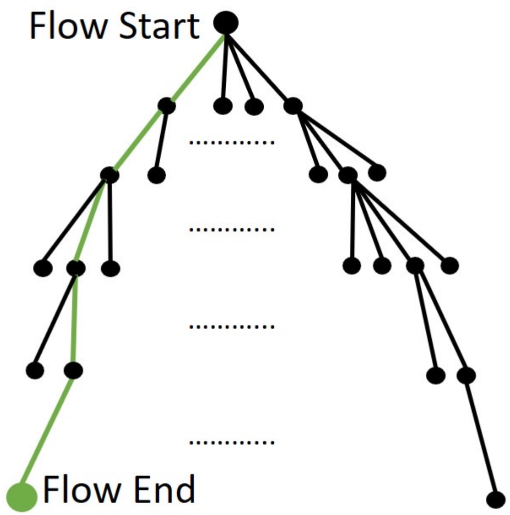
Solvers, Engines, Tools and Flows: The Next Wave for AI/ML in Physical Design

Andrew B. Kahng
UCSD CSE and ECE Departments

abk@ucsd.edu

https://vlsicad.ucsd.edu

AI/ML in Physical Design



Stages of ML Insertion

1. Mechanize, automate
2. Orchestration of search, opt
3. Pruning via predictors, models
4. Reinforcement learning, intelligence

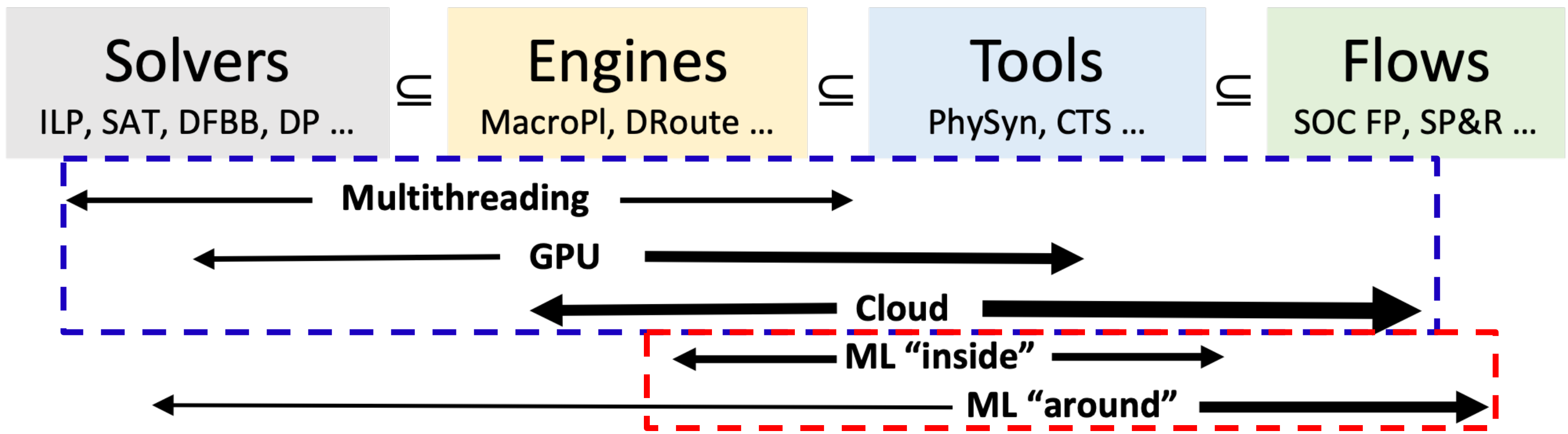
ISPD-2018

Main Threads:
- **Prediction**
- **Optimization**
- **Generation**

Since ISPD-2018

- **Successes**
  - Simple physics by regression: timing across corners, EM/IR, …
  - Black-box hyperparameter search: Cerebrus, DSO.ai
  - Use of ML for hints and ballpark starting points
- **Disappointments**
  - **Tool silos** are more closed
  - **No prospect** of companies sharing data, or of public foundation models
  - **Costs** (#machines, #licenses, #training/learning passes, …)
- **Surprises**
  - Rush to LLMs and Generative AI
- **Understanding of Challenges and Limits for AI/ML in PD**
  - Optimization QoR: Strong baselines
  - Scale: Problem decomposition
  - Chaos: Sampling
  - Accuracy: High bars for deployment
  - (+ Data, Generalization, Validation, Cost, …)

Lens: Solvers, Engines, Tools and Flows



ISPD-2024, Figure 3

- **Foundations that are orthogonal (but enabling) to AI/ML**
  - Speed: enables data, iterations
  - Scalability: enables parallel, D/Q
  - Stability: enables prediction, proxies
  - Quality: optimization cost, prediction accuracy
- **AI/ML in PD: context, culture will matter more than “porting”**
  - Open data, open source, standard APIs, reproducibility, benchmarking …

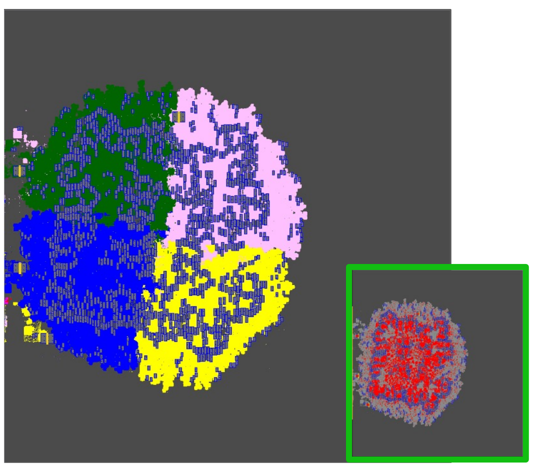
Section 2: PD Challenges and Levers

- **Challenges**
  - Design partitioning and block shaping
  - Placement-aware hierarchical floorplanning
  - Datapath-aware floorplanning
  - Drive for area reduction
- **Levers**
  - GPU-based speedups
  - Cloud
  - Sampling for stability
  - Multiple views in unit time: Tomography

Dataflow-Driven GPU-Accelerated RePlAce

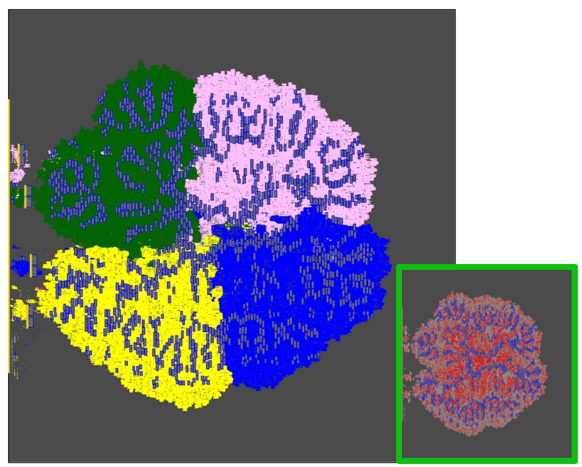
- Global placer for large-scale IP blocks
  - **Speed and scalability:** > 30X speedup vs. RePlAce; 10M insts in 30 min
  - **Quality:** dataflow-driven, physical hierarchy aware placement
  - **Accessibility:** permissive open source, integrated in OpenROAD (coming)
  - **Ease of use:** OpenROAD flow or plug into commercial flow

RePlAce



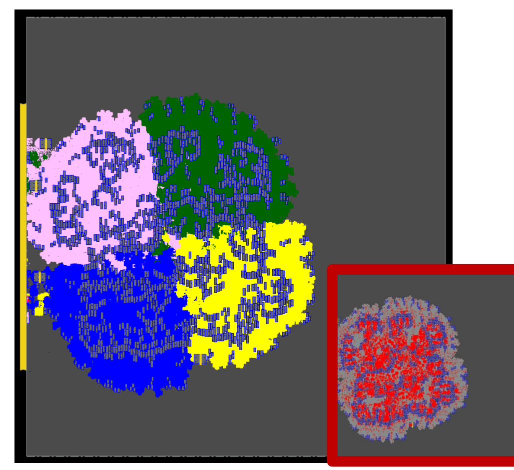
GP Runtime (s): 65381
HPWL/eGR WL (m): 325/404

Commercial



GP Runtime (s): 24561
HPWL/eGR WL (m): 414/478

New



GP Runtime (s): 1808
HPWL/eGR WL (m): 327/407

Testcase: MemPool Cluster, ETH Zurich **(9.5M cells, 1296 macros in NG45)**

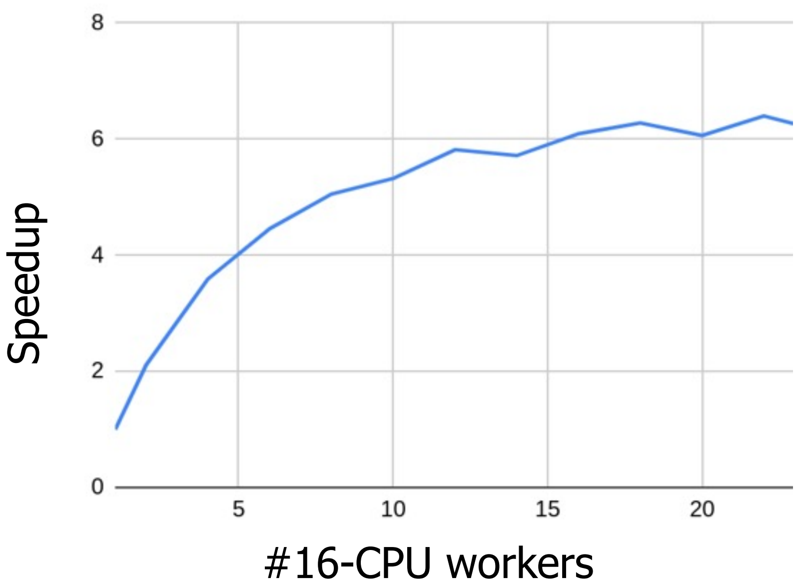
To Do: Fill in GPU-Accelerated PD Flow → *and unleash ML !?*

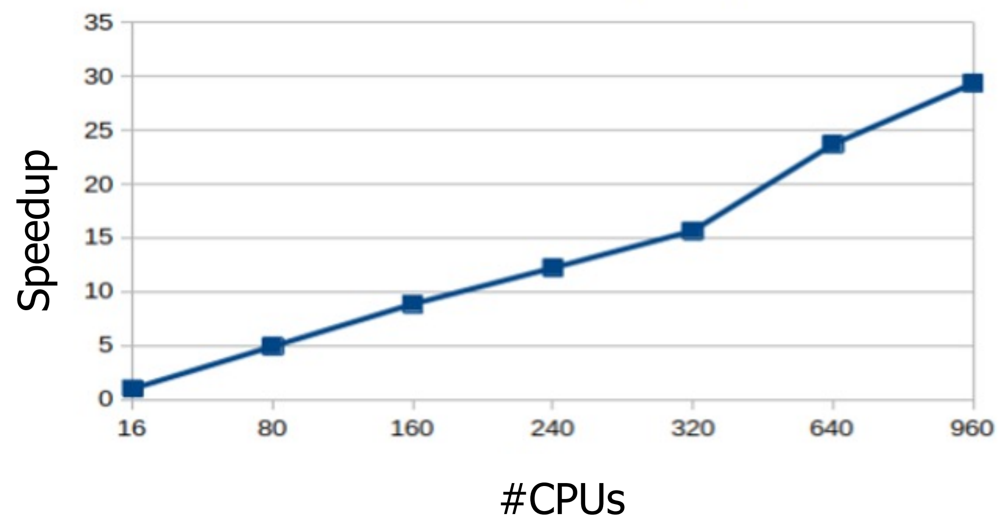
Physical Design Flow	Academic GPU-Accelerated Tools
RTL Simulation	RTLFlow [ICPP'22]
Logic Synthesis	CULS [DAC'23]
Macro Placement	AutoDMP [ISPD'23]
Global Placement	DREAMPlace [DAC'19] , Xplace [DAC'22]
Detailed Placement	ABCDPlace [TCAD'20]
Clock Tree Synthesis	Missing
Global Routing	GAMER [ICCAD'21], FastGR [DATE'22], GGR [ICCAD'22]
Detailed Routing	Missing
DRC Checker	OpenDRC [DAC'23]
STA Engine	[TCAD'23a], [TCAD'23b] (From Prof. Yibo Lin)
Design Closure, Opt	Missing

Cloud Deployment of Optimizers

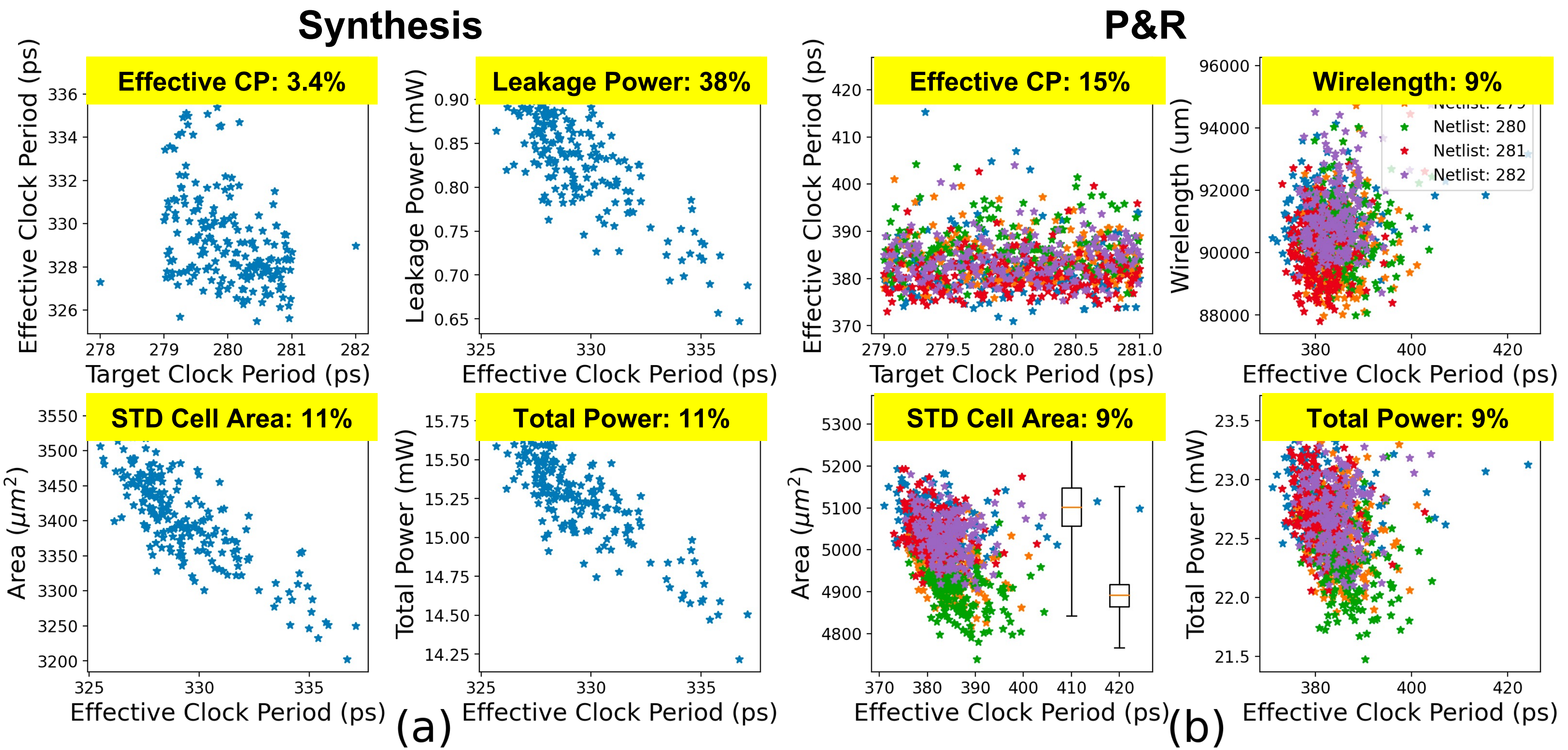
- Complements GPU acceleration; also a “low-hanging fruit”
- Poor alignment with traditional EDA business models

- Distributed incremental DR: ~100X speedup w/20 16-core workers
- Cloud-based pin access analysis: 30X speedup



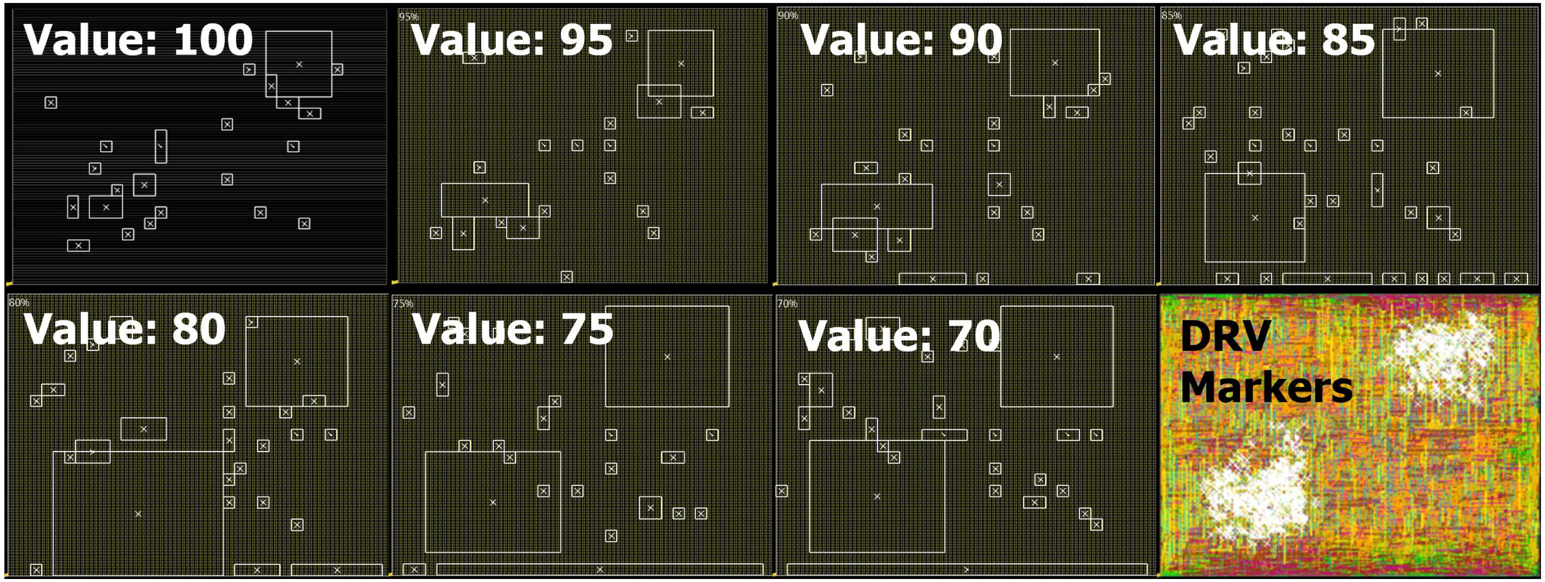


Chaos Demands Sampling (Example: GF12 AES)



Variation of metrics: $100 \times (\frac{max}{min} - 1)$

Multiple Views in Unit Time: Tomography



- Congestion report using Innovus eGR
  - Varying partial density value of the routing blockage from 70 – 100
  - Union of congestion reports enables better ML-based alignment with actual DRC markers
    - Partial density = 75 is closest match to actual markers, in this case
- Routing runtime = 1.5 hours; eGR runtime < 1 second

Section 3: Elements of a Next Wave

- Generative AI
- **ML at Interstices**

"Magic" at Interstices

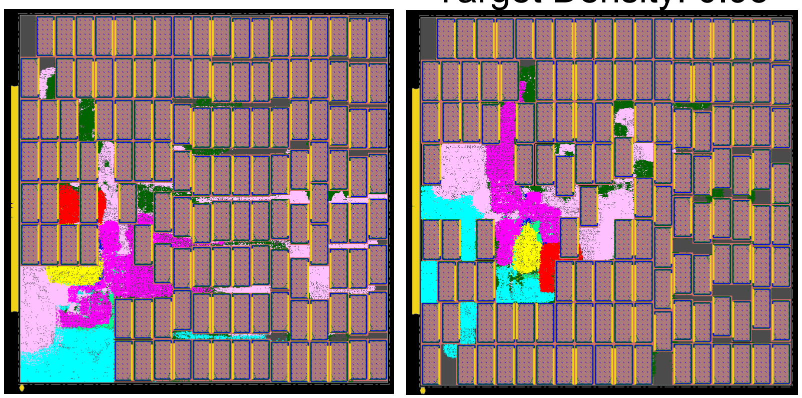
- **Co-evolutions, Co-optimizations are often at arm's length**
- **Interstices = opportunities for "*Conditioning Magic*" via ML**

Co-optimizations	**"Magic"**
• Netlist – Backend	• Netlist
• Hierarchy – Floorplan	• Netlist Partitioning
• Floorplan – SP&R	• Block shaping *+ boundaries*
• Synthesis – P&R	• Placement screens
• Place – Route	• Route screens
• GRoute – DRoute	• Route guides
	• Corners *+ endpoint SDCs*
	• Constraints
	• Tool/engine recipes
	• …

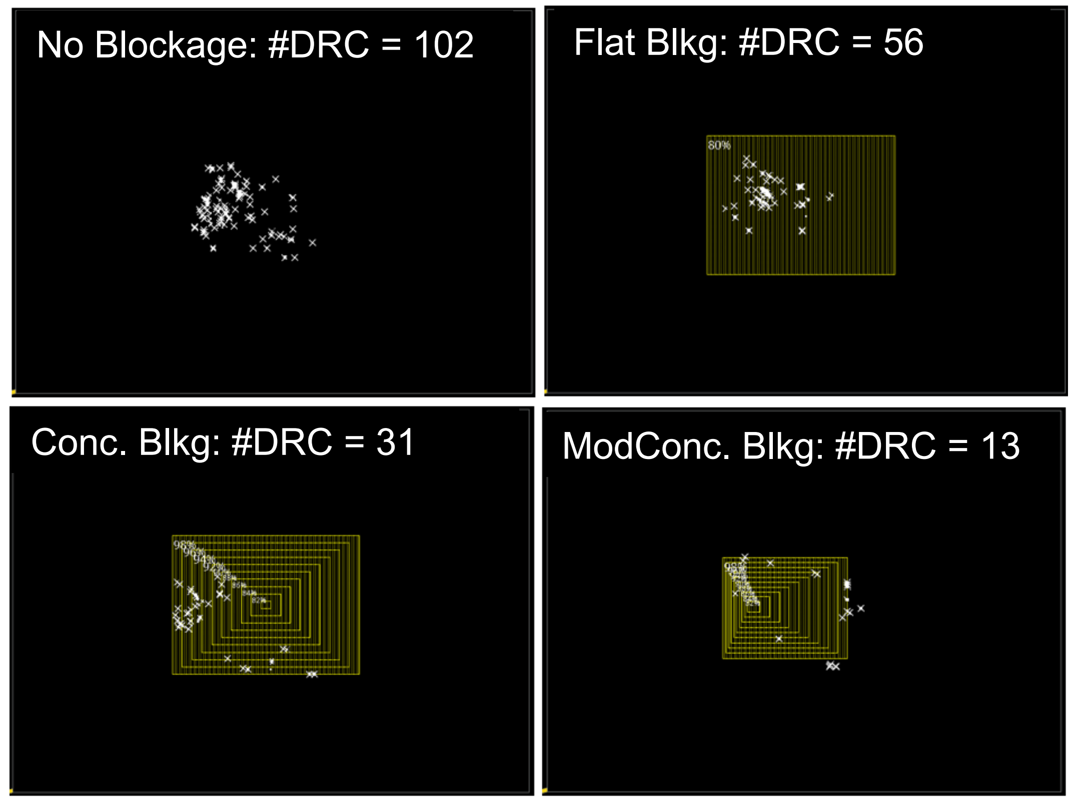
ML to Condition the PD Canvas

- Can "condition" PD with target density, cell padding, placement and routing blockages, …

Default | Target Density: 0.95



Target Density	Wirelength (um)	Total Power (mW)	WNS (ps)	TNS (ns)
Default	4897941	839.8	-450	-541.7
0.95	**4150554**	**812.1**	**-155**	**-154.6**



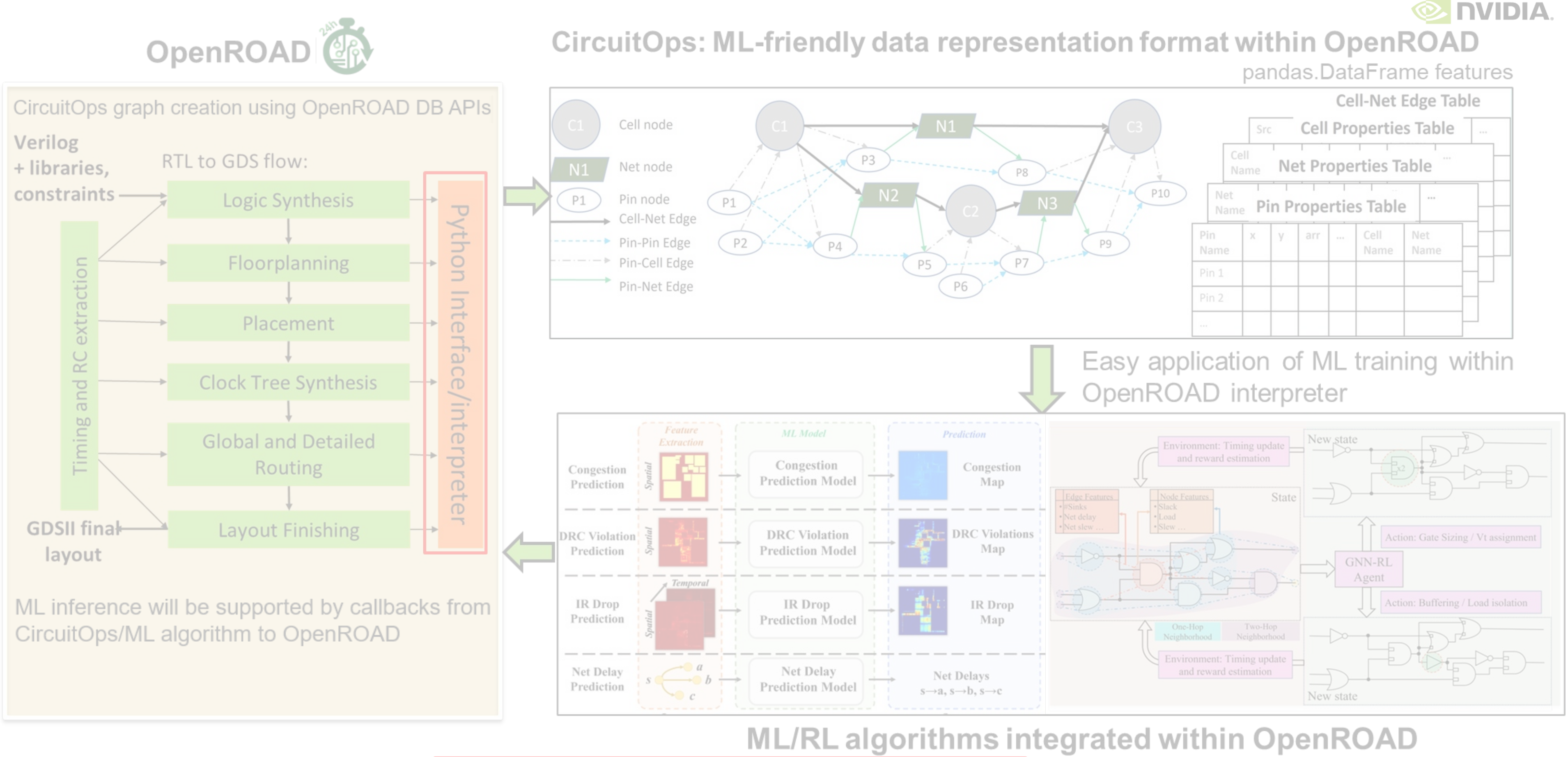
- "Magic screens": placement and routing blockages
- ML goal: Find best settings for these knobs

Section 3: Elements of a Next Wave

- **Generative AI**
- **ML at Interstices**
- **Infrastructure for ML**: platforms, proxies

AI/ML for PD: Infrastructure

https://github.com/NVlabs/CircuitOps



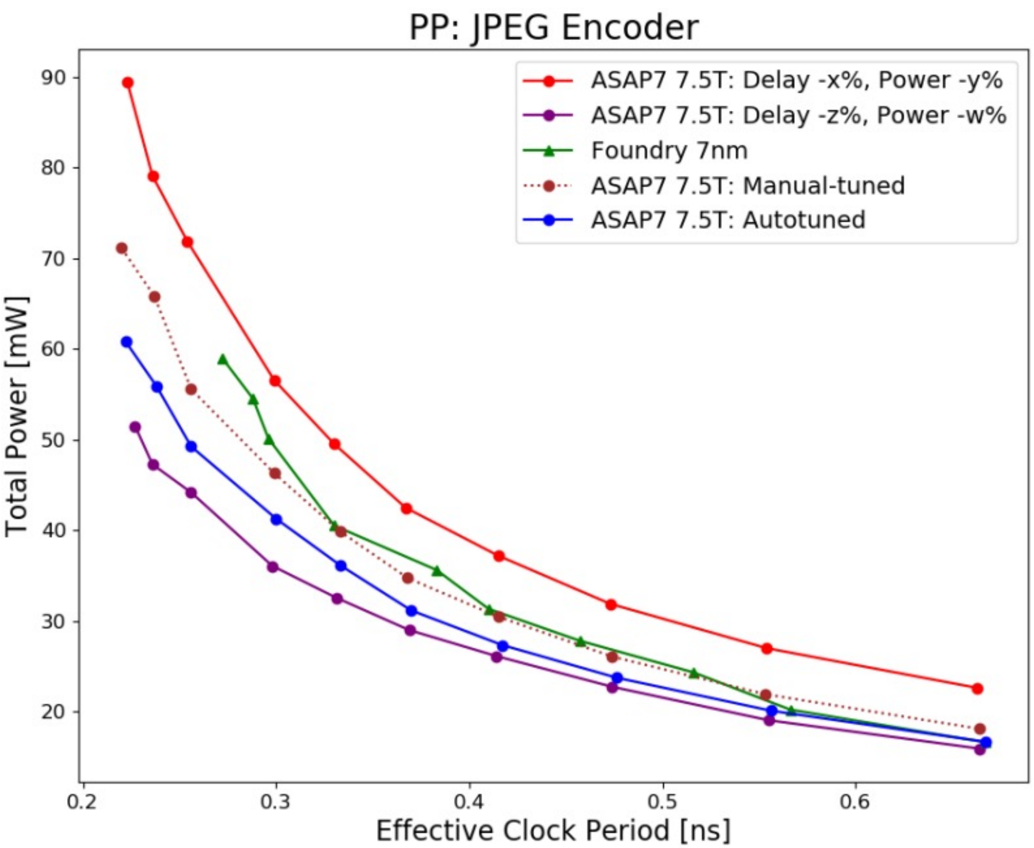
CircuitOps and OpenROAD: Unleashing ML EDA for Research and Education

Andrew B. Kahng, UCSD
Vidya A. Chhabria, ASU
Bing-Yue Wu, ASU

ASP-DAC 2024 Tutorial #8
https://github.com/ASU-VDA-Lab/ASP-DAC24-Tutorial

AI/ML and EDA: Data From Proxies

- **Mindset for AI/ML data: if it is not sharable, develop a proxy!**
  - **PDK:** ASAP7/5 **+ scaling, autotuning**



Power v. Effective CP hockey stick

- Foundry 7nm in **green**
- **Red**, **purple** bounds from simple scaling of ASAP7 delay, power
- **Autotuning** (Ray/Tune) with **(~11%)** loss = MAPE of power, fmax errors at 10 target clock periods
- Tuning parameters: delay, pin cap, internal/switching power, setup/hold …

Scripts are open-sourced in RDF-2023: https://github.com/ieee-ceda-datc/RDF-2023

Section 3: Elements of a Next Wave

- **Generative AI**
- **ML at Interstices**
- **Infrastructure for ML**: platforms, proxies
- **Culture changes**

Opening Doors and Minds

ieee-ceda-datc

CUHK EDA

YosysHQ

OpenROAD

Xyce

VERILATOR

XLS

"Next Wave of AI/ML in Physical Design" will be sparked by the accessibility, scale and velocity of open-source EDA (& more) !

ASP-DAC 2024

29th Asia and South Pacific Design Automation Conference

iEDA: An Open-source infrastructure of EDA

Xingquan Li, Zengrong Huang, Simin Tao, Zhipeng Huang, Chunan Zhuang, Hao Wang, Yifan Li, Yihang Qiu, Guojie Luo, Huawei Li, Haihua Shen, Mingyu Chen, Dongbo Bu, Wenxing Zhu, Ye Cai, Xiaoming Xiong, Ying Jian, Yi Heng, Peng Zhang, Bei Yu, Biwei Xie, Yungang Bao

Jan. 23 2024

"Open-source is not a goal but a way"

iPD: An Open-source intelligent Physical Design Toolchain

Importance of Open-Source EDA Tools for Academia

Open Letter on European Strategic and Funding Directions

To Whom It May Concern

March 8, 2024

Initial Signatories

Luca Benini
University of Bologna, Italy & ETH Zürich, Switzerland
Professor, Lead of the RISC-V PULP platform

Giovanni De Micheli
EPFL Lausanne, Switzerland
Professor and Director LSI lab

Marie-Minerve Louërat
Sorbonne University, France
Research Scientist, Coriolis Foundation hosted by CNRS Foundation

Harald Pretl
Johannes Kepler University Linz, Austria
Professor, Maintainer of IIC-OSIC-TOOLS

Stefan Wallentowitz
Hochschule München University of Applied Sciences, Germany
Professor, Director at FOSSi Foundation & Director at RISC-V

Signatories (62)

https://open-source-eda-letter.eu/

Thank You For Watching !

I look forward to seeing you soon in Taipei !

Acknowledgments: Many thanks to Sayak Kundu, Bodhisatta Pramanik, Zhiang Wang and Dooseok Yoon for their help with the figures and text in this paper. Discussions with Siddhartha Nath, Igor Markov, Chuck Alpert and Ilgweon Kang are also gratefully acknowledged. Research at UCSD is partially supported by DARPA, Samsung, the C-DEN center, and gifts from Google, Intel and others.