



ACCELERATING THE DESIGN OF NEXT GENERATION COMPUTING SYSTEMS WITH GPUS

Sameer Halepete, Mar 2022



EXPLOSION IN DEMAND FOR
UNIQUE SILICON DESIGNS



“NVIDIA PROVIDES THE CORE TECHNOLOGY RESHAPING INDUSTRY AND SOCIETY”

DIGINOMICA



GAMING



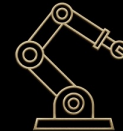
HPC



HEALTHCARE



SMART CITY



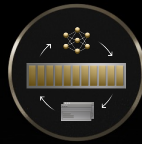
ROBOTICS



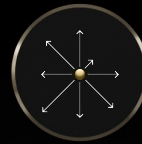
AUTOMOTIVE



vGPU



BASE
COMMAND



FLEET
COMMAND



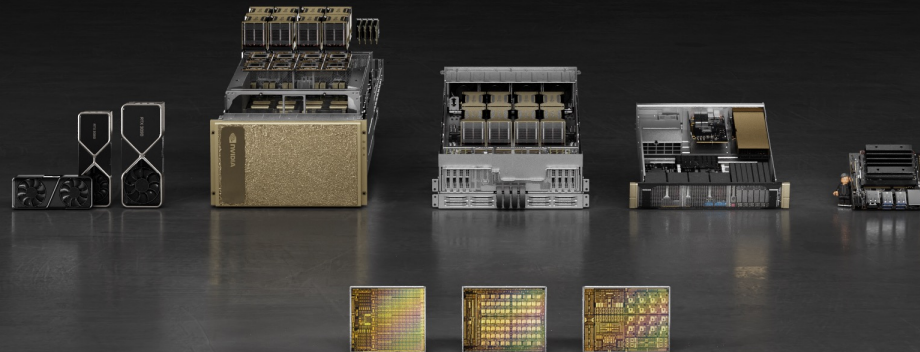
NVIDIA AI



OMNIVERSE

Incredible speed-ups take more than
just powerful chips

Full-stack invention: chips, systems,
algorithms, apps



“WITH ITS SERIES OF RTX GRAPHICS CARDS, NVIDIA IS WOWING THE GAMING WORLD”

TECHRADAR

NVIDIA Ampere architecture-powered devices deliver ray tracing at industry-best performance

RTX supported by more than 150 top games and creative apps



“GEFORCE NOW SHOWCASES THE LIMITLESS POTENTIAL OF CLOUD GAMING”

TECHRADAR

NVIDIA GeForce NOW™ is a GeForce® PC in
the cloud

Macs, PCs, iOS/Android mobile devices

11 million members have instant access to
more than 1,000 games





“OMNIVERSE ALLOWS THE ENTIRE PRODUCTION
PROCESS TO BE SIMULATED WITH
PHOTOREALISTIC DETAILS,
AND WITH PHYSICAL PROPERTIES”

WIRED

A platform for simulation, connecting to virtual worlds

Obeys laws of physics, simulates particles, fluids, materials

Train robots, design products, or create digital twins of buildings, factories, cities

“BMW AND NVIDIA DEVELOP ‘NEXT LEVEL’ VIRTUAL FACTORY PLANNING”

ROBOTICS AND AUTOMATION NEWS

BMW uses NVIDIA Omniverse™ for its factory
of the future

Simulated end to end

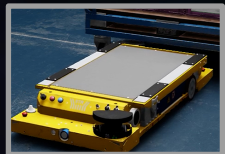
A digital twin of the factory BMW will build



FLOOR
PLANNING



ERGONOMICS
TRAINING



ROBOT
ORCHESTRATION



“THE UK’S MOST POWERFUL
SUPERCOMPUTER HAS GONE LIVE,
WITH POTENTIAL TO TRANSFORM
AI-BASED HEALTHCARE RESEARCH”

ZDNET

Cambridge-1 is a 400-petaflops AI supercomputer
for U.K. healthcare research

Based on NVIDIA DGX™ SuperPOD turnkey AI
supercomputing infrastructure



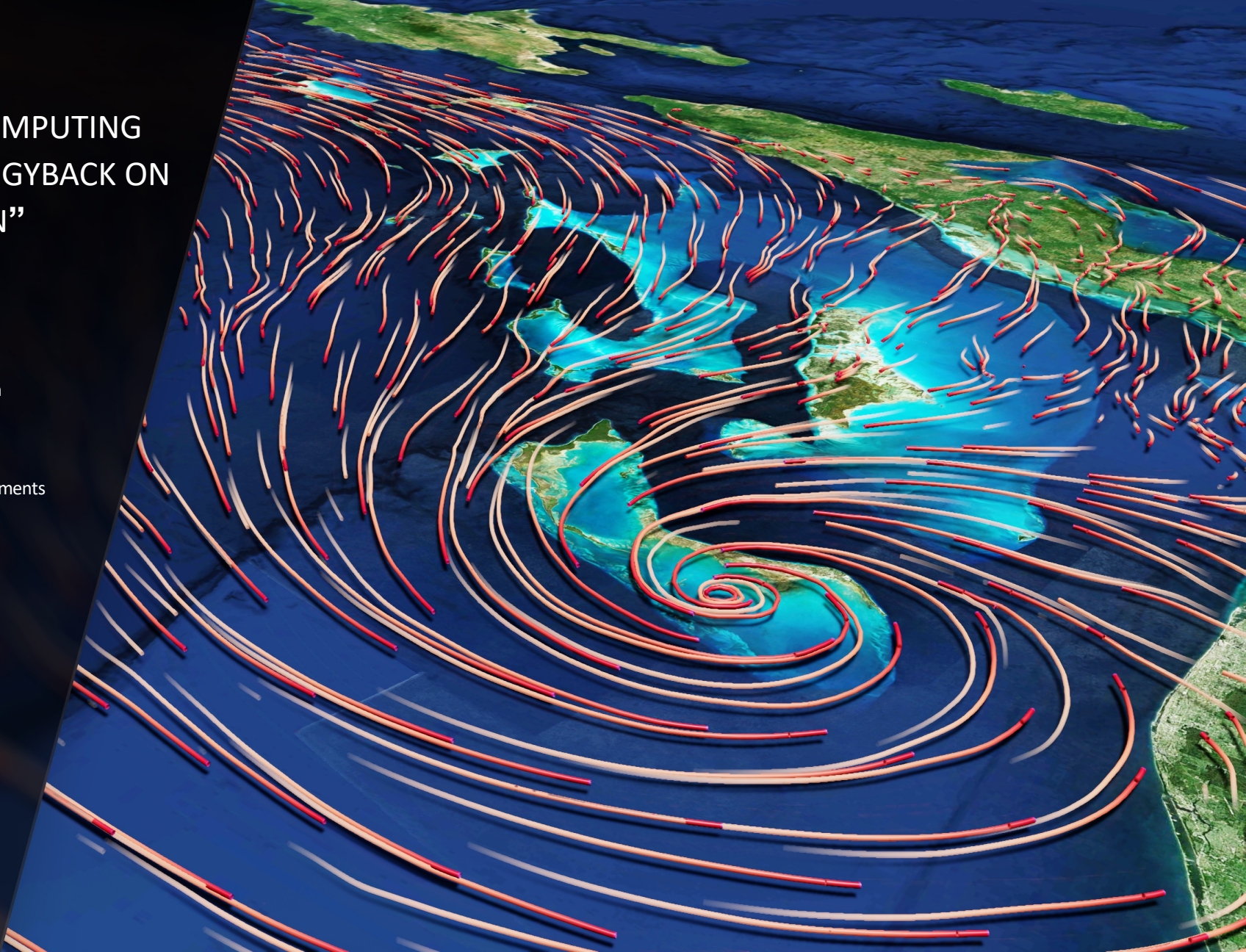
“HIGH PERFORMANCE COMPUTING
WORKLOADS GET TO PIGGYBACK ON
ALL OF THIS INNOVATION”

THE NEXT PLATFORM

NVIDIA accelerates 700+ apps; top 15 apps in
scientific computing

Full-stack optimization drives speed improvements
even without releasing new GPUs

Achieves 13X speed boost in 5 years



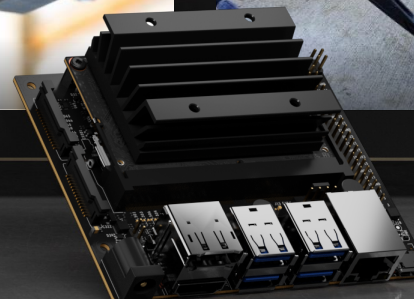
“NVIDIA JETSON NANO IS RESPONSIBLE FOR THE BIGGEST INDUSTRIAL IOT REVOLUTION THESE DAYS”

UPSWIFT

NVIDIA Jetson AGX Xavier™ powers embedded systems

Jetson Xavier NX enables millions of smaller,
low-power AI systems for embedded IoT apps

From edge to data center, all of NVIDIA's AI computers
run on the same CUDA-X™ AI software



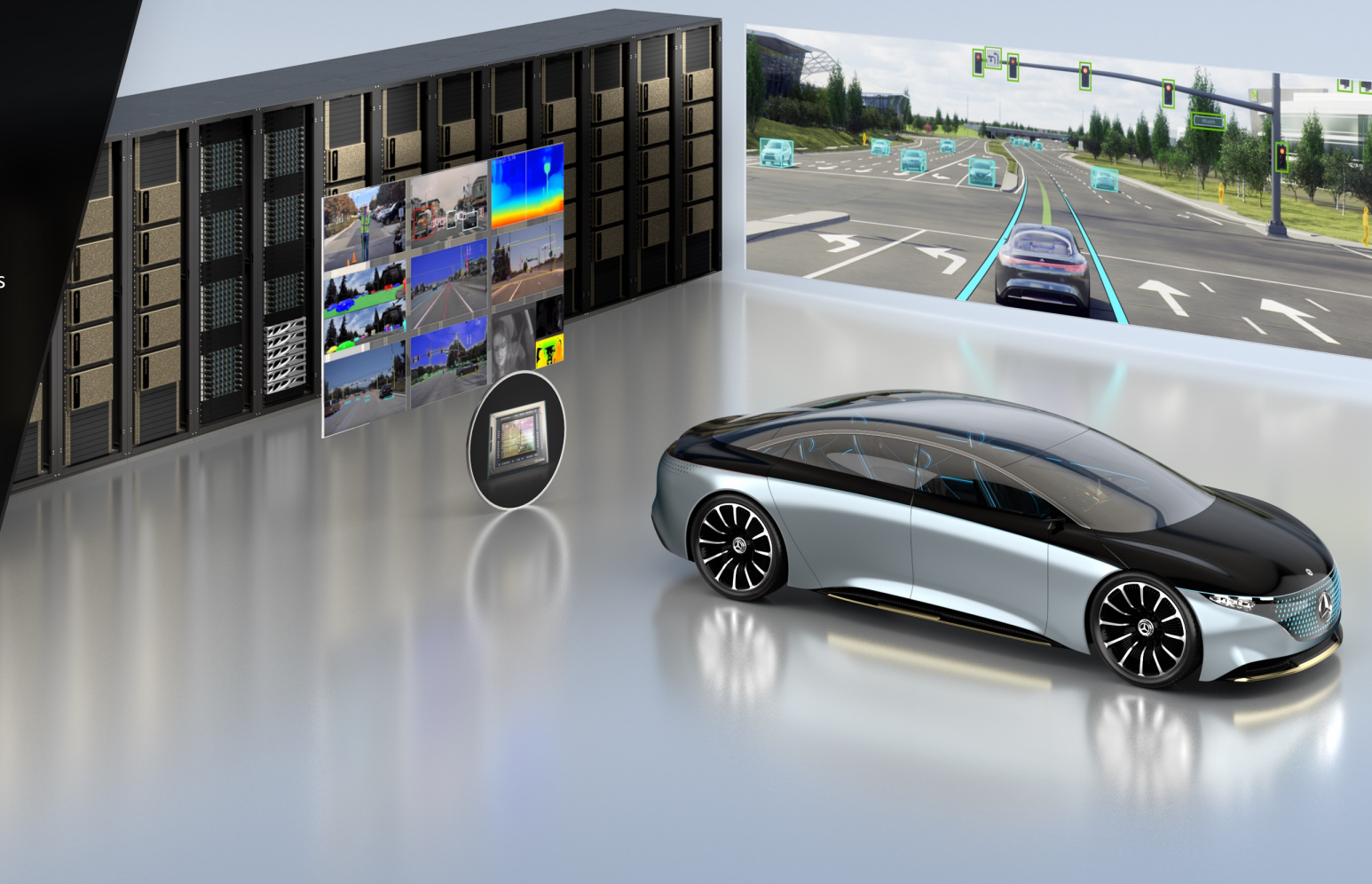
“MERCEDES-BENZ AND NVIDIA SIGN A DEAL TO MAKE CARS MORE LIKE IPHONES”

BLOOMBERG

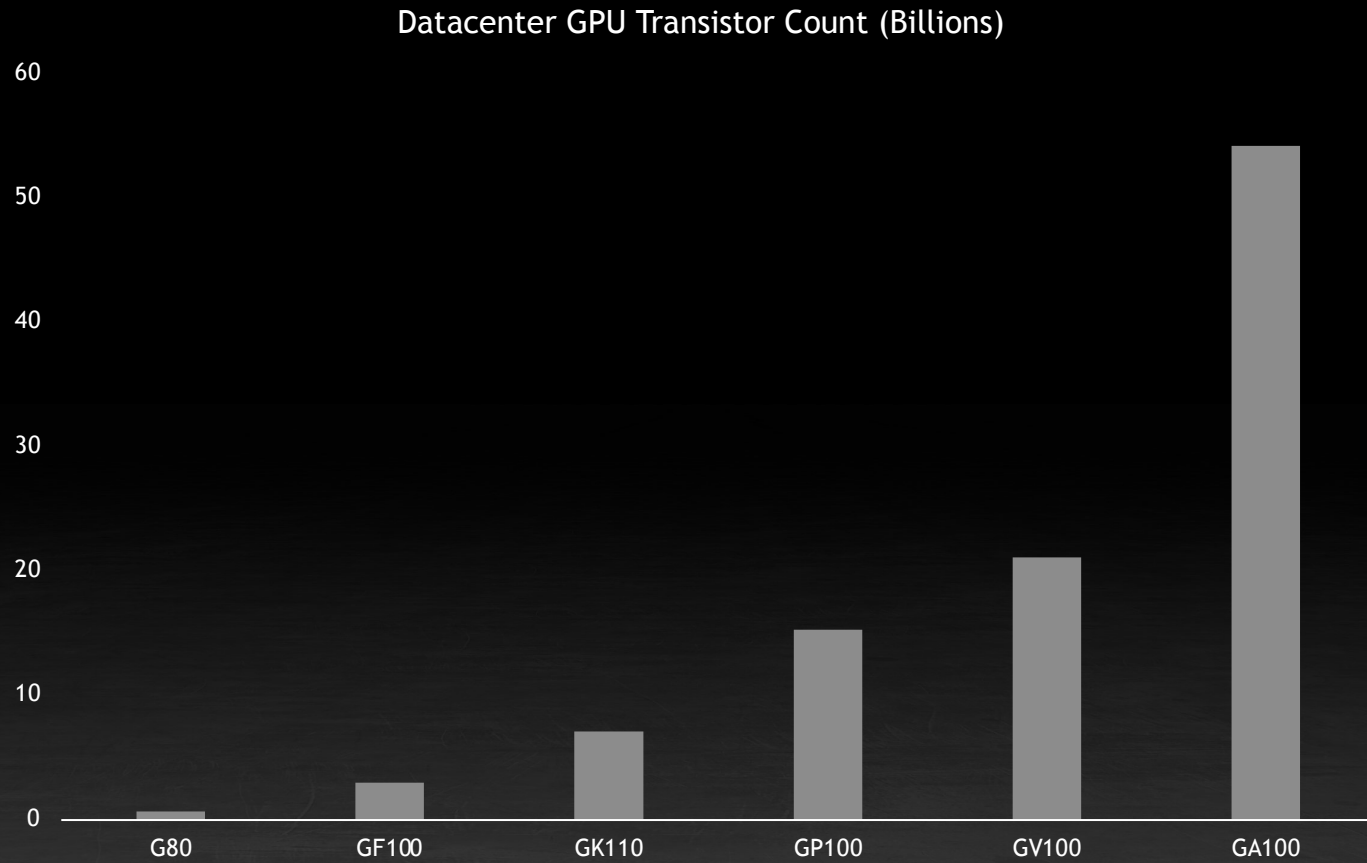
NVIDIA DRIVE™ is an end-to-end platform for developing autonomous vehicles

Supercomputer learns from road data, trains AI networks to drive

AV computer perceives, drives, learns, improves



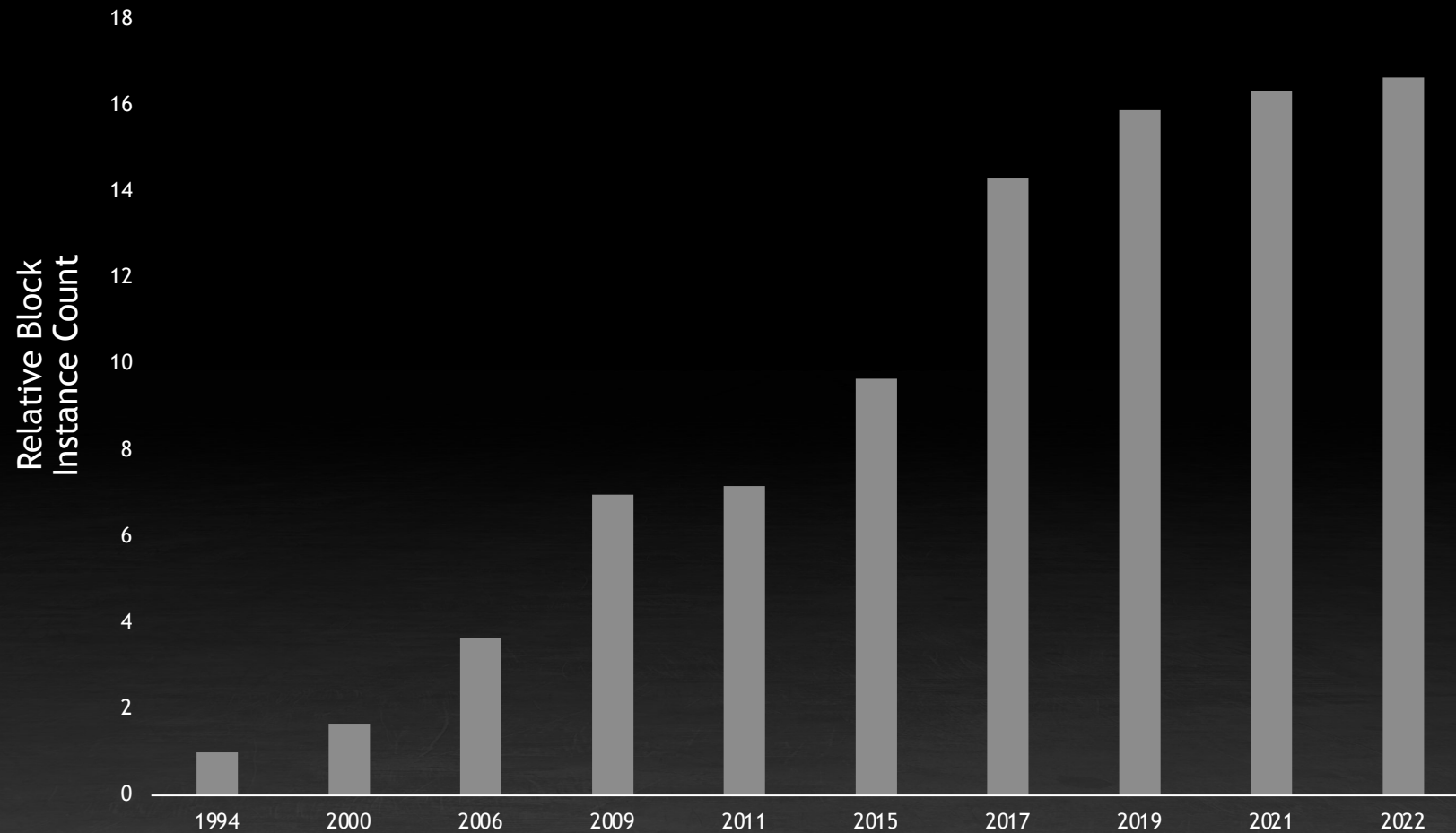
UNRELENTING INCREASE IN DESIGN SIZE/COMPLEXITY



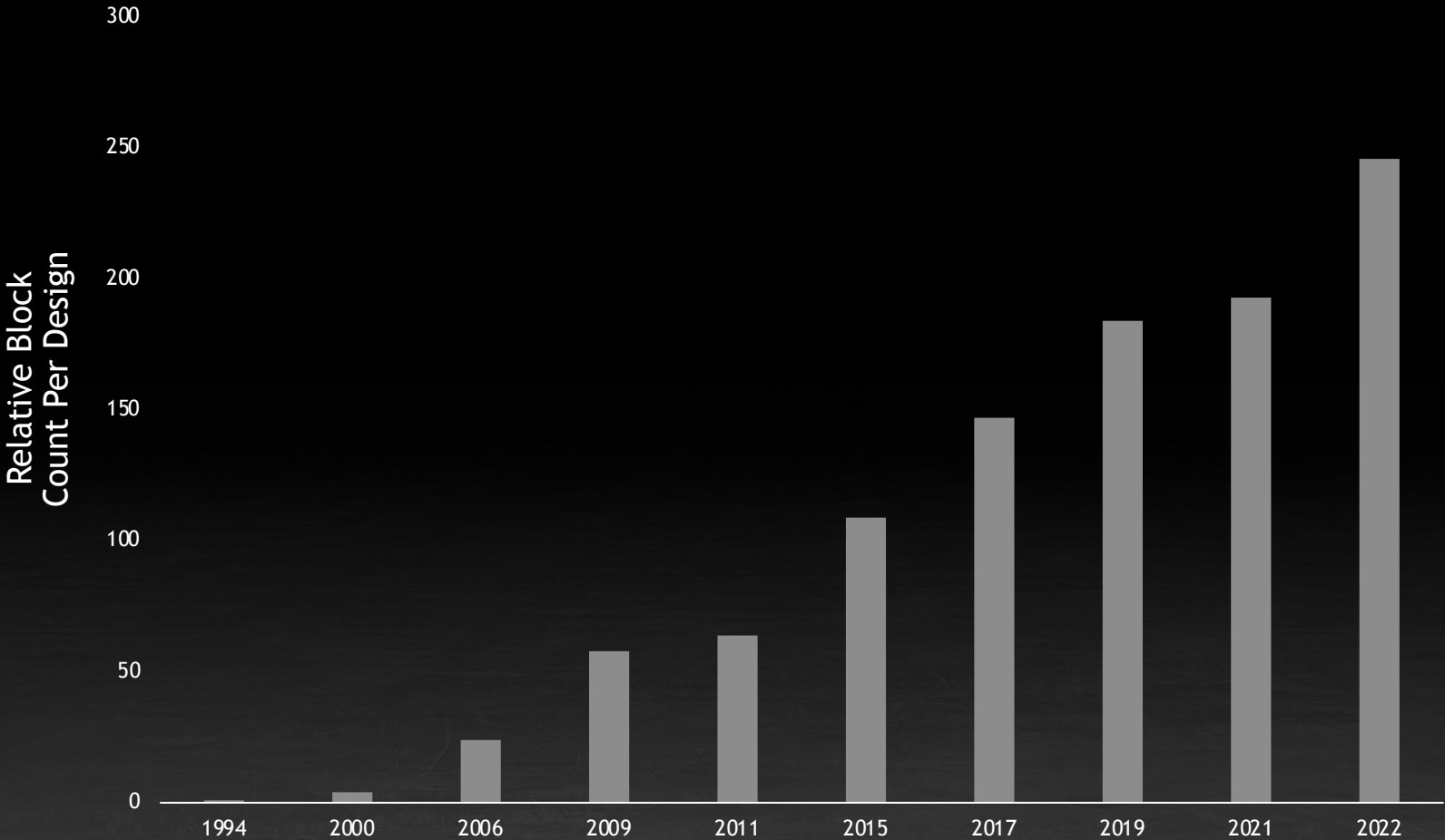


**SIGNIFICANT DROP-OFF IN VLSI
DESIGN PRODUCTIVITY**

AVERAGE SIZE OF DESIGN LEVELS OFF



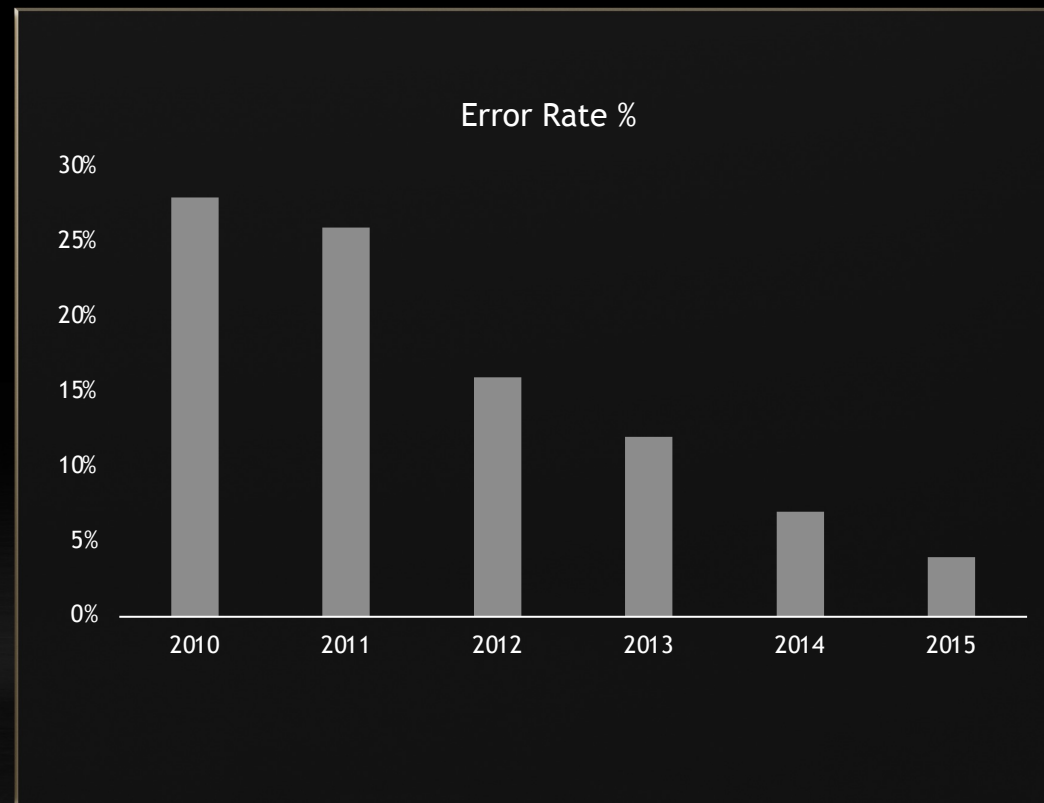
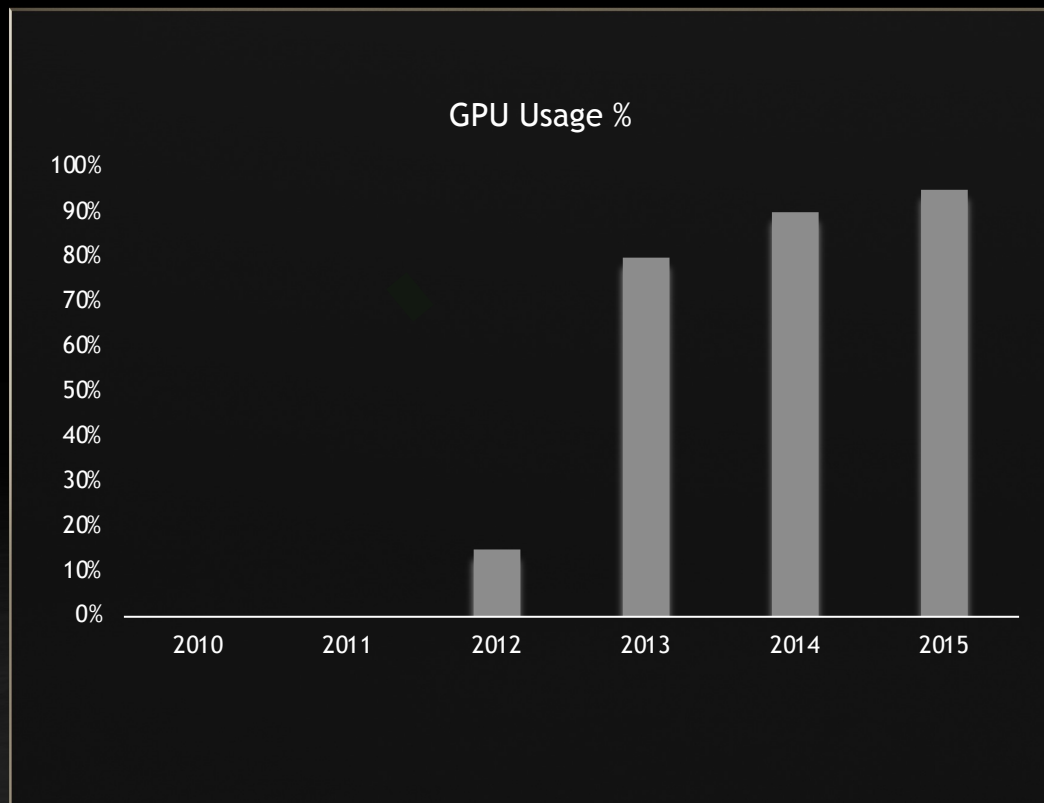
UNIQUE BLOCK COUNT SKYROCKETS

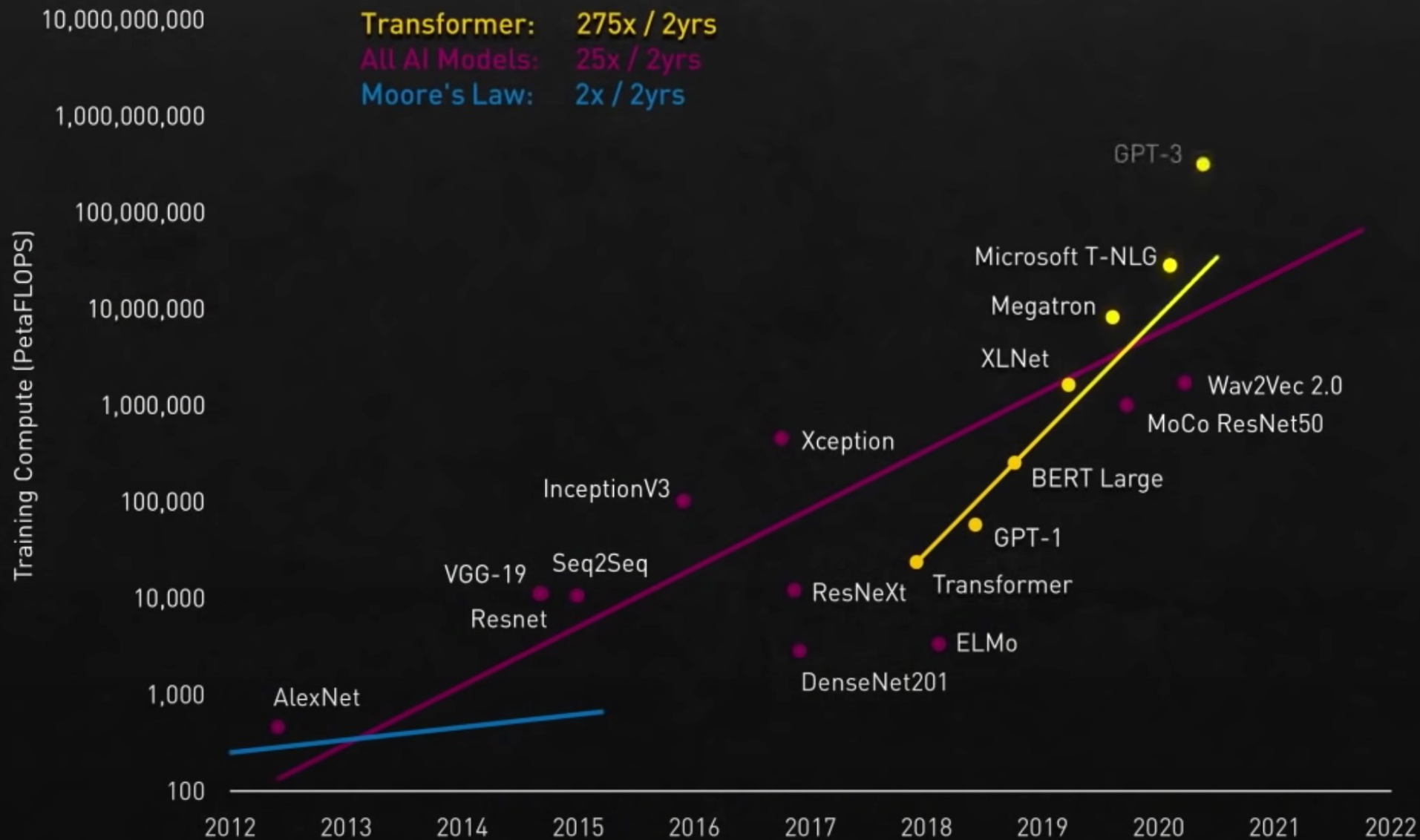




NEED TO RETHINK
EDA STRATEGIES

ALEXNET CAUSED A SEISMIC SHIFT IN COMPUTER VISION

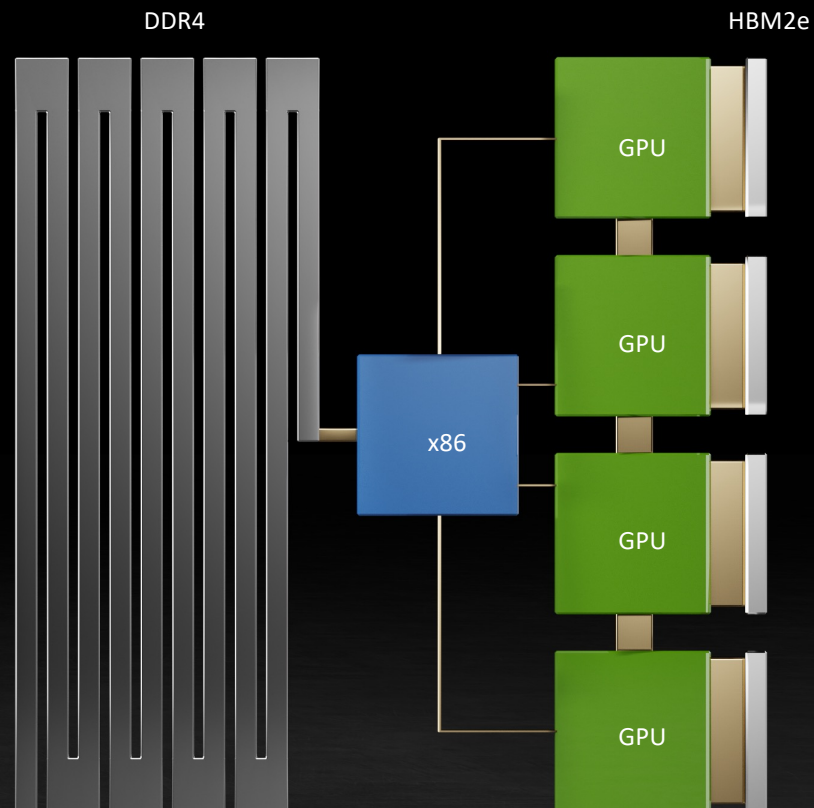




DATA-COMPUTE DEMAND GROWING FASTER THAN SYSTEM BANDWIDTH

GPU Starved by CPU Memory and PCIe Bandwidth

GPU	8,000	GB/sec
CPU	200	GB/sec
PCIe Gen 4	16*	GB/sec
Mem-to-GPU	64	GB/sec



* Effective bandwidth between CPU and each A100 GPU in half of a DGX A100 today illustrated

A NEW COMPUTING ARCHITECTURE FOR AI AND DATA SCIENCE

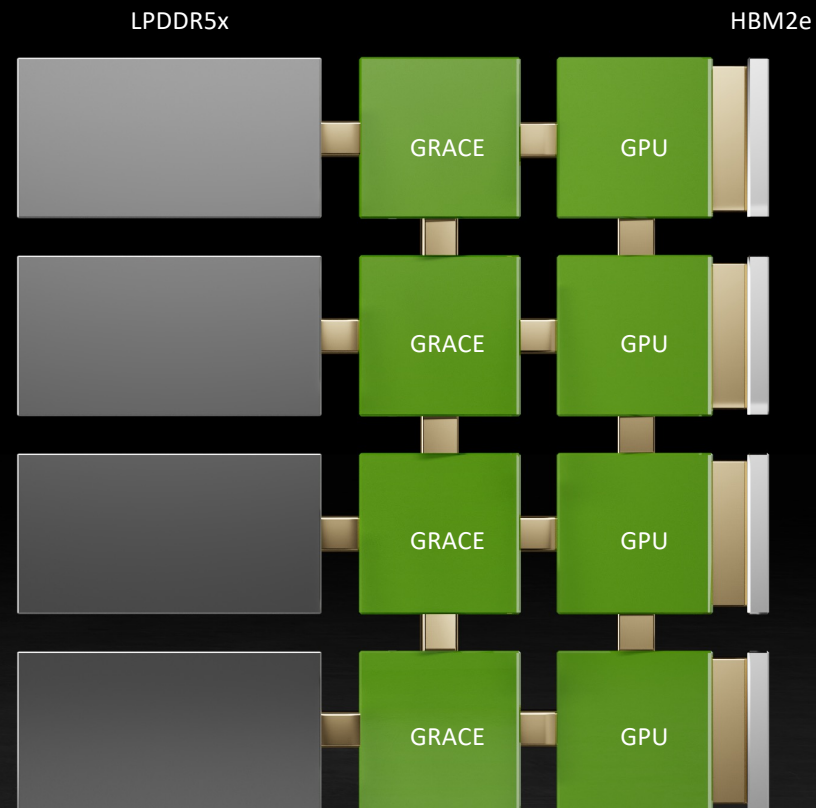
30X Increase System Memory to GPU

GPU	8,000	GB/sec
-----	-------	--------

CPU	500	GB/sec
-----	-----	--------

NVLINK	500	GB/sec
--------	-----	--------

Mem-to-GPU	2,000	GB/sec	30X
------------	-------	--------	-----





NVIDIA'S EDA RESEARCH STRATEGY

EDA RESEARCH STRATEGY

Understand longer-term potential for GPUs and AI in core EDA algorithms

Example – DREAMPlace:

RePIAce: Electrostatics-based
nonlinear global placement

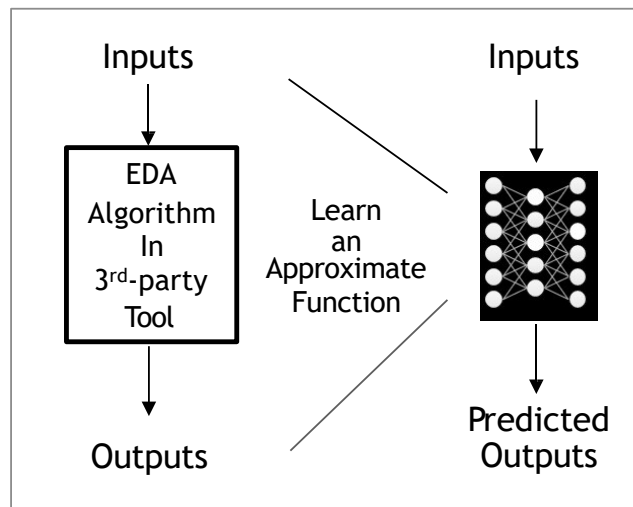


43x speedup on Global Placement

Accelerate known EDA algorithms
with our AI Accelerated Computing platform
Use GPU programming systems and
AI SW libraries for EDA

Potential Impact:

- (1) Direct acceleration of known methods
- (2) Scale up and scale out across a datacenter/cloud

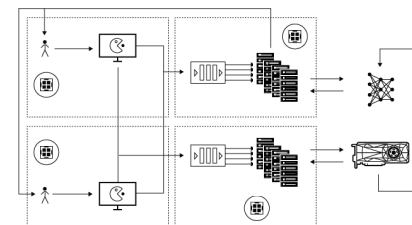


Trained ML models as predictors
Supervised learning with CNNs, GNNs, etc..
Applying latest NN models to EDA problems

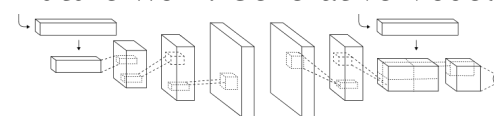
Potential Impact:

- (1) Augment traditional EDA w/ predictors
- (2) Differentiable functions in optimizations

Current Work: Reinforcement Learning



Future Work: Generative Models



Develop new EDA algorithms using AI
Combine VLSI design/domain knowledge with
AI methods for optimization

Potential Impact:

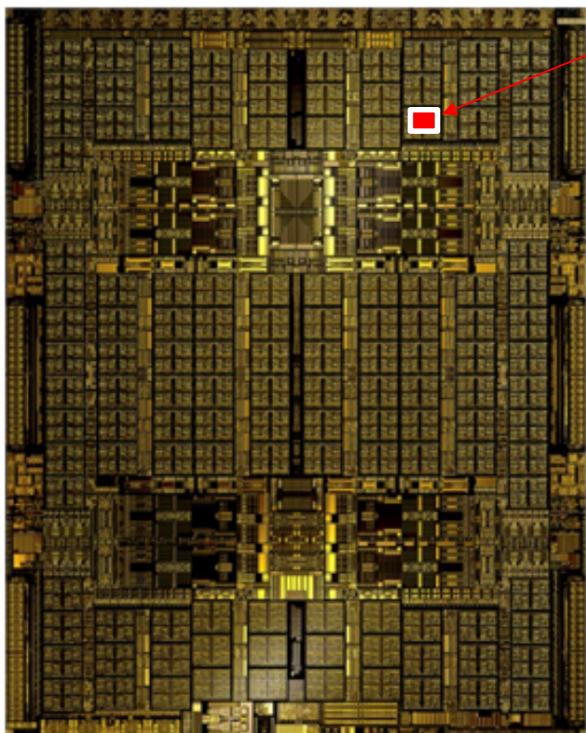
- (1) Automate manual tasks
- (2) Improve time-to-solution in EDA tools and flows with fewer iterations, faster steps

ACCELERATE PLACEMENT ON GPU WITH DL TOOLKITS

DREAMPlace (DAC 2019 best paper)

Collaboration between NVIDIA Research and UT-Austin

GA100 - Billions of standard cells

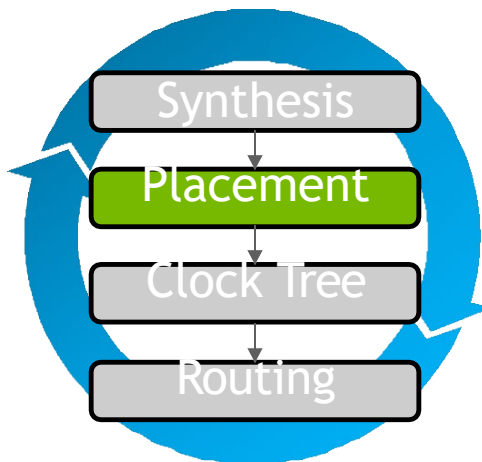


Large GPUs/SoCs divided into hundreds of unique *partitions*, with millions of cells per partition

How to improve scalability and dev productivity?

Use AI HW+SW platforms

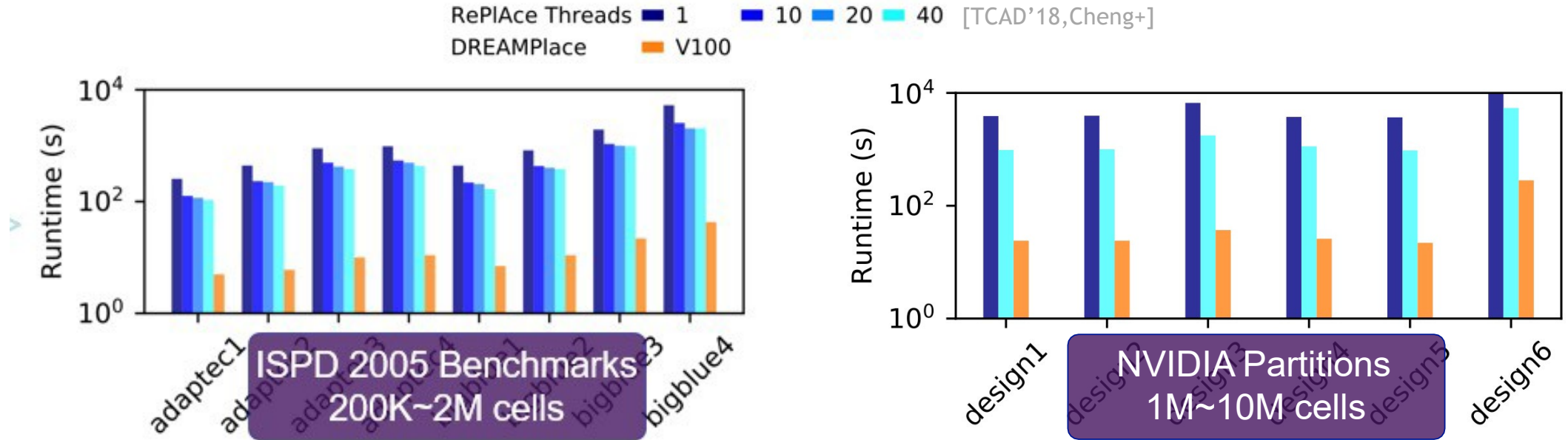
PyTorch TensorFlow



Iteration time through backend flow
Days for one partition,
Many iterations per tapeout

DREAMPLACE RESULTS

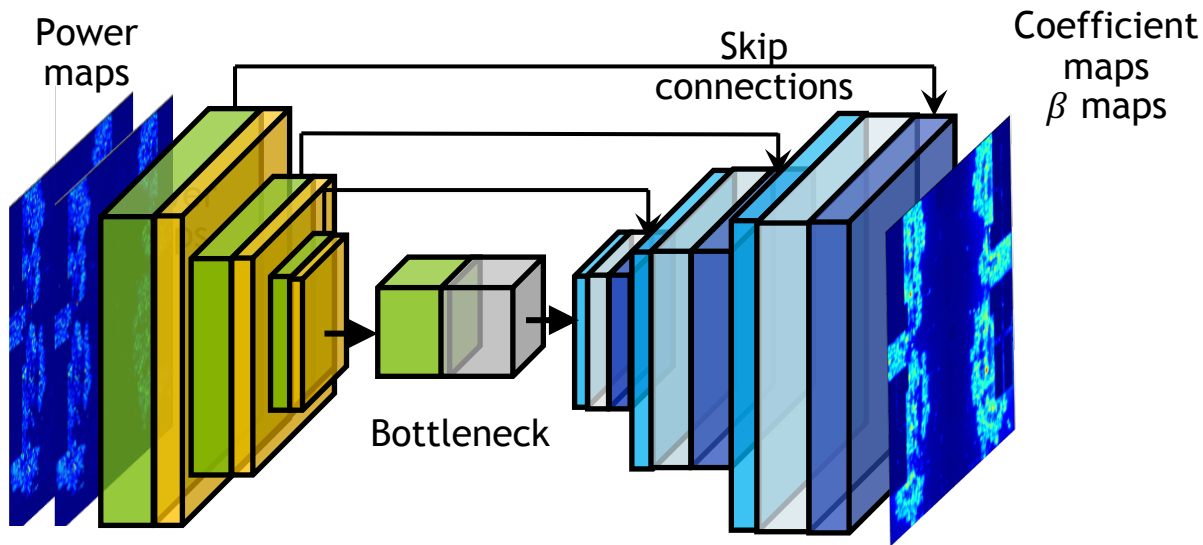
DREAMPlace vs. Best Academic Placer



up to 43x speedup: Enables 10M-cell chiplets to be placed in 5 minutes!

ABCDPlace, DREAMPlace 3.0 extended to detail placement, routerability optimization, multi GPU etc

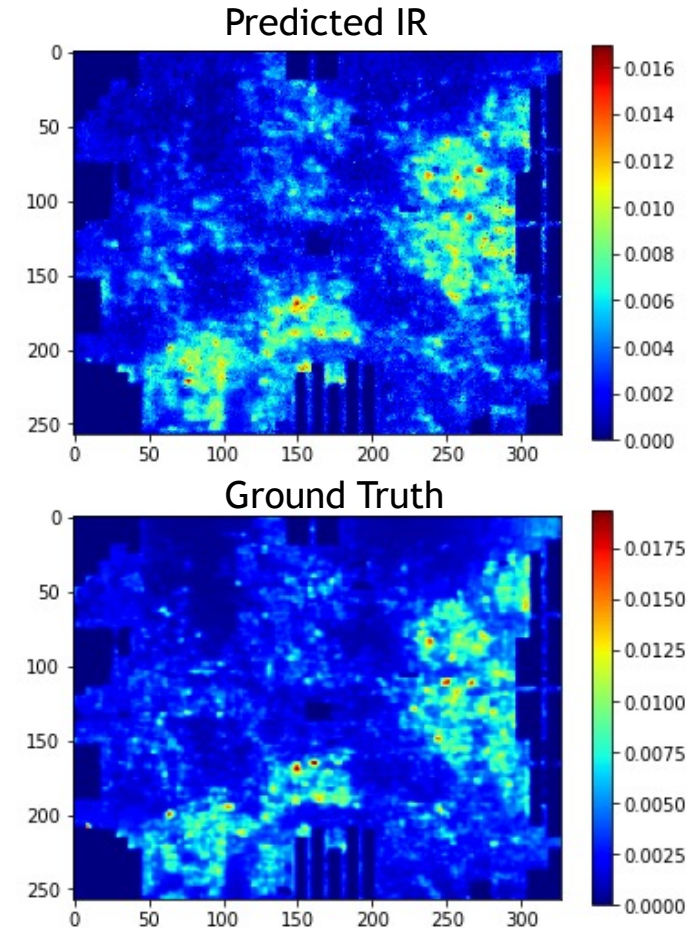
IR DROP PREDICTION WITH 3D-UNET



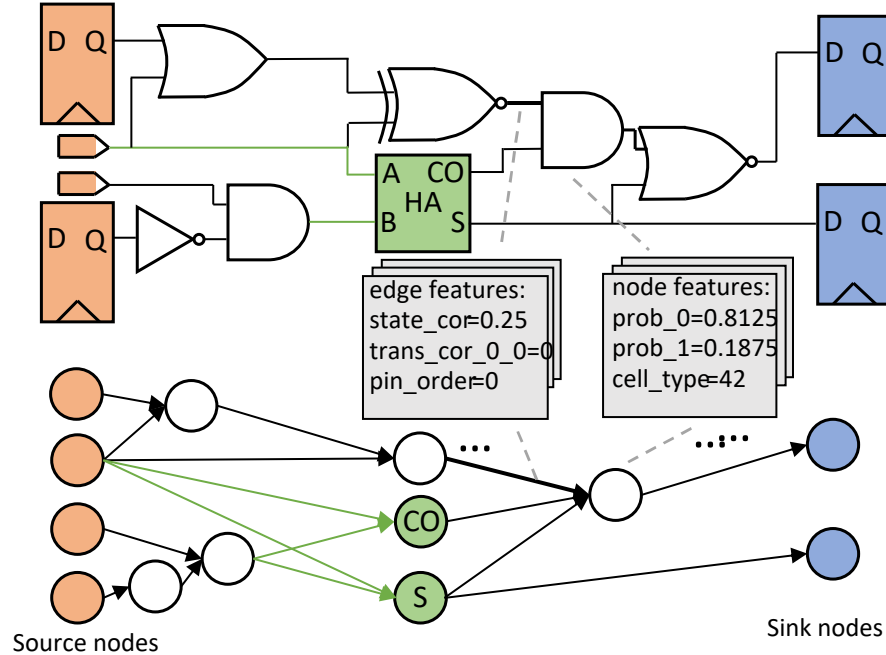
$$IR_{inst} = \beta_1 P_i + \beta_2 P_s + \beta_3 P_l + \beta_4 P_{sca} + \beta_5 P_{tot} + \beta_6 R + \beta_7 P_{ires}$$

Best of both worlds: Transferable and ML model learns the coefficients

94% accuracy in 3 second, 18 minutes including feature extractions
vs 3 hr in commercial tools



POWER DISSIPATION WITH GNNs

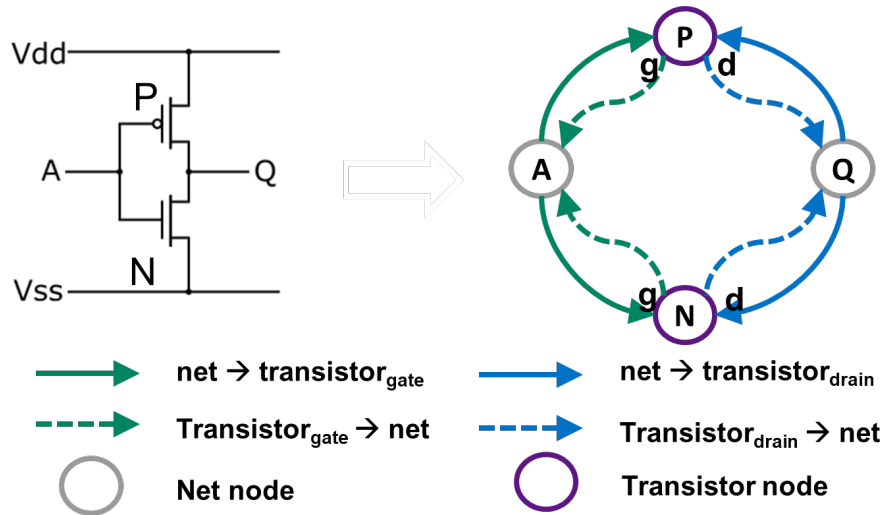


Netlist to Graph Conversion

Design	Batch Size=1 Inference Latency (s)	Max Windows Batch Size	Per Cycle Simulation Throughput (kHz)	GNN Inference Throughput (kHz)
qadd_pipe	0.293	2200	2.5	5050.9 (2044.9X)
qmult_pipe	0.401	1400	1.9	1679.5 (888.3X)
fadd	0.304	1200	8.8	3024.5 (343.7X)
fmult	0.342	950	6.2	1259.9 (202.5X)
NoCRouter	0.435	175	1.4	194.7 (141.0X)
RISC-V	0.703	30	2.4	44.4 (18.7X)

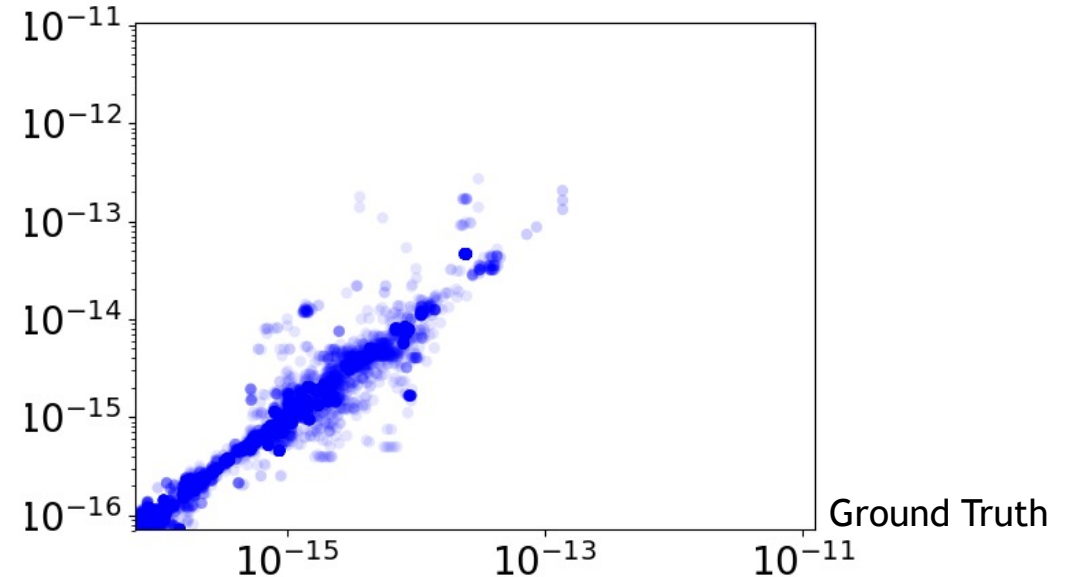
Design	Ground Truth Average Power (mW)	Baseline Predicted Power (mW)	GRANNITE Inference Power, Testing (mW)	GRANNITE Inference Power, Validation (mW)
qadd_pipe	0.553	0.550 (0.5%)	0.564 (2.0%)	0.563 (1.9%)
qmult_pipe	2.018	2.006 (0.6%)	2.156 (6.8%)	2.084 (3.3%)
fadd	0.068	0.103 (51.5%)	0.071 (4.0%)	0.068 (0.3%)
fmult	0.219	0.563 (157.1%)	0.215 (1.9%)	0.221 (1.0%)
NoCRouter	1.036	1.088 (5.0%)	0.897 (13.4%)	0.918 (11.4%)
RISC-V	0.923	0.904 (2.0%)	0.997 (8.0%)	0.873 (5.4%)
datapath_units		N/A (3.3%)	N/A (1.0%)	N/A (0.9%)
AVERAGE		(31.4%)	(5.3%)	(3.4%)

PARASITICS PREDICTION WITH GNNS



Circuit Schematics to Graph Conversion

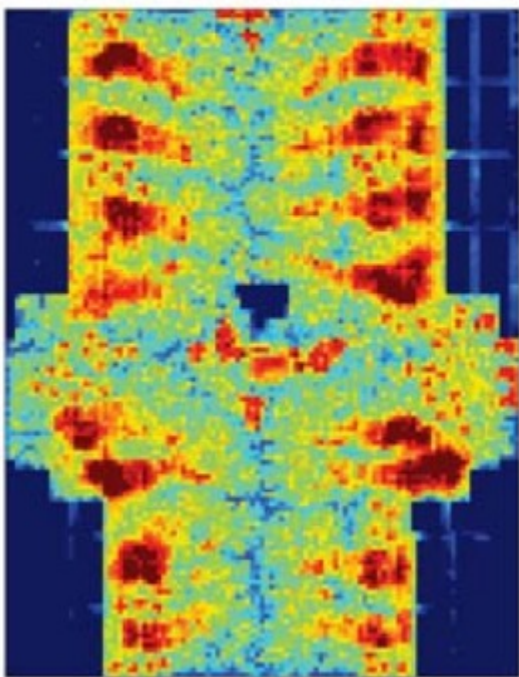
Cap Prediction (F)



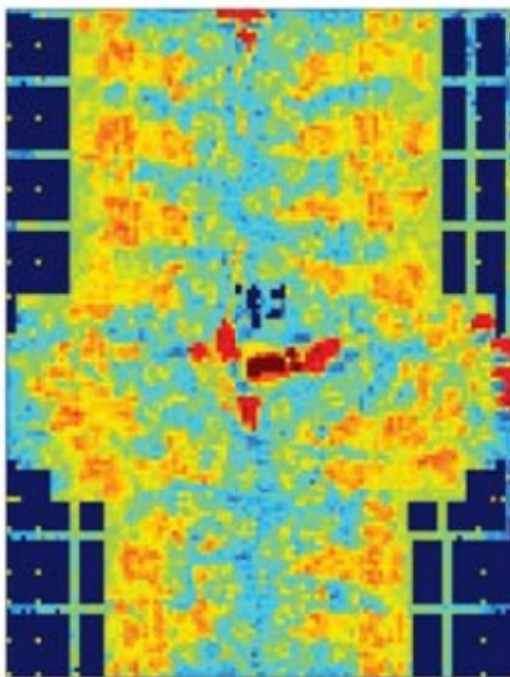
MAE=0.852fF MAPE=15%
Simulation error reduced to <10%

ROUTING CONGESTION PREDICTION WITH GNNS

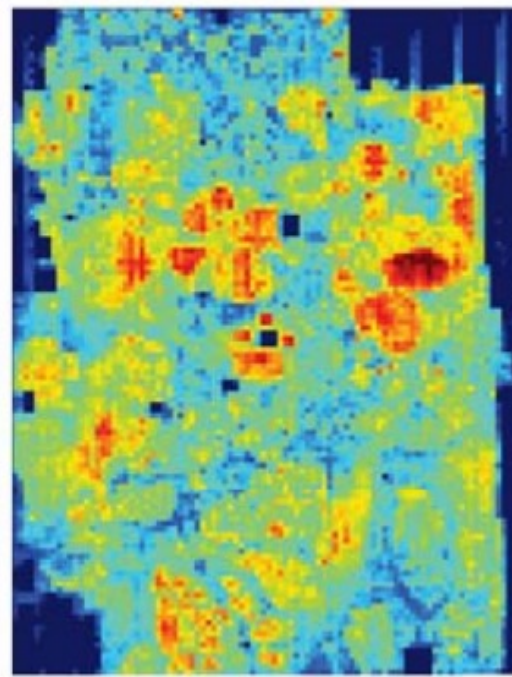
Partition_B
Actual Congestion



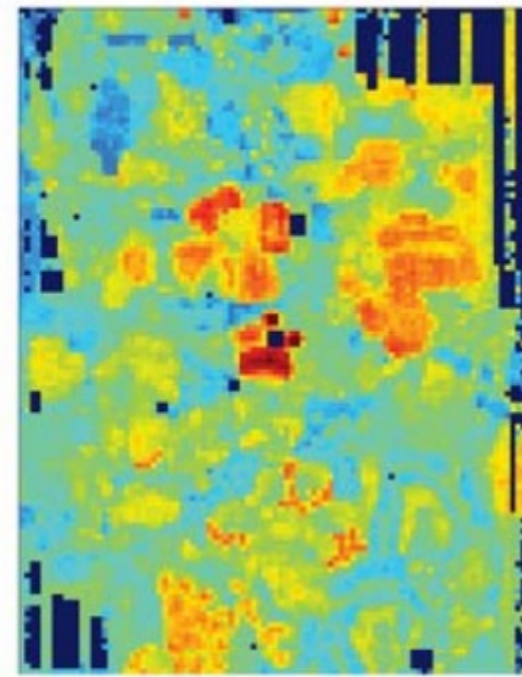
Partition_B
Predicted Congestion
Kendall Correlation : 0.61



Partition_F
Actual Congestion



Partition_F
Predicted Congestion
Kendall Correlation : 0.53



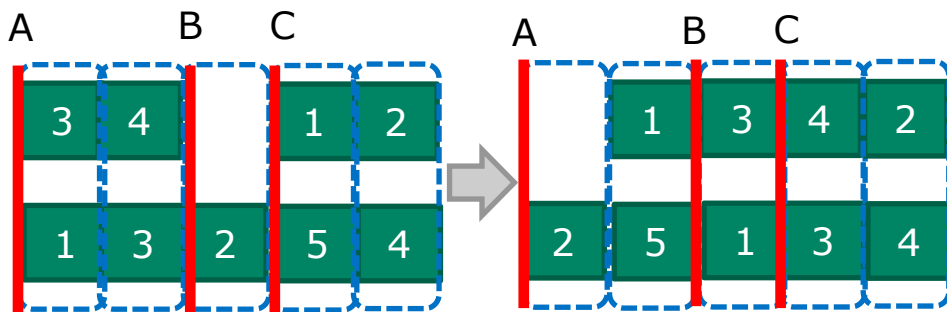
NVCELL: AUTOMATED STDCELL LAYOUT

(Ren et al., DAC 2021)

Simulated Annealing

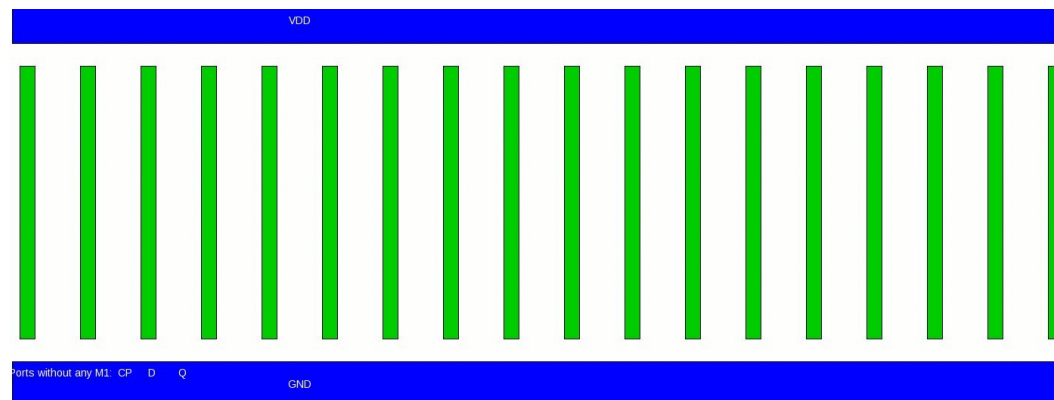
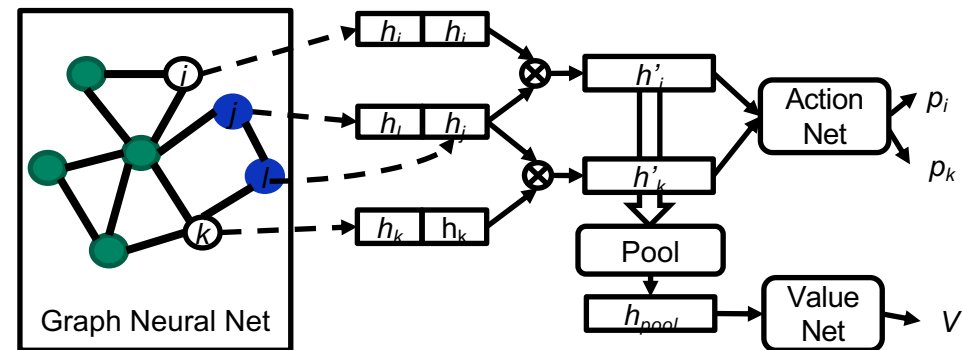
PMOS sequence: 3, 4, NA, 1, 2
NMOS sequence: 1, 3, 2, 5, 4
PIN sequence: A, NA, B, C, NA

Placement representation



Transforms: Swap PMOS/NMOS pair segment, etc

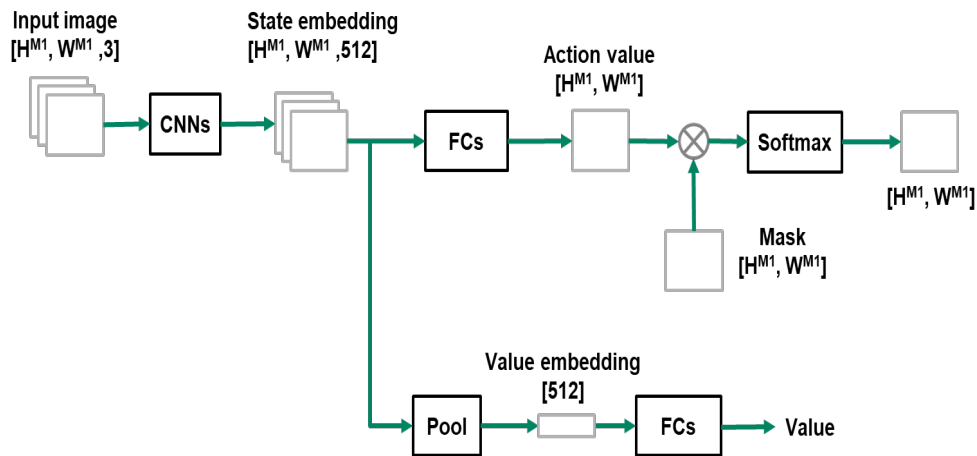
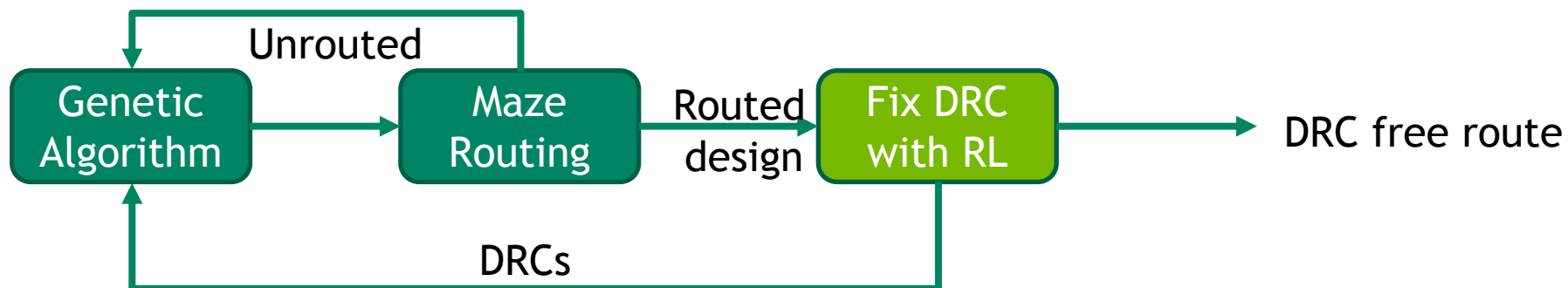
Reinforcement learning (in progress)



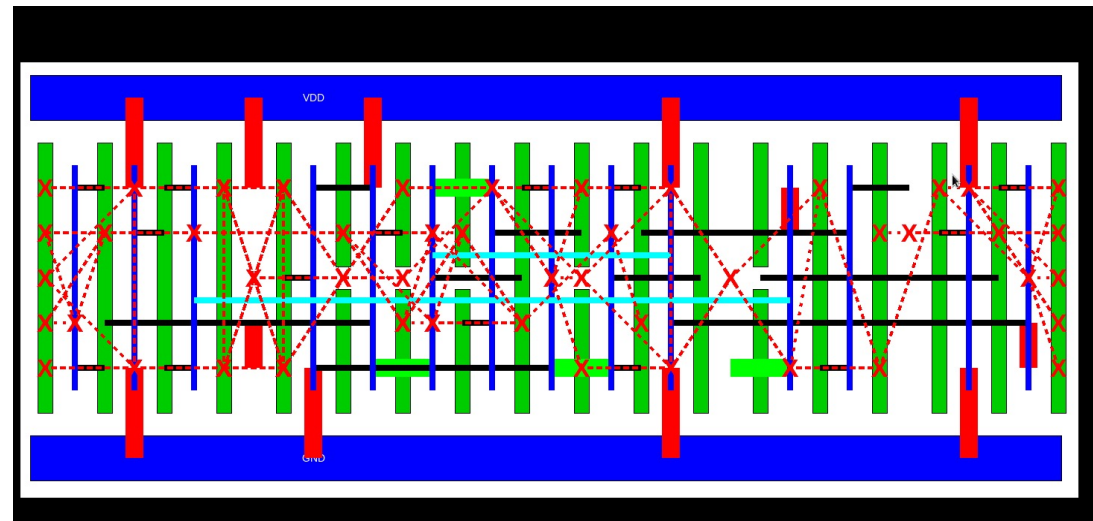
RL Placement Game Sequence (as good as SA on 97% cells)

NVCELL: AUTOMATED STDCELL LAYOUT

Routing (Ren et al., ASPDAC 2020)



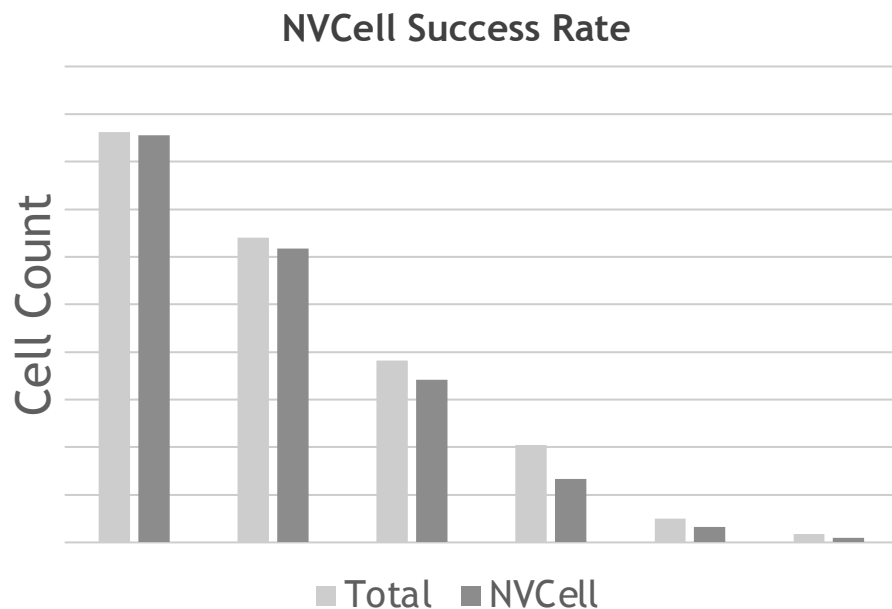
Transferable RL Model trained only one single cell



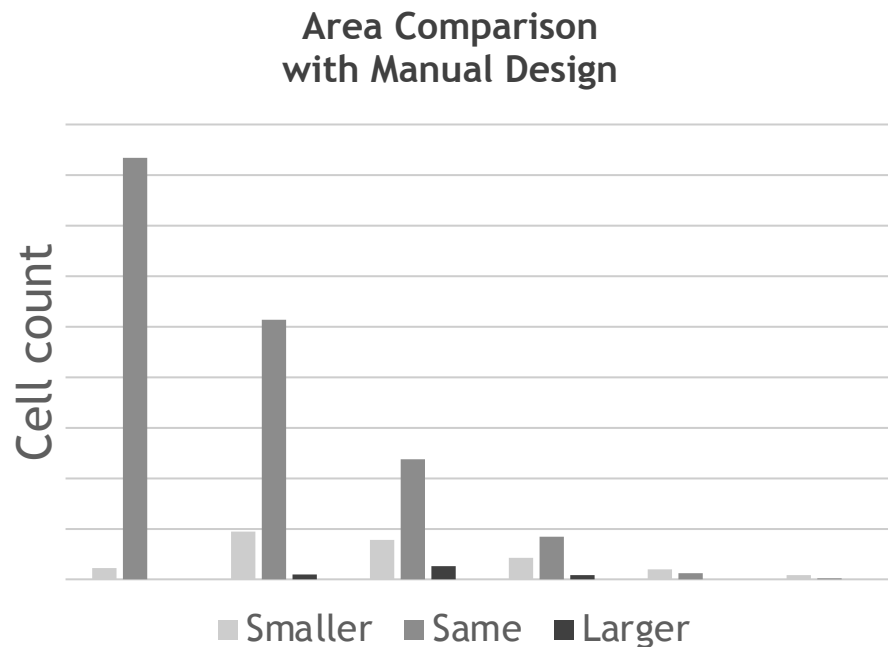
DRC Fixing with RL

NVCELL: AUTOMATED STDCELL LAYOUT

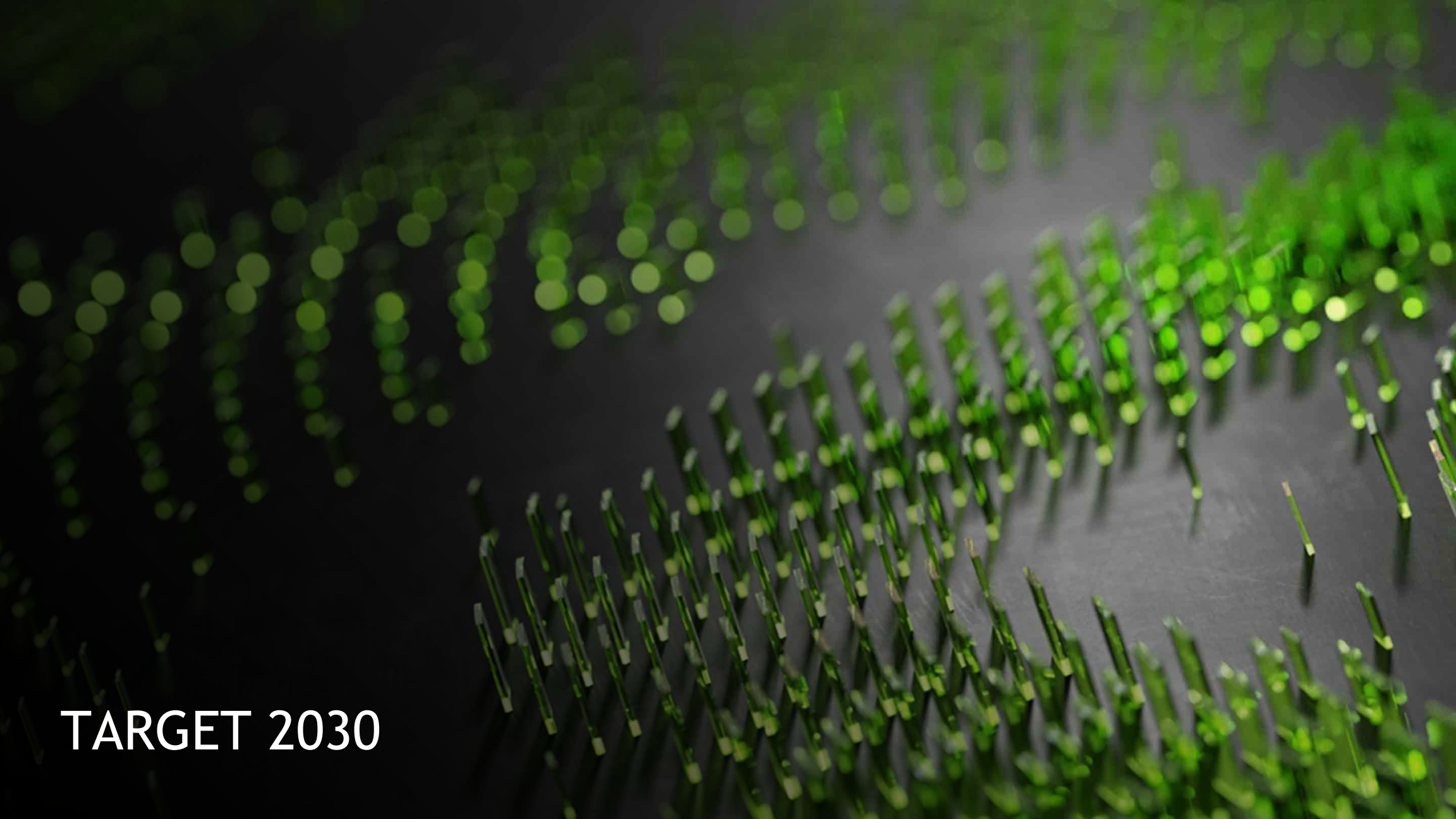
Production use in latest technology nodes, achieving 2X overall productivity improvement



92% single row cells are DRC/LVS clean
Support multirow. High success rate with tuning.



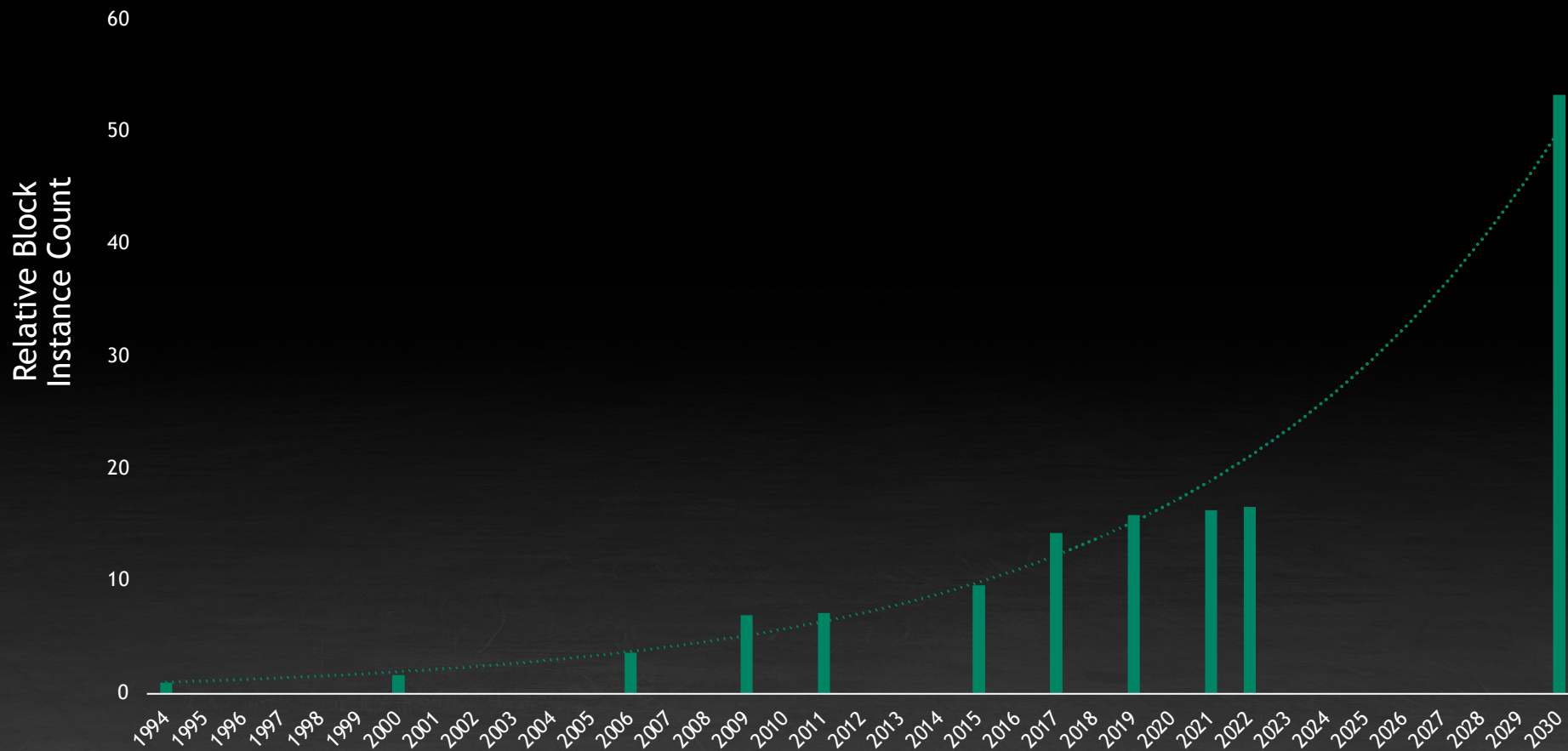
12% cells are smaller
Similar delay/power ($\pm 1-3\%$)

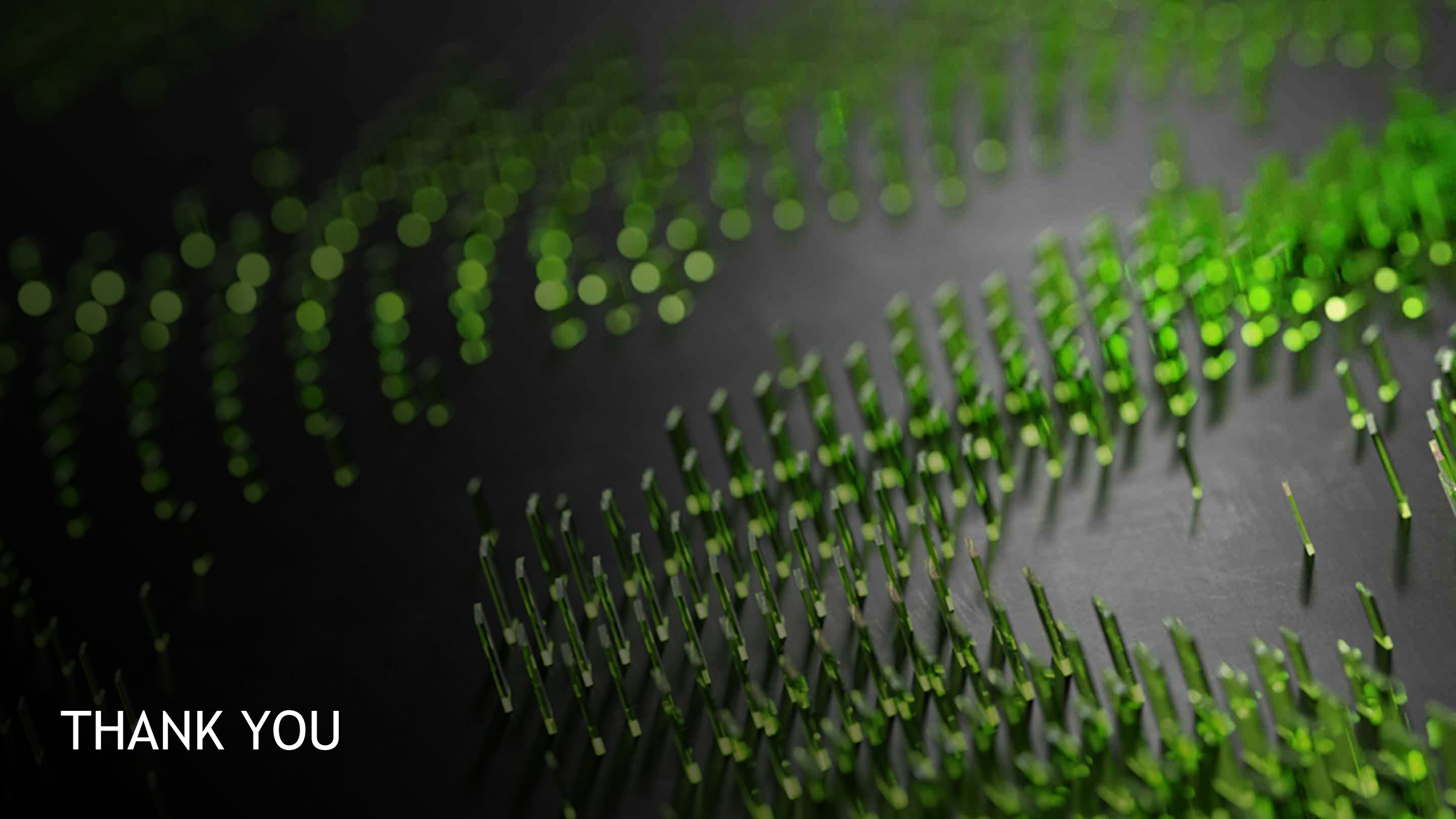


TARGET 2030

TARGET 2030

3X DESIGN SIZE BY THE END OF THE DECADE





THANK YOU