AI/ML, Optimization and EDA in TILOS, an NSF National AI Research Institute

Andrew B. Kahng, UC San Diego

The Institute for Learning-enabled Optimization at Scale tilos.ai







• What is TILOS?





Renn TEXAS Yale

NSF National AI Research Institute for Advances in Optimization

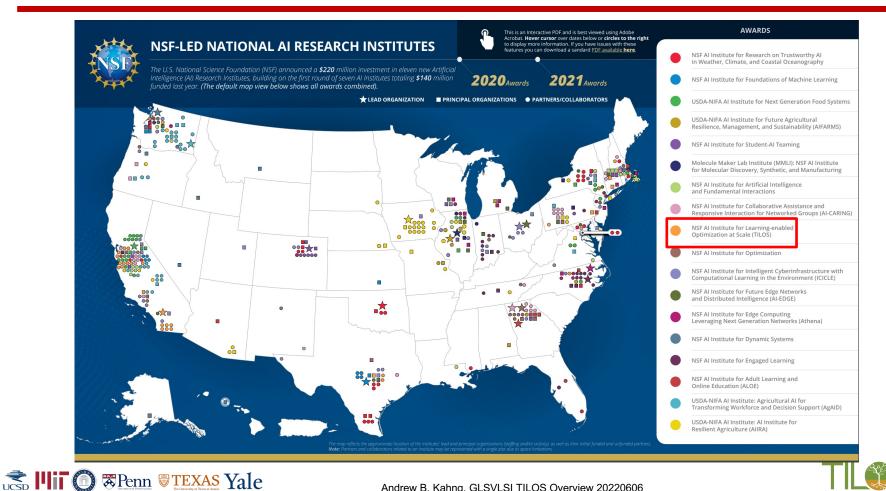
Mission: make impossible optimizations possible, at scale and in practice.

5-year grant, \$20M total funding from NSF (started November 1st !) Partial support is from Intel Corporation

UCSD is the lead institution



What is the National AI Research Institutes Program?



TILOS: Fishburn and Dunlop (ICCAD85)

TILOS: A POSYNOMIAL PROGRAMMING APPROACH TO TRANSISTOR SIZING

J. P. Fishburn¹ and A. E. Dunlop² AT&T Bell Laboratories Murray Hill, New Jersey 07974

Abstract

TEXAS Yale

A new transistor sizing algorithm, which couples synchronous timing analysis with convex optimization techniques, is presented. Let A be the sum of transistor sizes, T the longest delay through the circuit, and K a positive constant. Using a distributed RC model, each of the following three programs is shown to be convex: 1) Minimize A subject to T < K. 2) Minimize T subject to A < K. 3) Minimize AT^{K} . The convex equations describing T are a particular class of functions called posynomials. Convex programs have many pleasant properties, and chief among these is the fact that any point found to be locally optimal is certain to be globally optimal. TILOS (TImed LOgic Synthesizer) is a program that sizes transistors in CMOS circuits. Preliminary results of TILOS's transistor sizing algorithm are presented.

Introduction 1.

Given a synchronous MOS circuit of the form shown in Figure 1 with N transistors of sizes (channel widths) $x_1, x_2, ..., x_N$, the following question is considered: How can the circuit's performance be improved by adjusting the x? Two figures of merit are of special interest. T is defined to be the minimum



John P. Fishburn, PhD

away peacefully Saturday, April 24, 2021. as "Jack" by his friends and family, was be Muscatine, Iowa on May 21, 1951, the son Susan (Wooldridge) Fishburn.

Iack attended local schools in Muscatine a his undergraduate degree in Mathematics University of Iowa. Jack furthered his edu receiving his PhD from University of Wis Madison where he defended his thesis on algorithms including the parallelization of

Jack met his wife of 35 years, Lynne, after



The Institute for Learning-enabled Optimization at Scale

Optimization: Find a best-possible solution

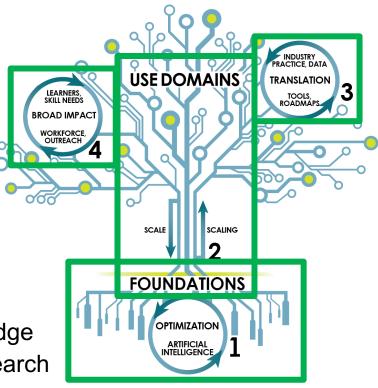
Fundamental challenges: scale and complexity

→ Nexus of AI/ML, optimization, use in practice

Vision: Four "virtuous cycles"

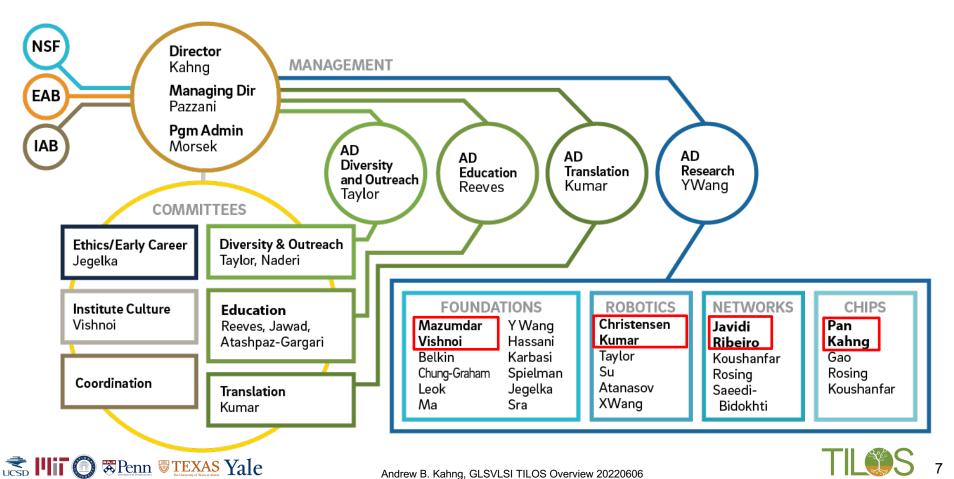
Penn TEXAS Yale

- 1. Foundations: Al and Optimization
- 2. Scaling: Foundations and Use Domains
- 3. Translation: Academia and Industry leading edge
- 4. Broad Impact: Education, Outreach, and Research





Structure

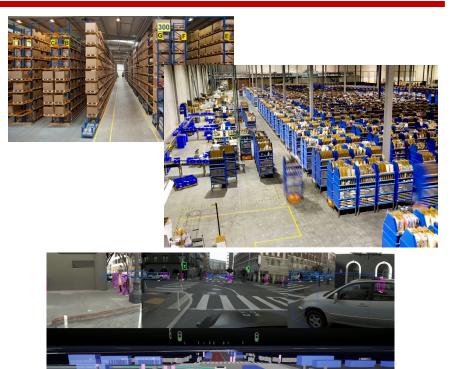


Robotics physical systems in the real world

Challenged by

- Dimensionality
- Structural and Dynamic constraints
- Dynamic world with a need to anticipate changes
- **Optimization target:** multi-robot interaction with these challenges
 - Reduce traffic congestion by 25+%
 - Perform efficient real-world learning
 - Deploy in regular homes

TEXAS Yale





Communication Networks the infrastructure of the information age

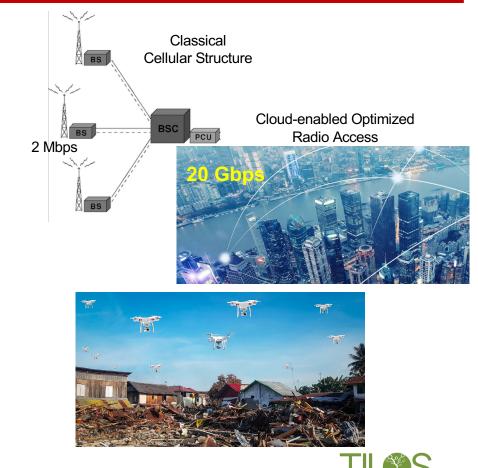
Challenged by

- Decentralized management/control
- Multiple design scales
 - Physical laws of signal propagation
 - Ubiquitous global connectivity
- Impossible to sustain
 - Overprovisioning wastes energy

• Optimization target:

TEXAS Yale

- Federated learning/optimization
- Automated (blackbox) optimization
- Integrated representation of physics



Challenged by:

Complexity billions of transistors, stack of abstractions, nanometer physics

Optimization target: 1000x speedups, scalability

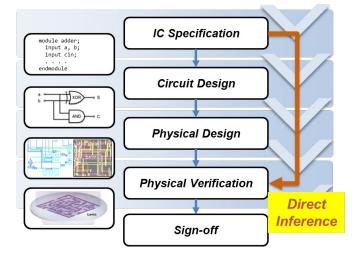
Direct inference of layout

TEXAS Yale

- Verification
- More system objectives "X": X = security, resilience, …

 Image: Chess
 Chip placement

 10²⁰ States
 10⁹⁰⁰⁰ states





Al advances \rightarrow pose new challenges, provide new tools for optimization

Bridging Discrete and Continuous

Distributed, Parallel, and Federated

Optimization on Manifolds

Dynamic Decisions under Uncertainty

Nonconvex Optimization in Deep Learning

New perspectives on classic problems \rightarrow watch this space!

Credit: N. Vishnoi, S. Jegelka, D. Spielman, Y. Wang +





IC Design and Design Automation: Challenges

Challenges

- 1. Scale
- 2. Representation
- 3. Uncertain objectives Multi-stage, Multi-scale Dynamics
- 4. Reliability, Generalization
- 5. Federated, Distributed



Credit: N. Vishnoi, S. Jegelka, D. Spielman, Y. Wang +



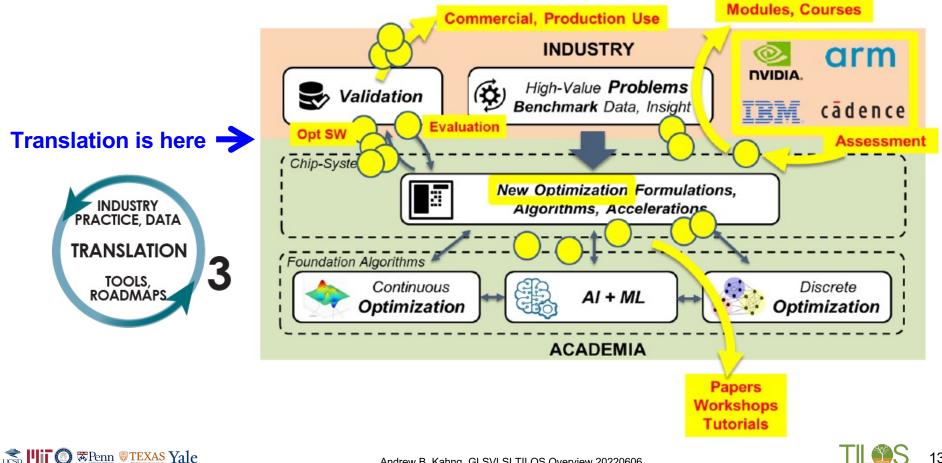
IC Design

Optimization

Problems



"Life Cycle" of Research and Translation



Agenda

- What is TILOS?
- Why TILOS?





Learning, Optimization, Scaling

 "Machine Learning (ML) is the part of AI studying how computer agents can improve their perception, knowledge, thinking, or actions based on experience or data."

Prof. Christopher Manning, Stanford, Sept. 2020 https://hai.stanford.edu/sites/default/files/2020-09/AI-Definitions-HAI.pdf

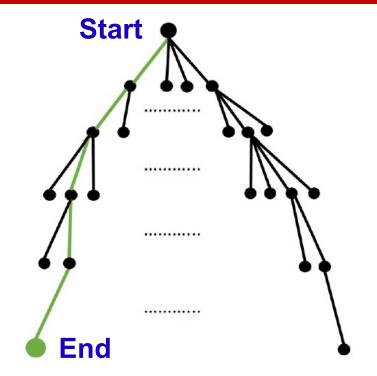
- **Optimization** is the universal quest to *do better*.
- Scaling is what drives all of us.







Challenge: Optimization (IC Design) "Lives in a Box"



Huge space of trajectories: architecture, enablement, IPs, tools, manual fix, ...

Start to End: expensive!

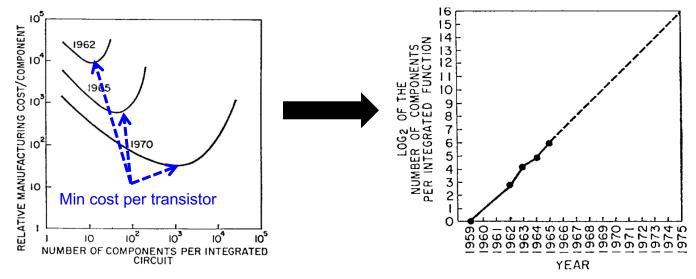
- O(year) for product
- O(weeks) for SP&R and Opt
- Goal: best possible End
- Constraint: stay in "Box"
 - {compute}
 - X {licenses}
 - X {people}
 - X {weeks}

UCSD Penn TEXAS Yale



Scaling: Delivers Value

Moore, 1965: "The complexity for minimum component costs has increased at a rate of roughly a factor of two per year"



- Scaling focus: "PPAC" power, performance, area, cost
- Moore's Law is a law of cost reduction 1% = 1 week
 - Corollary: greater reach of integration, more innovation within reach

Score Score



ML for EDA and IC Design: What

Predict

- Will RouteOpt finish with clean signoff, <1000 DRVs by tomorrow night?
- Classify
 - Out of these 50 floorplans + budgets, which 3 should go into trial SP&R?

Estimate

• How many hold buffers will tool eventually add into this post-CTS layout?

Guide / advise

- What P&R tool setup/script will obtain the best QOR within next 36 hours?
- More broadly: answer any question that is difficult for humans
 - Google Brain, 2020: "super-human macro placement" on arXiv
 - Overarching: "intelligent flow", "automated super-human expertise"
- More directly: regressions and image classifications (LSF, litho)



ML for EDA and IC Design: Why

- A. You need models to have predictions
- **B.** You need predictions to leverage in exploration
- C. What you can't predict, you guardband
- **D.** What you don't explore, you leave on the table
- E. C and D are bad for product quality and schedule
- We are in an Era of Optimization
 - Look for ML to win quality, schedule, cost
 - E.g., reduce analysis runtime, miscorrelation

→ We hope that ML will bring Scaling







4 Aspects of ML for EDA and IC Design

1. Mechanization and Automation

Create super-human robot engineers

2. Orchestration of Search and Optimization

Optimize the use of N robot engineers

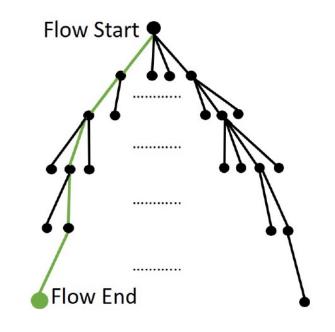
3. Pruning via Predictors, Models

Predict design-specific tool outcomes

Prune "doomed runs"

4. From Reinforcement Learning to Intelligence

Target: "MLDA", "self-driving tools and flows", "superhuman"



Huge space of tool, command, option trajectories through design flow

UCSD HIT O TEXAS Yale



Generic Need: Predict Doomed Runs

- Example: progression of DRC violations in commercial router
- Simple strategy: track and project key metrics as time series
- Example method: use Markov decision process (MDP): "GO" vs. "STOP" strategy card to terminate "doomed runs" early

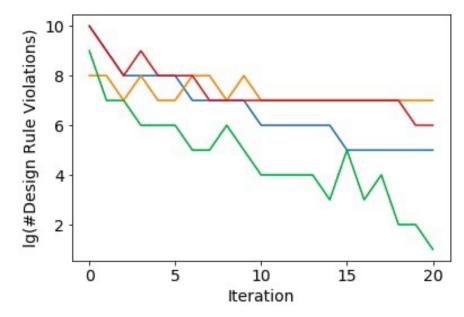


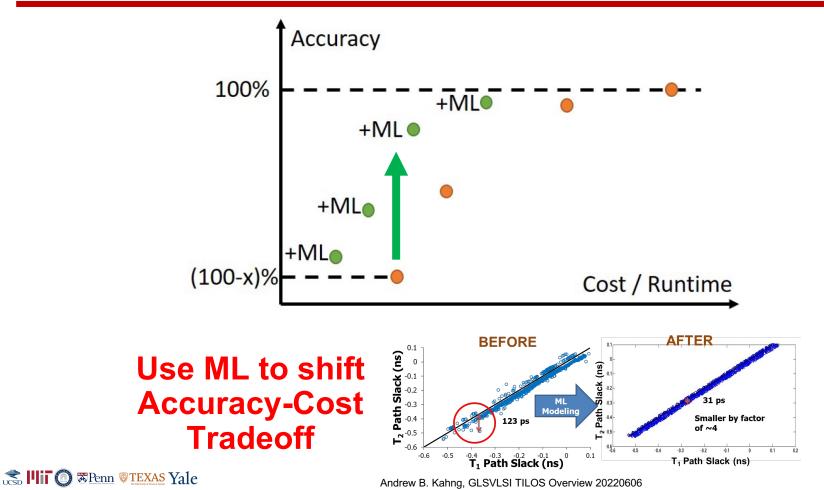


Figure from

link link link

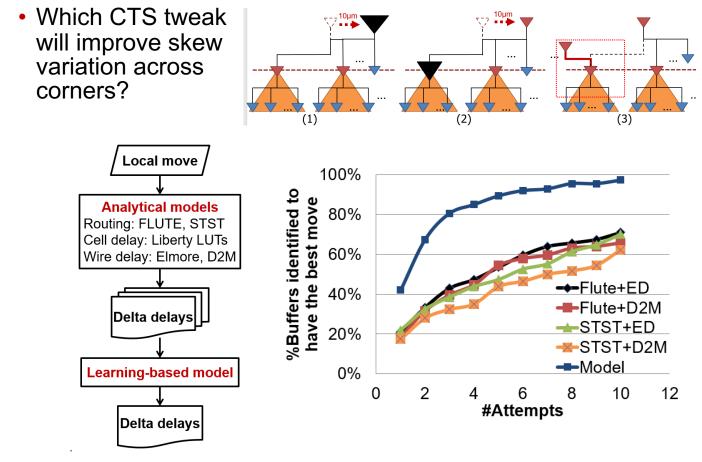
DATE14 <u>link</u> SLIP15 <u>link</u>

Generic Need: Shift Accuracy-Cost Tradeoff





Generic Need: Model-Guided Optimization



TEXAS Yale



Agenda

- What is TILOS?
- Why TILOS?
- TILOS Goals
 Refocusing: EDA is Optimization





EDA is Optimization

EDA is about optimizations and algorithms

- High stakes: performance, power, design closure
- Discrete, combinatorial formulations at huge scale
 - Optimizations: ILP, MCF, QAP, SAT/SMT, LR, ...
 - Algorithms: min-cost flow, high-dim DP, ...
- But need an answer overnight
- Reality under the hood: metaheuristics
 - Annealing, multi-start, PSO, NGSA-II, ripup-reroute, greed, ...
 - Convenient objectives, customer-/tech-specific tuning, ...
 - ... which comes at a cost

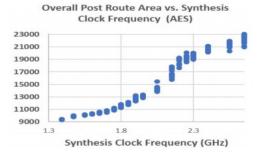


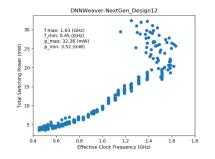
Never Enough Time: Heuristics and Chaos

"CAD tools are chaos machines"

-- Ward Vercruysse, Sun UltraSPARC III CAD manager, Physical Design Workshop 1996

- Push harder on a tool that is made up of heuristics stacked on top of heuristics → result becomes less predictable
- Change initial conditions slightly \rightarrow outcome can change a lot





Recurring theme in my group...

TEXAS Yale

- ISQED02 (Noise), ISQED10 (Chaos)
- DAC18 WIP (Multi-Armed Bandits)

Measurement of Inherent Noise in EDA Tools Andrew B. Kahng[†] and Stefanus Mantik

Methodology From Chaos in IC Implementation

[†] UCSD CSE and ECE Departments, La Jolla, CA 92093-0114 UCLA Computer Science Department, Los Angeles, CA 90095-1596 abk@ucsd.edu, stefanus@cs.ucla.edu Kwangok Jeong¹ and Andrew B. Kahng^{1,2} ¹ECE and ²CSE Departments, University of California at San Diego, La Jolla, CA, USA kieone@vlsicad.usca.dcu, abk@cs.uscd.edu

A No-Human-in-the-Loop Methodology Toward Optimal Utilization of EDA Tools and Flows



Andrew B. Kahng, Shriram Kumar and Tushar Shah, UCSD

N Kumar and Tushar Shan, UCSD Room CHIPS TO STSTEM

Simultaneously maximize reward
IC Design Problem
Each "arm" = a target frequency.

distributions of outcomes)

- Each "pull" = a run of the tool flow
- Three well-known sampling strategies
- Thompson Sampling

using T "pulls" (iterations)Involves an explore-exploit tradeoff

- <u>€</u>-Greedy Sampling
- Softmax Sampling
- (+ Naïve: uniformly sample from all available arms)

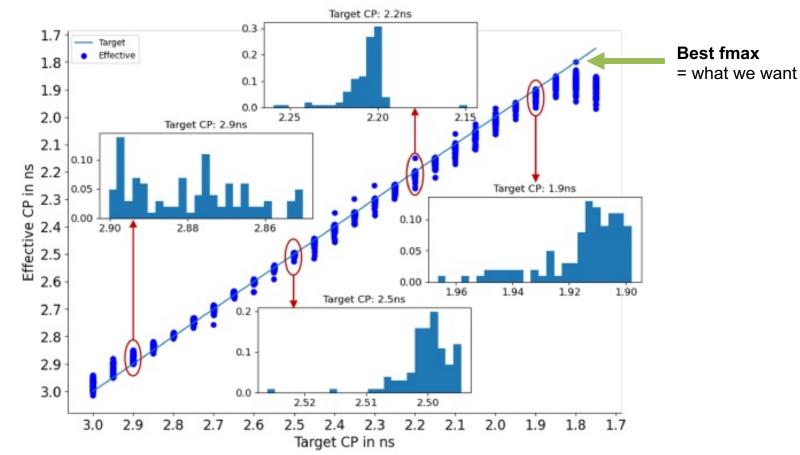
IC Design = "Multi-Armed Bandit" Problem • Multi Armed Bandit Problem (MAB): Given a slot

machine with N arms, maximize total reward obtained

Draw samples to learn model parameters (e.g.,



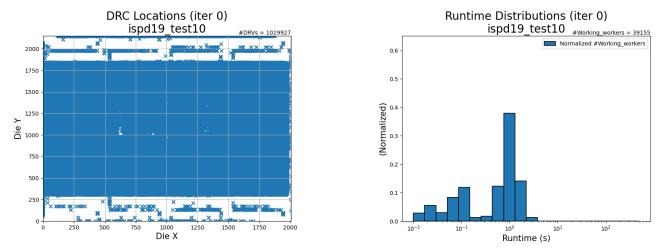
A Closer Look: "Actual" vs. "Target"







Next: Cloud-Deployed Learning and Prediction !



- ispd19_test10, TritonRoute [drt in OpenROAD]
 - Left: Tiles with DRVs (down to 21 at final iteration #64)
 - Right: Runtimes of workers (up to > 500s in late iterations)
- Prediction
 - Doomed runs? Hotspots, long tails? DRVs from placement, global route?
- Learning
 - Costing strategies for workers? Sampling / multi-start strategies for tiles?

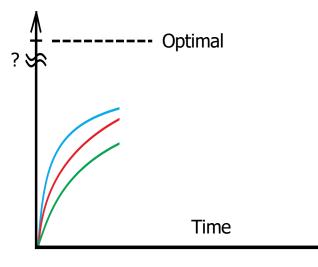
Score Score



Have We Lost Sight of the Suboptimality Gap?

Reality of optimization

- Better, faster, cheaper pick any two
- IC EDA: want all three at once
 - "Unfortunately, the runtime of ..."
- But the world has changed
 - Automation, cloud, ...



Question: If you give your SP&R flow 3 extra days of runtime, would it know how to use this extra time?

Question: If you could run 10,000 copies of your P&R tool at the same time, what QOR improvement **should** you expect?

Seen ©TEXAS Yale



Generic Need: Re-Focus on Suboptimality (to get closer to Optimality!)

Suboptimality ... in what sense?

• Need proper aiming points (which AI can help us learn)

Benchmarking

- "Real" benchmarks in EDA have been obfuscated, incomplete, non-vertical, old...
- "Artificial" benchmarks have tiptoed between realism, known optimal solution quality, scalability ...
- Enough (30+ years of this) is enough \rightarrow find a next level
 - E.g., "Underwriters Laboratories for IC Design Tools"

Comprehending modern compute resources cloud, GPU, accelerators, ...

- EDA Optimization + Learning **naturally** live in the cloud
- Many of today's EDA optimization implementations: "EDA1.0" from the 1980s
 - → Can TILOS help discover new, cloud-scalable "EDA2.0" foundations?

UCSD IIII 🔘 🛪 Penn 🖗 TEXAS Yale



Learning + Optimization = Scaling of EDA

How Learning helps Optimization

- "Modeling and prediction": prune doomed runs early
- "New cost-accuracy tradeoffs in analysis": less guardbanding
- Value: more optimization within the available box of resources
- Center of gravity for most of "ML in/around EDA" so far

How Optimization helps Learning

- Stochastic gradient descent, nonconvex optimization
- Distributed/parallel
- Meta-level: optimization of HW on which learning takes place
- Virtuous cycle that brings **Scaling** (of CAD/EDA and IC design)





Agenda

- What is TILOS?
- Why TILOS?
- TILOS Goals
 - Refocusing: EDA is Optimization
 - New Foundations of ML and Optimization





Al and Optimization: Key Directions to Watch

Al advances \rightarrow pose new challenges, provide new tools for optimization

Bridging Discrete and Continuous

Distributed, Parallel, and Federated

Optimization on Manifolds

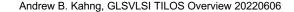
Dynamic Decisions under Uncertainty

Nonconvex Optimization in Deep Learning

New perspectives on classic problems \rightarrow watch this space!

Credit: N. Vishnoi, S. Jegelka, D. Spielman, Y. Wang +

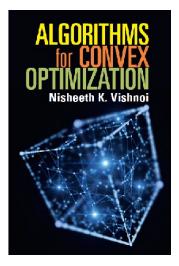


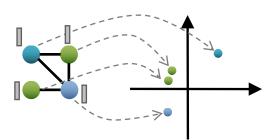




Bridging Discrete and Continuous

- Discrete domains: combinatorial explosion
 - Computational intractability
 - Discrete methods are fragile
- Continuous relaxations can provide robust and fast solutions
 - Continuous methods generalize well: inclusion of continuous properties; less dependence on problem assumptions
- Representation learning for discrete domains to interface continuous methods
 - Graph Neural Networks (GNNs), finding hidden structure
- Leverage discrete structure to speed up continuous methods





Credit: N. Vishnoi, S. Jegelka, D. Spielman, Y. Wang +

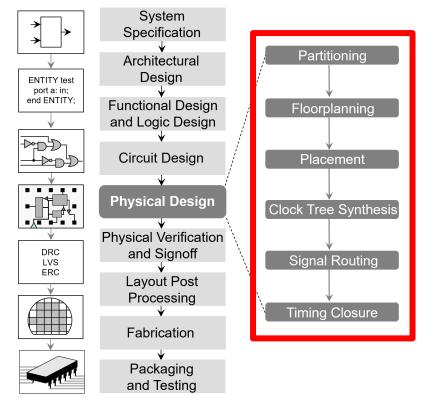




Bridging Discrete and Continuous in IC Design

- Improve solutions to problems of individual layers
 - E.g., partitioning, network flows
- Refine the problems
- Learn surrogate objective functions to smooth the composition of objectives and algorithms
- Representations of discrete objects (e.g., GNNs) facilitate learning parts of solutions

Chip Design Process

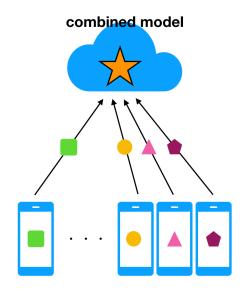


Credit: N. Vishnoi, S. Jegelka, D. Spielman, Y. Wang +



Distributed, Parallel, and Federated

- How we compute, learn and optimize quickly
- Parallelizing second-order methods
- Distributed submodular optimization
- Distributed with communication failures
- Federated
 - Balance communication and computation
 - Maintain privacy and security
- + Partitioning, Clustering, Sparsification, ...



Credit: N. Vishnoi, S. Jegelka, D. Spielman, Y. Wang +



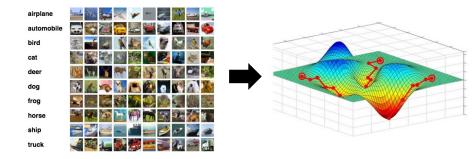
Dynamic Decisions Under Uncertainty Search or jump to ... Pull requests □ ieee-ceda-datc / aspdac-2022-tutorial Public Decision making in unknown and dynamic environments \rightarrow sequential and reinforcement Environment learning Outcome determined by environment or a complicated algorithm \rightarrow actions taken change the future State, Solutions are optimal distributions on actions, Action Reward rather than optimal actions Update the distributions Develop better sampling algorithms Leverage low-dimensional representations of the environment / state AI Credit: N. Vishnoi, S. Jegelka, D. Spielman, Y. Wang +

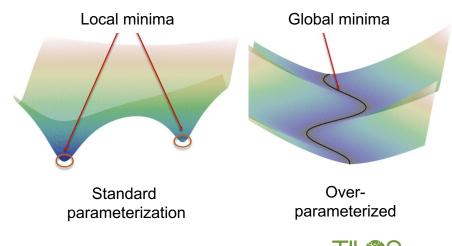


github.com/ieee-ceda-datc/aspdac-2022-tutoria

Nonconvex Optimization in Deep Learning

- Modern model training is not convex !
- Missing understanding of deep network training by nonconvex optimization
- Overparameterization and discovery of global optima
- Convergence of adaptive gradient methods
- Robustness of optimization to noise, errors, corruption





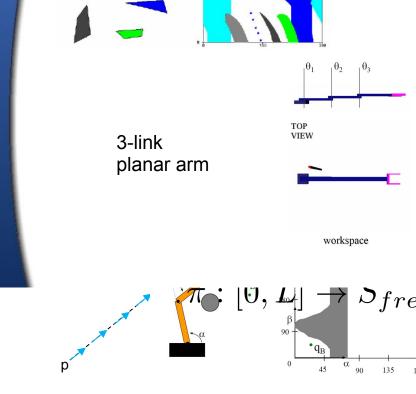
Optimization on Manifolds

• Projections and simplifications of data, representations of problems

• Sampling as optimization on manifolds

Geodesically convex sets

 Algorithms, representation and analysis for singular manifolds





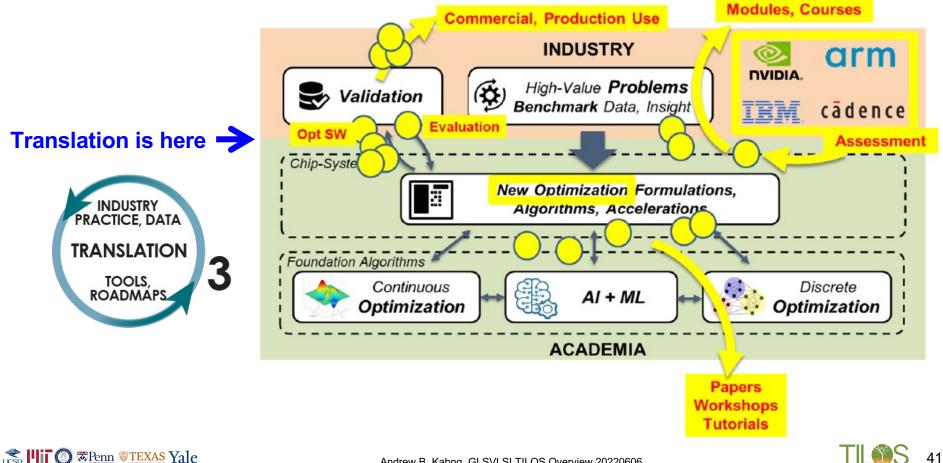
Agenda

- What is TILOS?
- Why TILOS?
- TILOS Goals
 - Refocusing: EDA is Optimization
 - New Foundations of ML and Optimization
 - A Next Level of Translation





"Life Cycle" of Research and Translation

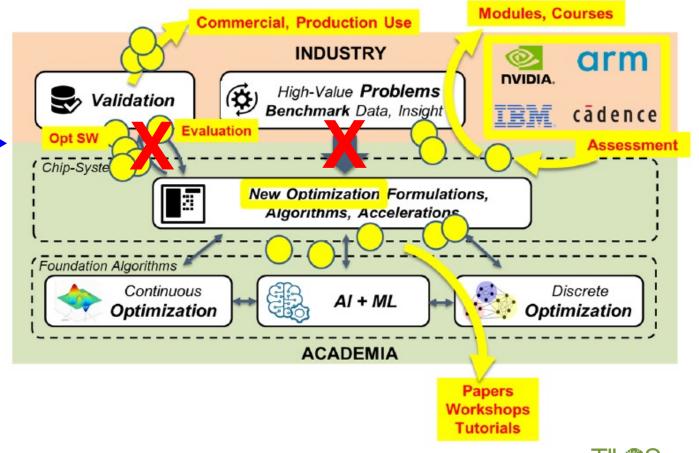


"Life Cycle" of Research and Translation

Translation is here ->

Translation requires solutions to the X's:

- data
- benchmarking
- roadmapping



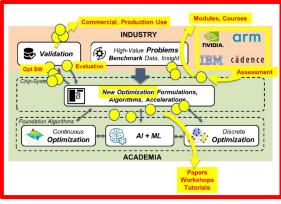
"Third Virtuous Cycle" in the Real World ...

TILOS goals include:

- **Democratization** of research at the leading edge
- **New norms** for transparency, reproducibility, translation
- New norms of benchmarking in high-stakes use domains
- Principled roadmapping to guide investment of time and \$
- = what a NAIRI Institute for Advances in Optimization should aim for !

TEXAS Yale







Example Research Questions on the path to Translation goals for TILOS

 Can't access or expose real data → Generate artificial circuit designs that are indistinguishable from real circuit designs from the perspective of optimizers

[Also: Can we learn from much less real data?] [Also: Can artificial-but-realistic instances help us quantify suboptimality gaps and distributions?]

- Can't access the best optimizers → Model of "strong optimizer" outcomes based on instance attributes and "weak optimizer" outcomes
- Can't reveal sources of data → Develop methods for privacy-preserving anonymization and obfuscation of design and related data
- Benchmarking brings risks of misuse → Develop ethical principles and validations to enable fair benchmarking

[Can TILOS establish principles for reporting and comparison of applied ML and Optimization?]

Can't identify the most crucial learning, optimization goals → Roadmaps + Drivers

S Penn [™]TEXAS Yale

44

Vision: TILOS will pioneer a next level of translation

Data + benchmarking + roadmapping

- Bring industry practitioners and academic researchers closer together
- Bring attention to relevant problems and performance targets
- Problem roadmaps with benchmarks (generators), best-known solutions

Basic research

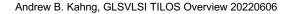
- Many facets of "data virtual reality"
- Anonymization, obfuscation
- Ethics of benchmarking and reporting

Community engagement and change

- Software releases in TILOS organization GitHub, plus impact metrics
- Published mechanisms that democratize research in high-stakes use domains
- Standards of software quality, testing, support
- Industry roundtables \rightarrow core problem formulations + metrics of progress



Seen ©TEXAS Yale





https://github.com/TILOS-AI-Institute -- watch this space

A TILOS-AI-Institute / MacroPlacement Private		
<> Code () Issues 11 Pull requests	O Actions ☐ Projects ① Security	🗠 Insights 🕸 Settings
🐉 main 🗸 🐉 2 branches 🚫 0 tags		Go to file Add file - Code -
abk-tilos Update README.md		4c08f9d 4 hours ago 🕚 107 commits
CodeElements	Update README.md	4 hours ago
Enablements	Updating enablements	8 hours ago
ExperimentalData	Updated directory organization	yesterday
Flows	Adding ariane 136 and ariane133	8 hours ago
Testcases	Updating directory structure	8 hours ago
	Create LICENSE	2 days ago
README.md	06051447	8 hours ago

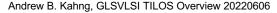
i = README.md

TEXAS Yale

Ø

MacroPlacement

MacroPlacement is an open, transparent effort to provide a public, baseline implementation of Google Brain's Circuit Training (Morpheus) deep RL-based placement method. We provide (1) testcases in open enablements, along with multiple EDA tool flows; (2) implementations of missing or binarized elements of Circuit Training; (3) reproducible example macro placement solutions produced by our implementation; and (4) post-routing results obtained by full completion of the place-and-route flow using both proprietary and open-source tools.





Missing Infrastructure

Data and ML for IC designers

- Model encapsulation and application
- IP, privacy protections to enable model sharing
- Data and ML for EDA tools/flows
 - Standard data model, names, semantics
- Data for ML and Optimization
 - Real designs, Artificial designs and "Eyecharts"
 - Calibration data ("Underwriters Lab")
 - Share the cost of developing big data = grid computing paradigm
 - Year 2000: SETI@Home Year 2020: Tool X on PDK Y with IP Z ?
 - + Challenges and incentives: "Kaggle for ML in IC design"
- IEEE CEDA DA Technical Committee: "Metrics4ML" <u>https://github.com/ieee-ceda-datc/datc-rdf-Metrics4ML</u>





- What is TILOS?
- Why TILOS?
- TILOS Goals
 - Refocusing: EDA is Optimization
 - New Foundations of ML and Optimization
 - A Next Level of Translation

• Learning-enabled Optimization at Scale: So Much To Do !!!





• Learning to Optimize (L2O)

• Models, predictors and objectives; sampling; RL; hybridized optimizers

• Scaling the reach of optimizers

- Partitioning, clustering, sparsification; cloud/parallel; multi-{dims,objs}
- Optimal solvers as well (e.g., optimal peephole/clip P&R&Opt)

System/Arch/SoC PPAC exploration

- Learning to cluster+shape+pack+plan; pathfinding with confidence
 - Stack of abstractions: device, circuit, memory, integration fabrics

• And more ... (what are your targets and potential collaborations?)



Strategic Plan: Four Chips Research Thrusts

• TILOS "Strategic & Implementation Plan" accepted by NSF on March 1

- Research Teams \rightarrow Thrusts \rightarrow Topics \rightarrow Projects with identified lead faculty
- "SMART" goals and metrics: multi-disciplinary, multi-organization collaboration, knowledge transfer, impact (best-ever results, software releases, etc.)

• Thrust 1: Layout Kahng, Pan

- Optimal embedding: deep learning, end-case optimizers, GPU acceleration
- Modern partitioning: constraints and objectives
- Nexus of sampling, sequential decision-making, L2O, cloud → elements of "DA 2.0"

• Thrust 2: Verification Gao

- Better search methods for derivative-free optimization → bridge the gap between search-based and modeling-based (e.g., Bayesian optimization) methods
- Interior search for nonlinear SMT and applications in verification



Strategic Plan: Four Chips Research Thrusts

- Thrust 3: Quantifying the Cost of "X" (e.g., X = Security) Koushanfar
 - Open-source data anonymity
 - Ensuring the integrity of training data
 - Privacy-preserving learning and optimization

• Thrust 4: Data, Benchmarking, Roadmapping Kahng

- Artificial netlist generation
- Roadmapping of CAD optimization
- Ethical and fair benchmarking

System Driver: IoT Networks Rosing

- Challenge and test new learning and optimization methods in complex, real environment
- Spans all use domains, axes
 - (scale, hierarchy, physics, security, autonomy, robustness, performance, energy, ...)

UCSD Plin @ TEXAS Yale



Broad Industry Support for Proposal (December 2020)



E XILINX

TEXAS Yale

Western Digital



Andrew B. Kahng, GLSVLSI TILOS Overview 20220606

TILOS Chips Team





Andrew B. Kahng, GLSVLSI TILOS Overview 20220606



TILOS Big Team





































CSD Penn TEXAS Yale







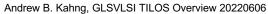




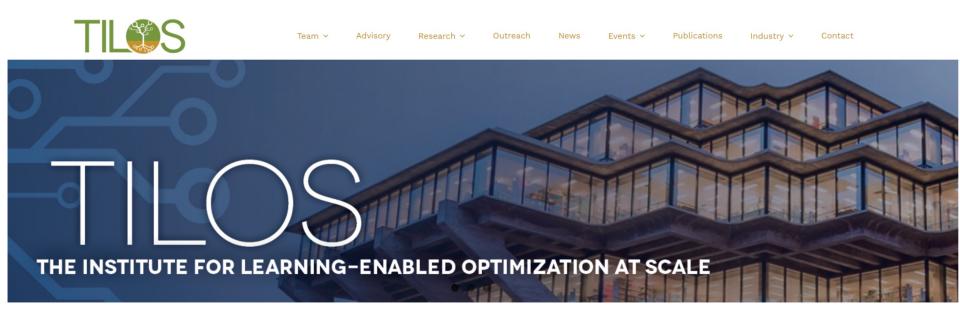








TILOS: An NSF AI Research Institute for Advances in Optimization



- Partially supported by Intel Corporation
- Launched on November 1, 2021



• Partnerships, collaborations welcome! tilos@eng.ucsd.edu

Score Score

Andrew B. Kahng, GLSVLSI TILOS Overview 20220606



THANK YOU !

• TILOS AI Institute: NSF CCF-2112665

TEXAS Yale

- Questions/Feedback (ABK): <u>abk-tilos@eng.ucsd.edu</u> , <u>abk@eng.ucsd.edu</u>
- TILOS General Inbox: tilos@eng.ucsd.edu
- Research at UCSD ABKGroup is supported by NSF, DARPA, Samsung, NXP, Mentor Graphics and the C-DEN center.

