AI, EDA and Disruptive Innovation

CEDA Distinguished Lecture
June 25, 2024

Andrew B. Kahng
UCSD CSE and ECE Departments

abk@ucsd.edu

https://vlsicad.ucsd.edu
Takeaways

• **EDA = optimization + automation** holds main levers for scaling

• Closed AI/ML silos/platforms \(\square\) academia must enable itself
  • Action: curated data + domain knowledge, culture (rewards, badges, …)

• AI/ML in EDA is **difficult**: optimization QOR, data needs, ML ops
  • Action: baselines, replication in the open
  • Action: high-value target selection

• Innovation beyond a “toy”: **shared**, sustained, patient efforts
  • Action: infrastructure with professional staff, longer-term support horizons

• Proxies have gaps – need coordination to close these!
  • Action: Design enablement: PDK scalers, foundation IP, calibrations
  • Action: Design tools: formal and physical verifications; DFT; HLS-PS-LS
  • Action: Stackable multiphysics solvers, compact models

• Optimization: many rich vistas + can race faster in the open
Agenda

• EDA and Scaling
Scaling

- **Scaling** = *getting better results with less resources*
  - People, money, time, energy, area, …
  - *Moore’s Law: 1 week = 1%*
Scaling is **Design-Based**

Mark Papermaster, keynote, Design Automation Conference, July 2022
Barriers: Cost, Expertise, Risk

Andreas Olofsson, keynote, Intl. Symp. on Physical Design, March 2018
Design Scaling = EDA Scaling

- **Scaling** = *getting better results with less resources*
  - People, money, time, energy, area, …
  - *Moore’s Law:* 1 week = 1%

- **Design** scaling requires EDA scaling
  - More designers and designs → more EDA engineers
  - Differentiated design capability → bespoke EDA

**Must scale design, EDA and people**
EDA = Automation, Optimization for Design

- A supplier industry to the semiconductor, electronics industries
- An enabling technology for creation of new IC products

The core challenge of EDA and IC design: OPTIMIZATION
- \(O(\text{year})\) for a new chip
- \(O(\text{weeks})\) for synthesis, place-and-route, opt “flow”
- Optimization: best-possible End s.t. resource bounds

Resource “box”: Compute \(\times\) Licenses \(\times\) People \(\times\) Weeks

EDA = huge, high-stakes, intractable optimizations
Takeaways

- EDA = optimization + automation holds main levers for scaling
Agenda

• EDA and Scaling
• AI and EDA: Looking Back
New Directions for Learning-Based IC Design Tools and Methodologies

ASP-DAC 2018 Session 5A

Andrew B. Kahng
CSE and ECE Departments
UC San Diego
http://vlsicad.ucsd.edu/~abk/
abk@ucsd.edu

Quality, Schedule, and Cost: Design Technology and the Last Semiconductor Scaling Levers

Andrew B. Kahng
CSE and ECE Departments
UC San Diego
http://vlsicad.ucsd.edu
ASP-DAC 2018 Keynote

- ML = scaling: modeling + prediction; optimization objectives
- ML = Foundation #1 for “Last Levers” of scaling

Machine Learning Gives Us Scaling!

• High-value opportunities in and around EDA

Modeling and Prediction
- Predict tool outcome = F(design, constraints)
- How to run tool “optimally” for given design
- Avoid “failed runs” → reduce iterations in design flows
- Dream: one-pass design flow
- Model analysis errors (crude vs. golden)
- Reduced guardbands and pessimism → better tooling

Optimization (ML models = objectives)
- Better use of resources (tools, schedule)
- Project-level prediction, adaptive schedules
- Today: the major focus for IC industry
- U.S. DARPA IDEA program: automated

Takeaways

- Quality, Schedule, Cost are “the last levers for semiconductor scaling”
  - Accessibility of hardware / semiconductor design
  - Continue semiconductor value trajectory (for a while longer)

- Foundation #1: machine learning in, around EDA
  - Pervasive ML → Drive down iterations, margins
  - Cloud-targeted, large-scale optimizations → drive down TAT

- Foundation #2: open-source EDA
  - Will a “Linux of EDA” be possible this time around?

- Foundation #3: partitioning and cloud EDA
  - Also part of schedule reduction
  - Design Capability Gap is a crisis for the industry
  - Need all hands on deck!
ASP-DAC 2018 Invited: Learning-Based EDA

- **ML EDA opportunities: modeling, prediction, correlation**
- **Challenges: technology and industry**

### Opportunities
- Reduce schedule, optimize resource usage
  - Better downstream prediction → fewer iterations
  - Models of TAT and uncertainty → better decisions
- Improve design QoR
  - Analysis correlation → reduced design cycles
  - Downstream flow prediction → more accurate designs
- “No human in the loop” IC design

### Challenges
- **Technology challenges**
  - “Small data” problem alongside “big data” problem
  - Huge implementation space, difficult parameter identification
  - Complicated by tool versions, design versions, technology changes (pictures of cats and trees don’t change every year)
  - Possibly helpful: EDA folks know what’s in their tools!
- **Industry challenges**
  - Who (fabless, EDA or foundry) will drive ML into design enablements and production flows?
  - How will models be standardized and (partially) shared?
    - EDA {doesn’t like to, doesn’t know how to} model itself
    - Depend on customers and data to understand needs
    - Is reduction of design schedule and resource good for EDA?
- **Can we realize “METRICS”?**
  - [http://vlsicad.ucsd.edu/GSRC/metrics](http://vlsicad.ucsd.edu/GSRC/metrics)
Observation: Because of the red X’s, many “target” and “todo” items from 2018 are still open today. 😞
Since 2018 …

- **Successes** (*many were low-hanging fruits*)
  - Simple physics by regression: timing across corners, EM/IR, …
  - **Black-box hyperparameter search:** Cerebrus, DSO.ai
  - Use of ML for early hints and ballpark starting points

- **Disappointments** (*many are unsurprising*)
  - **Tool silos** are more closed (but with data platforms such as JedAI)
  - **No prospect** of companies sharing data, or of public foundation models
  - **Costs** (#machines, #licenses, #training/learning passes, …)

- **Surprises**
  - Rush to LLMs and Generative AI (train, debug, testbench, copilot, manpages, …)
  - Near-total sway of (EDA) Suppliers in ecosystem

- **Challenges and Limits of ML in EDA are real**
  - Optimization QOR
  - Data
  - Scalability
  - Generalization
  - Validation
  - Cost
Messages

• **EDA = optimization + automation = the main levers for scaling**

• **Closed AI/ML platforms** academia must enable itself
  - Action: curated data, domain knowledge

• **AI/ML is difficult** – optimization QOR, data needs, ML ops
  - Action: baselines, replication in the open
  - Action: high-value target selection
Agenda

• EDA and Scaling
• AI and EDA: Looking Back
• Innovation
Innovation

“implementation of creative ideas in an economic setting”

• Not the same as creativity or invention

• Requires innovators (people)

• Requires the right conditions
Disruptive Innovation  
Christensen, *The Innovator’s Dilemma*, 1987

- **Two entry points:**
  - Low-end markets (needs are exceeded)
    
    *I don’t need a 3nm-capable P&R tool*
  - New markets (needs are unserved)
    
    *I need a 3D chiplet integration planning tool*

- **Trajectory of disruption**
  - Initial package of attributes is not valued by existing customers
  - But, the attributes that existing customers value improve rapidly …

“The next big thing always starts out dismissed as a toy”

– Chris Dixon, 2010
Agenda

• EDA and Scaling
• AI and EDA: Looking Back
• Innovation
• Paths Forward
  • Infrastructure for learning
60 Years Ago: 1st SHARE DA Workshop, 1964

SHARE = Society to Help Avoid Redundant Effort
AI/ML for Chip Design: Infrastructure

OpenROAD

CircuitOps: ML-friendly data representation format within OpenROAD

Easy application of ML training within OpenROAD interpreter

ML/RL algorithms integrated within OpenROAD

CircuitOps and OpenROAD: Unleashing ML EDA for Research and Education

VTS-2024 paper

OpenROAD and CircuitOps: Infrastructure for ML EDA Research and Education

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

Vidya A. Chhabria1, Wenjing Jiang2, Andrew B. Kahng3, Rongjian Liang4, Haining Ren1, Sachin S. Sapatnekar2, and Bing-Yue Wu4
1Arizona State University; 2University of Minnesota; 3University of California, San Diego; 4NVIDIA Corporation

Andrew B. Kahng, 240625 IEEE CEDA Distinguished Lecture
SLICE: Shared Infrastructure for ML EDA

- MLCAD-2023 invited talk, Prof. Jiang Hu, Texas A&M University.
- SLICE website: A one-stop shop for ML EDA infrastructure, with pointers to datasets, EDA tool flows, contests, and proxy PDKs.
Takeaways

• EDA = optimization + automation holds main levers for scaling

• Closed AI/ML silos/platforms ☐ academia must enable itself
  • Action: curated data + domain knowledge, culture (rewards, badges, …)

• AI/ML in EDA is difficult: optimization QOR, data needs, ML ops
  • Action: baselines, replication in the open
  • Action: high-value target selection

• Innovation beyond a “toy”: shared, sustained, patient efforts
  • Action: infrastructure with professional staff, longer-term support horizons
Agenda

• EDA and Scaling
• AI and EDA: Looking Back
• Innovation
• Paths Forward
  • Infrastructure for learning
  • Barriers to data proxies
Barriers Demand Proxies

- If it is not sharable, need a proxy!
Proxy PDK

• If it is not sharable, need a proxy!
• PDK: ASAP7/5 + scaling, autotuning

<table>
<thead>
<tr>
<th>Scaled Param</th>
<th>RVT</th>
<th>LVT</th>
<th>SLVT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delay</td>
<td>-0.40</td>
<td>-0.39</td>
<td>-0.23</td>
</tr>
<tr>
<td>Internal Power</td>
<td>-0.27</td>
<td>-0.28</td>
<td>-0.24</td>
</tr>
<tr>
<td>Setup/Hold Time</td>
<td>-0.08</td>
<td>-0.35</td>
<td>-0.09</td>
</tr>
<tr>
<td>Input Pin Cap</td>
<td>+0.31</td>
<td>+0.46</td>
<td>+0.30</td>
</tr>
<tr>
<td>BEOL Cap</td>
<td></td>
<td></td>
<td>-0.52</td>
</tr>
<tr>
<td>BEOL Res</td>
<td></td>
<td></td>
<td>-0.45</td>
</tr>
</tbody>
</table>

• Power v. Effective CP hockey stick
  • Foundry 7nm in Gray
  • Autotuned ASAP7 in Orange
  • BEOL RC and cell-level tuning params
  • Autotuning (Ray/Tune) achieves ~2.3% loss = MAPE of power, fmax errors at 9 target CP values

Scripts are open-sourced in IEEE CEDA DATC RDF-2023
https://github.com/ieee-ceda-datc/RDF-2023
Proxy Enablement

- If it is not sharable, need a **proxy**!
- **PDK:** ASAP7/5 + scaling, autotuning
- **Enablement:** PROBE3.0++

Scripts are open-sourced:
https://github.com/ABKGroup/PROBE3.0
https://github.com/ieee-ceda-datc/RDF-2023
Proxy Designs

• If it is not sharable, need a proxy!
  • PDK: ASAP7/5 + scaling, autotuning
  • Enablement: PROBE3.0++
  • Designs: PULP, Chipyard, ANG

• Real chip design data: sparse, expensive
  □ Artificial data can save cost, improve quality

• IDEA: small artificial netlist ~ “mini-brain”
  Cf. high-throughput drug discovery
  □  10x smaller □ flow is 15-20x faster
  Large target design (288k instances)


Artificial design (28k instances)
Artificial Netlist for PPA Matching

- Artificial netlist that matches relevant PPA behavior of target netlist stays clear of IP issues, helps STCO/DTCO exploration at scale.
- Current focus: Size reduction + “ANG2.0” + Autotuning
- Faster technology evaluation: 10X smaller artificial design → 15-20X smaller P&R runtime
- Research on generator knobs, autotuning loss functions, validations …
- Continuum: {ROs, Critical Path models} \{mini-brains\} \{Arm CPUs, massive DTCO/STCO $$\}$
Artificial Netlist for Flow Exploration

- High value: finding ‘optimal’ flow hyperparameters
  - Big, real design incurs substantial costs: TAT, compute+licenses, human effort
- Freedom = Use a small artificial netlist (*mini-brain*) to find an ‘optimal’ flow for a target large netlist (*big-brain*)
- Can do this IF the Pareto front of flow options for the *mini-brain* correspond to those of the *big-brain*

**Pareto Front**

- **Effective Clock Period [ps]**
- **Total Power [mW]**

---

**mini-brain P&R Flow Exploration**

- **Effective Clock Period [ps]**
- **Total Power [mW]**

**big-brain P&R Flow Exploration**

- **Effective Clock Period [ps]**
- **Total Power [mW]**
Proxy Tools

• If it is not sharable, need a proxy!
  • PDK: ASAP7/5 + scaling, autotuning
  • Enablement: PROBE3.0++
  • Designs: PULP, Chipyard, ANG
  • Tools: OpenROAD, iEDA, Verilator …

• Clarity: Leading edge and baselines become visible and well-defined
  • Commercial EDA license terms: tools cannot be benchmarked

• Better science: Advances are reproducible and reusable
  • Avoid controversy that wastes time and energy → mature research culture

• Efficiency and velocity: Less reinventing of wheels → field advances faster and is more attractive
  • Students today waste months on reconstruction, reimplementation
Zeitgeist of Openness: Worldwide Investment

iPD: An Open-source Intelligent Physical Design Toolchain

iEDA: An Open-source infrastructure of EDA

“Open-source is not a goal but a way”

https://www.nsfc.gov.cn/publish/portal0/tab434/info92532.htm China National NSF program announcement on EDA for chiplets (April 2024): 8 topics specify open-sourcing of results!
Zeitgeist of Openness: EU Open Letter

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, École Polytechnique Fédérale de Lausanne (EPFL), 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 (499)

https://open-source-eda-letter.eu/
Open-Source EDA Birds-of-a-Feather Session

at DAC 2024

Tuesday, June 25, 2024 6:30pm-9:30pm, Moscone West, Room 3001

The “Open-Source EDA, Data and Collaboration Summit” Birds-of-a-Feather session is the fifth in a series that began with the DAC 2018, 2019, 2022 and 2023 Open-Source Academic EDA Software Birds-of-a-Feather sessions. These sessions include a number of participants who also attend the WOSET workshop. This session at DAC 2024 serves as an informal meeting point for anyone who would like to hear or share ideas or latest updates on the following topics:

1. The global ecosystem and landscape of open-source EDA tools and industrial usage
2. The use of open-source EDA tools for Education and Workforce Development
3. Applications and Data for AI/ML-boosted EDA and design
4. Benchmarks, benchmarking and research practices as enabled by open source
5. Community goals, principles and practices for the worldwide open-source EDA comm
Takeaways

• EDA = optimization + automation holds main levers for scaling

• Closed AI/ML silos/platforms → academia must enable itself
  • Action: curated data + domain knowledge, culture (rewards, badges, …)

• AI/ML in EDA is difficult: optimization QOR, data needs, ML ops
  • Action: baselines, replication in the open
  • Action: high-value target selection

• Innovation beyond a “toy”: shared, sustained, patient efforts
  • Action: infrastructure with professional staff, longer-term support horizons

• Proxies have gaps – need coordination to close these!
  • Action: Design enablement: PDK scalers, foundation IP, calibrations
  • Action: Design tools: formal and physical verifications; DFT; HLS-PS-LS
  • Action: Stackable Multiphysics, solvers, compact models !!!
Agenda

• EDA and Scaling
• AI and EDA: Looking Back
• Innovation
• Paths Forward
  • Infrastructure for learning
  • Barriers to data proxies
  • Renewed attacks on optimization
IC Design and EDA = Optimization

Location protected by TCS!
Location protected by EULA!

Area (Power)

Clock Period (1/fmax)

Open-source / Proxy EDA
Cloud, GPU, AI/ML, Transparency, Benchmarking, …
+ Innovation Velocity

Commercial EDA

“Optimal”
 Optimization: Faster, Better **and** Cheaper ?!??

- “Faster, Better, Cheaper – pick any two” (it’s the law !)

- **Question:** Can open-source EDA give us **all three** at once?

---

**Diagram:**

- Faster → Better → Cheaper
- Samples ↑
- Operational NRE ↓ Design iterations, QOR ↑
- Attempts ↓ QOR, value ↑
- Clustering, Algorithms, GPU-/Cloud-native
- Open source, Community
- AI/ML, Algorithms, “Heavy” optimizers

---
1. Chaos: GF12 AES (Synthesis)

Variation of Effective CP: **3.4%**

Variation of Leakage Power: **38%**

Variation of Cell Area: **11%**

Variation of Total Power: **11%**
Chaos: GF12 AES (Place-and-Route)

Input netlists are synthesized with SDC clock periods of 278, 279, 280, 281, and 282ps

Variation of Effective CP: **15%**

Variation of Wirelength: **9%**

Variation of Total Cell Area: **9%**

Variation of Total Power: **9%**

Target Clock Period (ps)

Effective Clock Period (ps)

Effective Clock Period (ps)

Total Power (mW)
2. Early Design Space Exploration (Arch, RTL)

- Can we better explore architecture, RTL, and SoC floorplan design spaces?
  - Ideal: ultra-fast, yet match actual implementation
- **Hier-RTLMP** (src/mpl2 in OpenROAD): RTL- and dataflow-driven, human expert-like results ([arXiv])
Hier-RTLMP vs. Commercial Macro Placer

- **TABLA01 (GF12)**: 760 macros

<table>
<thead>
<tr>
<th>Macro Placer</th>
<th>Std Cell Area ($mm^2$)</th>
<th>Power (mW)</th>
<th>WNS (ns)</th>
<th>TNS (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hier-RTLMP</strong></td>
<td>0.160</td>
<td>640</td>
<td>-0.085</td>
<td>-0.417</td>
</tr>
<tr>
<td><strong>Comm</strong></td>
<td>0.165</td>
<td>689</td>
<td>-0.370</td>
<td>-92.246</td>
</tr>
</tbody>
</table>
3. Dataflow-Aware GPU-Accelerated RePLAce

<table>
<thead>
<tr>
<th>Global Placer</th>
<th>WL</th>
<th>Power</th>
<th>WNS</th>
<th>TNS</th>
<th>GP (s)</th>
<th>TAT (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RePLAce</td>
<td>1.00</td>
<td>1.00</td>
<td>-0.123</td>
<td>-108.15</td>
<td>387</td>
<td>653</td>
</tr>
<tr>
<td>DREAMPlace</td>
<td>0.92</td>
<td>0.98</td>
<td>-0.023</td>
<td>-2.623</td>
<td>61</td>
<td>88</td>
</tr>
<tr>
<td>DG-RePLAce</td>
<td>0.90</td>
<td>0.97</td>
<td>-0.014</td>
<td>-0.078</td>
<td>32</td>
<td>200</td>
</tr>
</tbody>
</table>

Testcase: BlackParrot RISC-V (Quad-Core) (evaluator: INVS 21.1) (827K stdcells, 196 macros in GF12LP)
Speed Enables **Autotuning** *(NVIDIA AutoDMP)*

### Step 1: Specify hyperparameters

**Hyperparameters (specified in configspace.json)**
- coarsening_ratio: range = [6, 20], type = int
- max_num_level: range = [1, 2], type = int
- virtual_iter: range = [1, 8], type = int
- num_hops: range = [1, 8], type = int
- halo_width: range = [1.0, 3.0], type = float
- target_density: range = [0.5, 0.8], type = float

### Step 2: MOTPE Bayesian Opt tuner

### Step 3: Run INVS P&R for Pareto candidates

<table>
<thead>
<tr>
<th>RUN_ID</th>
<th>WL</th>
<th>Power</th>
<th>WNS</th>
<th>TNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>default</td>
<td>0.90</td>
<td>0.972</td>
<td>-0.014</td>
<td>-0.078</td>
</tr>
<tr>
<td>14</td>
<td>0.86</td>
<td>0.967</td>
<td>-0.002</td>
<td>-0.007</td>
</tr>
<tr>
<td>17</td>
<td>0.85</td>
<td>0.971</td>
<td>-0.014</td>
<td>-1.048</td>
</tr>
<tr>
<td>18</td>
<td>0.86</td>
<td>0.968</td>
<td>-0.012</td>
<td>-0.216</td>
</tr>
<tr>
<td>26</td>
<td>0.85</td>
<td>0.969</td>
<td>-0.027</td>
<td>-1.794</td>
</tr>
<tr>
<td>27</td>
<td>0.86</td>
<td>0.970</td>
<td>-0.007</td>
<td>-0.139</td>
</tr>
</tbody>
</table>

Post-route layout of RUN_ID = 14

**Demo:** swerv_wrapper (NG45)

A. B. Kahng, 240625 IEEE CEDA Distinguished Lecture 45
4. More Data in Same Walltime: “Tomography”

- Idea: Many images / views can be taken in unit time
- Congestion report using Innovus eGR: $< 1$ second
  - Detailed Routing runtime: 1.5 hours
- Example of “ML magic” at interstices (place $\square$ route)
Routing Blockage Generation: Human vs. ML

- Human runs DRoute, ECO Route many times
- Tomography uses postCTS eGR features
  - **DRVNet** model predicts layer-wise routing blockage regions
  - **BlkgComp** model predicts better blockage configuration
  - Better results in much less time
Human vs. Tomography

- Layer-wise blockages generated using Tomography
- Similar or better than Human while not using any post-route result
  - Human takes at least one extra iteration of post-route result
- Improves routed wirelength and power in GF12

<table>
<thead>
<tr>
<th></th>
<th>NOVA-NG45</th>
<th>CA53-GF12</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Blkg</td>
<td>Human Layer-wise</td>
</tr>
<tr>
<td>DRC</td>
<td>2,003</td>
<td>898</td>
</tr>
<tr>
<td>WL [mm]</td>
<td>3,874</td>
<td>4,016</td>
</tr>
<tr>
<td>WNS [ns]</td>
<td>-0.634</td>
<td>-0.752</td>
</tr>
<tr>
<td>TNS [ns]</td>
<td>-419</td>
<td>-512</td>
</tr>
<tr>
<td>Power [mW]</td>
<td>257.3</td>
<td>259.8</td>
</tr>
</tbody>
</table>

Human vs. Tomography

A. B. Kahng, 240625 IEEE CEDA Distinguished Lecture
5. Openness Accelerates Progress!

Partitioning + Floorplanning and Beyond!

Placement + GPU + Autotuning

- AutoDMP (2023)
- DREAMPlace (2019)
- RePlAce (2018)
- Die-to-Die 3D placement (2024)
- DG-RePlAce (2024)
- ePlace (2015)
- GPU-GR + DG-RePlAce (2024++)
- 3D Placement (2024++)
- Hier-RTLMP (2023)
- 3D placement (2024++)
- Multi-FPGA CAD flows (2024++)
- BlobPlace (2024)
- Hier-RTLMP (2023)
- 2.5D / chiplets (2024++)
- TritonPart (2023)
- RTL-MP (2022)
- K-SpecPart (2023)
- SpecPart (2022)
- TritonPart (2023)
- Multi-FPGA CAD flows (2024++)
- BlobPlace (2024)
- Hier-RTLMP (2023)
- 2.5D / chiplets (2024++)
- TritonPart (2023)
- RTL-MP (2022)

Placement + GPU + Autotuning

- AutoDMP + DG-RePlAce (2024++)
- 3D Placement (2024++)
- GPU-GR + DG-RePlAce (2024++)
- 3D placement (2024++)
- Multi-FPGA CAD flows (2024++)
- BlobPlace (2024)
- Hier-RTLMP (2023)
- 2.5D / chiplets (2024++)

Partitioning + Floorplanning and Beyond!

- AutoDMP (2023)
- DREAMPlace (2019)
- RePlAce (2018)
- Die-to-Die 3D placement (2024)
- DG-RePlAce (2024)
- ePlace (2015)
- GPU-GR + DG-RePlAce (2024++)
- 3D Placement (2024++)
- Hier-RTLMP (2023)
- 3D placement (2024++)
- Multi-FPGA CAD flows (2024++)
- BlobPlace (2024)
- Hier-RTLMP (2023)
- 2.5D / chiplets (2024++)

- AutoDMP + DG-RePlAce (2024++)
- 3D Placement (2024++)
- GPU-GR + DG-RePlAce (2024++)
- 3D placement (2024++)
- Multi-FPGA CAD flows (2024++)
- BlobPlace (2024)
- Hier-RTLMP (2023)
- 2.5D / chiplets (2024++)

- AutoDMP + DG-RePlAce (2024++)
- 3D Placement (2024++)
- GPU-GR + DG-RePlAce (2024++)
- 3D placement (2024++)
- Multi-FPGA CAD flows (2024++)
- BlobPlace (2024)
- Hier-RTLMP (2023)
- 2.5D / chiplets (2024++)
Takeaways

- **EDA = optimization + automation** holds main levers for scaling

- Closed AI/ML silos/platforms → academia must enable itself
  - Action: curated data + domain knowledge, culture (rewards, badges, …)

- AI/ML in EDA is **difficult**: optimization QOR, data needs, ML ops
  - Action: baselines, replication in the open
  - Action: high-value target selection

- Innovation beyond a “toy”: **shared**, sustained, patient efforts
  - Action: infrastructure with professional staff, longer-term support horizons

- Proxies have gaps – need coordination to close these!
  - Action: Design enablement: PDK scalers, foundation IP, calibrations
  - Action: Design tools: formal and physical verifications; DFT; HLS-PS-LS
  - Action: Stackable multiphysics compact models

- **Optimization**: many rich vistas + can race faster in the open
Takeaways

- **EDA = optimization + automation** holds main levers for scaling
- Closed AI/ML silos/platforms must academia must enable itself
  - Action: curated data + domain knowledge, culture (rewards, badges, …)
- AI/ML in EDA is **difficult**: optimization QOR, data needs, ML ops
  - Action: baselines, replication in the open
  - Action: high-value target selection
- Innovation beyond a “toy”: **shared**, sustained, patient efforts
  - Action: infrastructure with professional staff, longer-term support horizons
- Proxies have gaps – need coordination to close these!
  - Action: Design enablement: PDK scalers, foundation IP, calibrations
  - Action: Design tools: formal and physical verifications; DFT; HLS-PS-LS
  - Action: Stackable multiphysics compact models
- Optimization: many rich vistas + can race faster in the open
LLM Agents

- Long-running, increasingly complex tasks
- Code generation, planning, reasoning, tool use
- Multimodal
- Learning from feedback / playbooks
Agenda

• EDA and Scaling
• AI and EDA
  • Looking back to 2018
  • Looking back from 2030

Where Are We Headed?

• “ML-enabled DA”, “EDA2.0” are the next waypoints in the 60+ year trajectory of design automation and scaling of design
  • + “cloud-native”, “open-source”, “optimization-centric”?

• Enabled by learning and optimization technologies

• Some elements arrive before others
  • E.g., autotuning and sampling before fair benchmarking

• Can we deliver on this picture within 5-10 years?
An Observation About “Chips Acts”

Training the Future Workforce

Better onramps

No friction (Ex.: “skin” creation)

Accelerate Innovation

Same Cadence Tcl script!

Open-Source EDA

Logos and images are property of respective owners. The “skin” script was written and developed by students at UCSD; however, the underlying commands and reports are copyrighted by Cadence. We thank Cadence for granting permission to share our research to help promote and foster the next generation of innovators.

A. B. Kahng, 240625 IEEE CEDA Distinguished Lecture
OpenROAD and LLMs: EDA Corpus

**EDA Corpus**: A Large Language Model Dataset for Enhanced Interaction with OpenROAD [https://github.com/OpenROAD-Assistant/EDA-Corpus](https://github.com/OpenROAD-Assistant/EDA-Corpus)

- **1500+** datapoints of prompt-script and question-answer pairs

**Examples**:

**Prompt:**
Give me the sample code to get a list of all the output pins of instances in a design using OpenROAD Python API.

**Response:**

```python
# Get the current design block from the OpenROAD database
block = ord.get_db_block()
# Retrieve all instances present in the design block
insts = block.getInsts()
output_pins = []
# Iterate over each instance in the design block
for inst in insts:
    # Retrieve the pins of the instance
    pins = inst.getITerms()
    # Iterate over each pin of the instance
    for pin in pins:
        # Check if the pin is an output signal
        if pin.isOutputSignal():
            # Append the output pin to the list
            output_pins.append(pin)

return output_pins
```

**Prompt:**
What is PDNGEN?

**Response:**
The PDNGEN module (pdn) in OpenROAD aims to simplify the process of adding a power grid into a floorplan. A user can specify a small set of power grid policies to be applied to the design, such as the layers to use, stripe width, and spacing, and then pdn will generate the actual metal straps. Power grid policies can be defined over both the standard cell area and areas occupied by macros.

<table>
<thead>
<tr>
<th>Model</th>
<th>Prompt-script</th>
<th>Question-answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChatGPT3.5</td>
<td>0 True 50 False</td>
<td>4 True 17 Partial 9 False</td>
</tr>
<tr>
<td>ChatGPT4</td>
<td>3 True 47 False</td>
<td>6 True 17 Partial 7 False</td>
</tr>
<tr>
<td>ChatGPT3.5 fine-tuned with prompt-script</td>
<td>22 True 28 False</td>
<td>- - -</td>
</tr>
<tr>
<td>ChatGPT3.5 fine-tuned with question-answer</td>
<td>- -</td>
<td>26 True 3 Partial 1 False</td>
</tr>
</tbody>
</table>

Fine-tuning ChatGPT3.5 with EDA Corpus
2018-2020: Standardized Messages

Refocusing on Optimization

- Reality of Optimization
  - Better, faster, cheaper – pick any two
- IC EDA: want all three at once
  - “Need an answer overnight”
  - “Unfortunately, the runtime of …”
- But the world has changed …
  - Automation, cloud, ML, GenAI, system design, multi-physics, quantifiable assurance, ...

“Machine Learning in EDA”: The Who

<table>
<thead>
<tr>
<th></th>
<th>EDA USER</th>
<th>EDA SUPPLIER</th>
<th>ACADEMIA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Owns Design+Flow</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Can Collect Data</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Owns Tool Details</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>ML “inside” EDA</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>ML “around” EDA</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

- Academic research: All of the pieces … or None?

- Academic research leading edge: Pioneering, unifying force … or Anklebiter?

- Challenges of talent, funding, techno/design relevance, feasible scale, …
- Hard to not question academic research relevance and impact
2020: Whither “Machine Learning in EDA”?

“End-to-end, Deep RL”

Moving the Fulcrum:
The Bitter Lesson
EDA Users (“magicians”)
EDA Suppliers (“better wands”)
Point/incremental ML research
Scaling slowdown, better pathfinding
EDA Research (e.g., Opt; U+S+A)
Open-source EDA

“Episodic ML”

Naïve
ABKGroup history
DAC/ICCAD, CDNS/SNPS, ...
Everyone rolls their own
Low-hanging fruits harvested
Silos have limits (by definition)

• Who is going to move the fulcrum, in which directions?
  • What changes if/when “End to end, Deep RL” wins?
• How should academic research contribute to ML in EDA?
THANK YOU!

abk@ucsd.edu

Acknowledgments: Many thanks to Sayak Kundu, Bodhisatta Pramanik, Zhiang Wang and Dooseok Yoon for their help with these slides. Past 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.