,
    "met5": 365
},
"PIN": {
    "NUM": 749,
    "MET2": 385,
    "MET3": 362,
    "POWER_PIN": "2 met4 met5"
},
"POWER": {
    "INTERNAL": "2.34e-02",
    "SWITCHING": "8.31e-03",
    "LEAKAGE": "6.52e-08",
    "TOTAL": "3.17e-02"
},
"CLOCK_PERIOD::8": {
    "PL_WNS_MAX": "3.63",
    "PL_WNS_MIN": "0.10",
    "RT_WNS_MAX": "3.60",
    "RT_WNS_MIN": "0.09"
}
```

(b) physical implementation

Textual Representation of Features (excerpts)



Layouts and Congestion Maps in RioSet (partial)



RioSet: available features and extraction stages

LLM FOR HDL CLASSIFICATION:

(a)-(c) and (d)-(e) are obfuscate and enhance operations respectively. A key insight is the crucial role of top module names in classification outcomes.

Correct: 12	core:2, storage: 1 communication: 2 function module: 7	Operation (a)	100% Correct
		Operation (b)	83.3% Correct
		Operation (c)	41.7% Correct
Incorrect: 5	core: 1 storage: 1 communication: 1 function module: 2	Operation (d)	80% Incorrect
		Operation (e)	10% Incorrect

Relevant dataset and auto-generation scripts are open-source and available on:

GitHub: https://github.com/b224hisl/RioSet

Email: zhuyf20@mails.tsinghua.edu.cn

TBSI 清华-伯克利深圳学院 Tsinghua-Berkeley Shenzhen Institute

清华大学 Tsinghua University

Berkeley University of California