CircuitOps (GenAI Infrastructure)

Rongjian Liang, Anthony Agnesina, Geraldo Pradipta, Vidya Chhabria*, Haoxing Ren

&

AutoDMP (Macro Placement)

Anthony Agnesina, Puranjay Rajvanshi, Tian Yang, Geraldo Pradipta, Austin Jiao, Ben Keller, Brucek Khailany, Haoxing Ren

*Vidya is from Arizona State University, and all other authors are from Nvidia

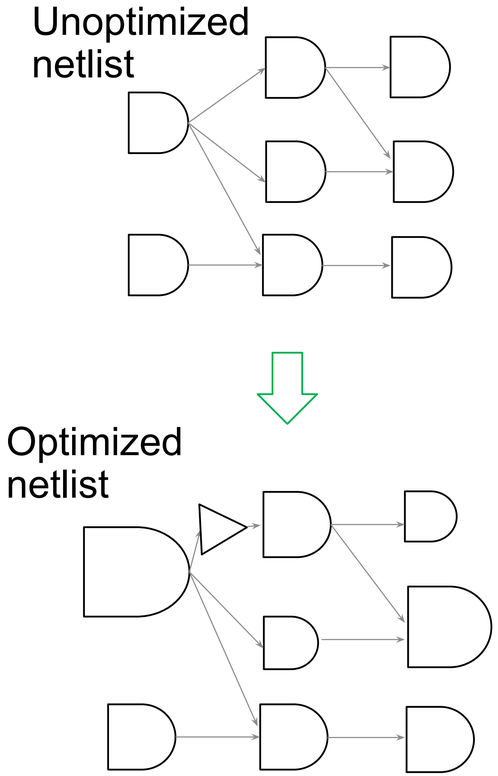
Rongjian Liang, Nvidia

Agenda

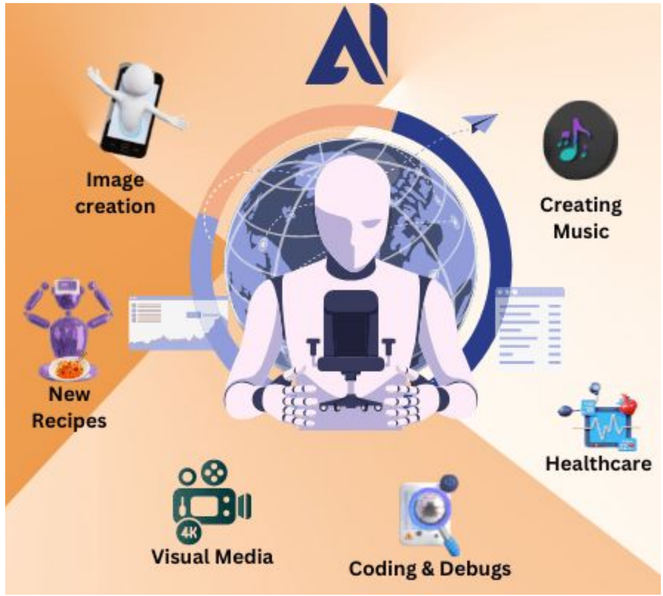
- CircuitOps: GenAI infrastructure for VLSI netlist opt.
- AutoDMP: Automated macro placement
- Conclusions

CircuitOps

Motivations: barriers between GenAI and VLSI netlist Optimization



- **AI-unfriendly data structures and interfaces** of EDA tools, hindering high volumes of data queries, transfer and processing
- **Steep learning curve of EDA knowledge and tools** for AI experts
- **Lack of shared IR** (intermediate representation), hindering data reuse across projects



VLSI netlist optimization tasks (e.g., buffering and sizing): **generating optimized netlists from unoptimized netlist**

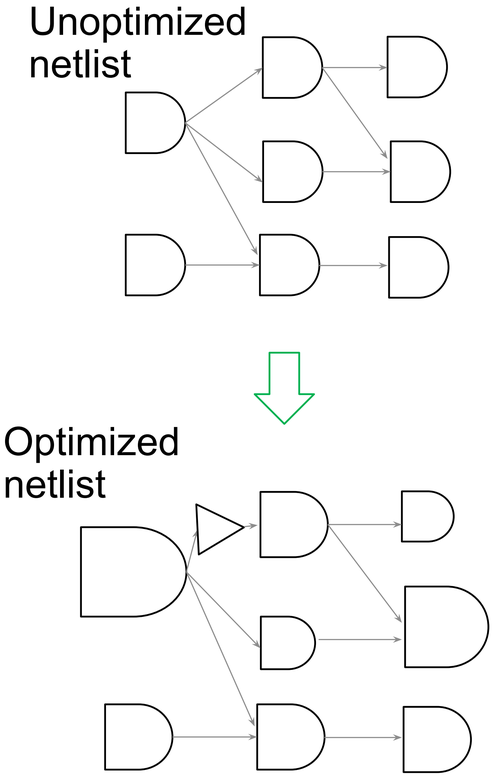
Barriers: Challenges in **data query, transfer, processing and sharing** for VLSI netlist opt. tasks

Generative AI techniques:

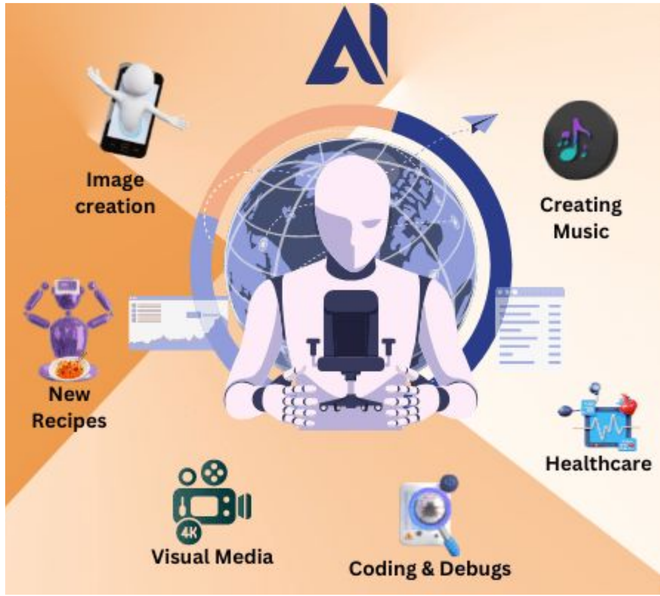
Data is the new oil for GenAI

CircuitOps

Contributions: bridging GenAI to VLSI netlist optimization



- Flexible and AI-friendly shared IR
- Tool-agnostic IR generation
- Customizable dataset generation
- Inference with gRPC-based data transfer



VLSI netlist optimization tasks (e.g., buffering and sizing): **generating optimized netlists from unoptimized netlist**

Bridge: An **infrastructure to facilitate data query, transfer, processing and sharing** for VLSI netlist opt.

Generative AI techniques:

Data is the new oil for GenAI

CircuitOps Overview

- The formats are supported by popular commercial and academic tools, making our IR generation process **tool agnostic**

- Operating on the AI-friendly LPG to generate customized dataset

Standard EDA files (DEF, LEF, .v, …) and timing attribute tables

EDA tools

CircuitOps

IR Generation

LPGs

Dataset Generation

Labeled Property Graph as IR

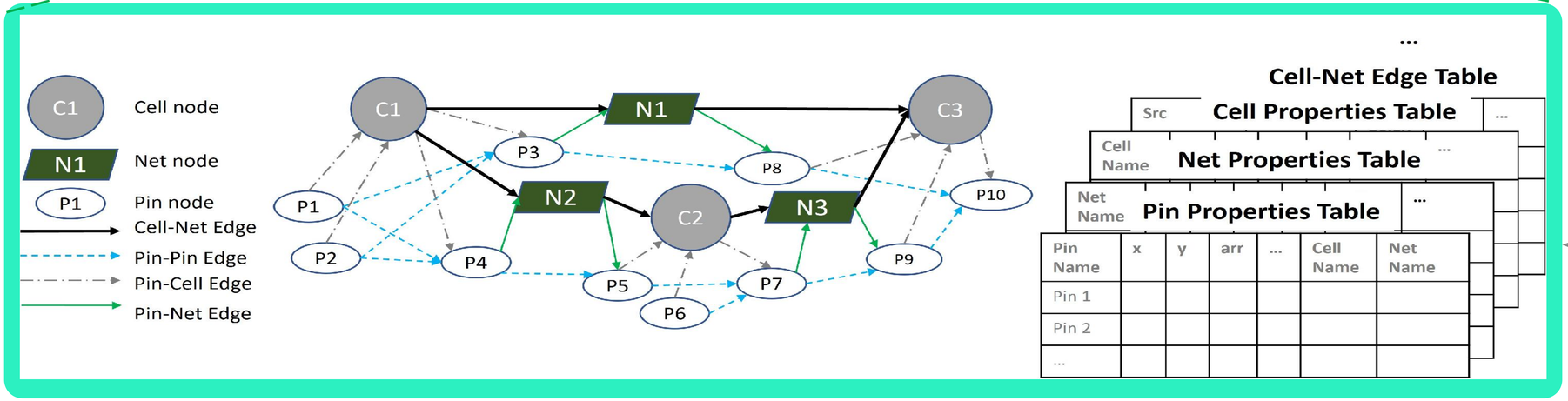
Sizing dataset, buffer tree

GAI circuit optimization training

gRPC

GAI circuit optimization inference

- Facilitating model deployment into product

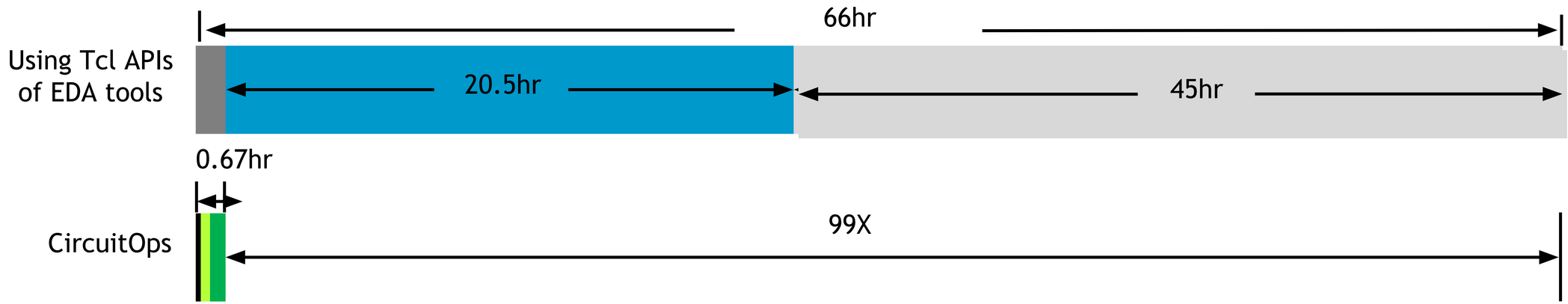


- **Flexible** way to describe netlist info.
- Backend relational tables offer opportunities for **massive parallelism by matrix operations**

NVIDIA

Experimental Results

- **Dataset Generation for Sizing and Buffering**
  - Sample 260K buffer trees and 200K timing paths and on a circuit with ~2.3M cells



Loading design in EDA tool

Buffer trees sampling and feature extraction

Timing paths sampling and feature extraction

Loading standard EDA files

LPG generation

In C++

Dataset generation for sizing and buffering on LPG

In Python, leveraging parallel operations in pandas, numpy, dgl and graph-tool

- **gRPC-Based Data Transfer for Buffering Inference**
  - CPU server <-> GPU server
  - Throughput: 75K nets/sec. Send one net = 5us, Receive one net = 8us

CircuitOps Open-Source (In Progress)

NVlabs/CircuitOps (github.com)

- Done:
  - OpenROAD tcl APIs based IR generation
  - Buffer tree sampling for a few open-sourced circuits
  - gRPC-based data transfer

- Todo:
  - More designs and platforms
  - Gate sizing dataset generation scripts
  - OpenROAD C++ APIs for IR generation
  - Code cleaning and documentation

AutoDMP

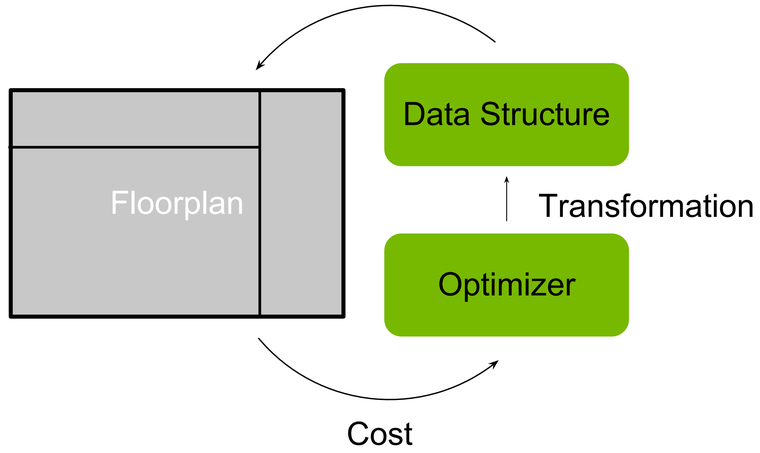
Motivations

- Macro placement is essential for Power-Performance-Area (PPA)

- Mixed-size analytical is **SOTA** macro placement
  - place macros and standard cells concurrently
  - Limited Design Space Exploration with commercial tools

- DREAMPlace is a **superfast** mixed-size analytical placer
  - Accelerate ePlace/RePlace with GPUs/PyTorch
  - Treats macros & standard cells similarly → macro **legalization issues**
  - High influence of parameter settings on optimization quality

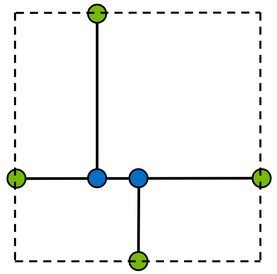
- **AutoDMP → generate quickly various high-quality macro placements**
  - Enhance DREAMPlace: fix legalization issues & expand the design space
  - Tune DREAMPlace settings with Multi-Objective Bayesian Optimization



DREAMPlace Extensions

- Weigh the smooth half-perimeter wirelength (HPWL) of nets
  - Improve correlation with RSMT during global placement
  - Based on pin count (RISA)

$$\mathrm{WL}(\mathbf{z}) = \sum_{e \in E} w(|e|)\mathrm{WL}(e;\mathbf{z})$$



HPWL

- Greedy macro-orientation refinement during detailed placement

- RUDY + Macros + Gaussian filter
  - Congestion score = average of top-10% most congested bins

$$\mathrm{RUDY}(g) = \frac{\sum_{e \in E} \frac{\mathrm{OA}(e,g)}{\mathrm{BBOX}(e)}}{s(g) - \sum_{m \in M} \alpha(m)\mathrm{OA}(m,g)}$$

Net Demand

Routing Supply

Macro Demand

- Miscellaneous fixes

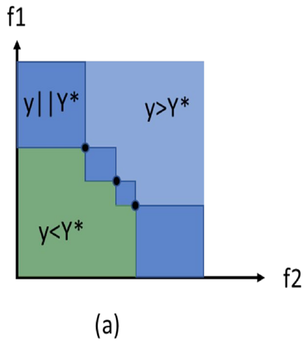
Anthony Agnesina, Puranjay Rajvanshi, Tian Yang, Geraldo Pradipta, Austin Jiao, Ben Keller, Brucek Khailany, and Haoxing Ren, "**AutoDMP: Automated DREAMPlace-based Macro Placement**", International Symposium on Physical Design (ISPD)

Parameter Space and Multi-Objective Bayesian Optimization

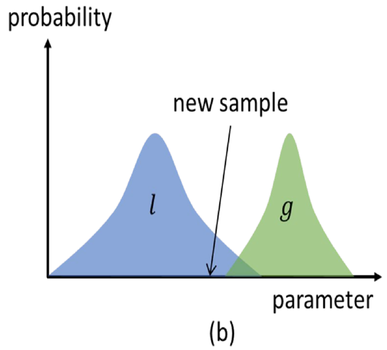
- Base DREAMPlace parameters
  - Optimization-related
  - Density target
- Initial macro/cell locations → placement **diversity**
- Macro halos → **fix legalization** issues
- Large parameters effects

Parameter	Search Range	$\widehat{c_v}$ (%) RSMT	Cong.	Divg. Rate
*horiz. initial position	[0.2, 0.8] (%)	2.2	0.9	0.0
*vert. initial position	[0.2, 0.8] (%)	2.0	1.1	0.0
*horiz. macro halo	technology dep.	1.8	1.3	0.0
*vert. macro halo	technology dep.	1.7	1.2	0.0
target density d_{target}	$[a_{\text{util}} - 0.2, a_{\text{util}}]$ (%)	-	-	-
density weight	$[1e^{-6}, 1.0]$	3.1	1.7	0.0
smooth HPWL model	{LSE, WA}	0.7	1.1	0.0
smooth HPWL initial γ_0	[0.10, 0.50]	5.1	1.9	0.0
GD initial LR lr_0	$[1e^{-4}, 1e^{-2}]$	1.4	1.0	0.0
GD LR decay	[0.99, 1.0]	6.7	2.3	53.2
GD optimizer	[Adam, Nesterov]	1.2	0.8	54.2
# horiz. global bins	{256, 512, 1024, 2048}	1.3	0.9	0.0
# vert. global bins	{256, 512, 1024, 2048}	3.1	1.3	21.1
λ update lower coeff. L	[0.90, 0.99]	4.2	1.9	0.0
λ update upper coeff. U	[1.01, 1.15]	27.0	7.5	1.8
λ update $\Delta\,\text{HPWL}_{REF}$	$[1.5e^5, 5.5e^5]$	2.3	1.2	0.0

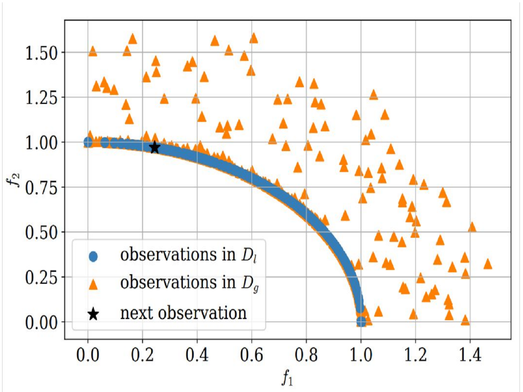
- Search for **multiple competing objectives**:
  - RSMT, density, congestion
- True multi-objective: **Pareto**-front modeling of PPA trade-offs
  - Tree-structured Parzen Estimator **(MOTPE)** for kernel density estimation of good/bad samples



(a)



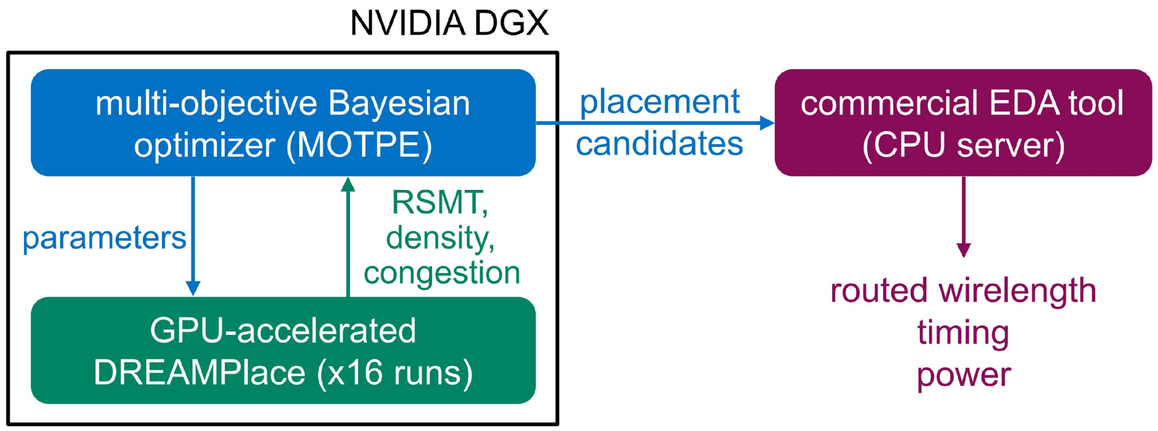
(b)



Y. Ozaki, Y. Tanigaki, S. Watanabe, and M. Onishi. Multiobjective Tree-Structured Parzen Estimator for Computationally Expensive Optimization Problems

Tilos results

- External validation by UCSD on GF12
  - Limited search: 200 samples with 2 GPUs, one candidate



Design Enablement	Macro Placer	Area (μm^2)	rWL (mm)	Power (mW)	WNS (ps)	TNS (ns)
Ariane GF12	CT	0.138	1.000	1.000	-0.145	-123.4
	CMP	0.139	**0.865**	0.990	-0.159	-142.3
	RePlAce	0.140	1.042	1.015	-0.168	-197.4
	SA	0.139	0.925	0.995	-0.155	-178.0
	AutoDMP	**0.137**	0.885	0.985	**-0.130**	**-90.5**
	Human	**0.137**	1.064	**0.981**	-0.139	-106.6
BlackParrot GF12	CT	0.179	1.000	1.000	0.001	0.000
	CMP	0.178	0.593	0.918	0.001	0.000
	RePlAce	0.178	0.798	0.959	0.000	0.000
	SA	0.178	0.731	0.944	0.000	0.000
	AutoDMP	0.178	**0.587**	**0.917**	0.000	0.000
	Human	0.178	0.642	0.928	0.000	0.000
MemPool GF12	CT	0.410	1.000	1.000	-0.195	-1849.4
	CMP	0.405	**0.821**	**0.895**	-0.197	-1961.3
	SA	0.412	0.991	1.000	-0.187	-2442.7
	AutoDMP	**0.402**	0.843	**0.895**	-0.213	**-1015.7**
	Human	0.406	0.888	0.920	**-0.149**	-1766.5

Cheng, C.K., Kahng, A.B., Kundu, S., Wang, Y. and Wang, Z., 2023. Assessment of Reinforcement Learning for Macro Placement. arXiv preprint arXiv:2302.11014.

Conclusions

- CircuitOps: A GenAI infrastructure for VLSI netlist optimization
- Open-Source: NVlabs/CircuitOps (github.com)

- AutoDMP: Automated DREAMPlace-based macro placement
- Open-Source: NVlabs/AutoDMP (github.com)


NVIDIA