AI HARDWARE SUMMIT

SYNOPSYS®
Silicon to Software™

Enabling On-Device Intelligence

Opening Keynote

AI Acceleration On-Device & Industry Trends & Dynamics

Dr. Thomas Andersen

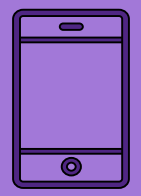
Head, Machine Learning and AI, Design Group

Artificial Intelligence – Avenue to Innovation

ICs Across All Markets to Include AI Capabilities

MOBILE

All Smartphones will integrate AI Processing Capabilities by 2021



DATA CENTER

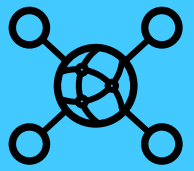
More than 50% of enterprises will deploy AI accelerators in their server infrastructure by 2022

AUTOMOTIVE

Volume production of autonomous vehicles will begin in 2020

IOT

More than 20% of IoT devices will have AI processing Capabilities by 2022



An Explosion of Abundant-Data Computing

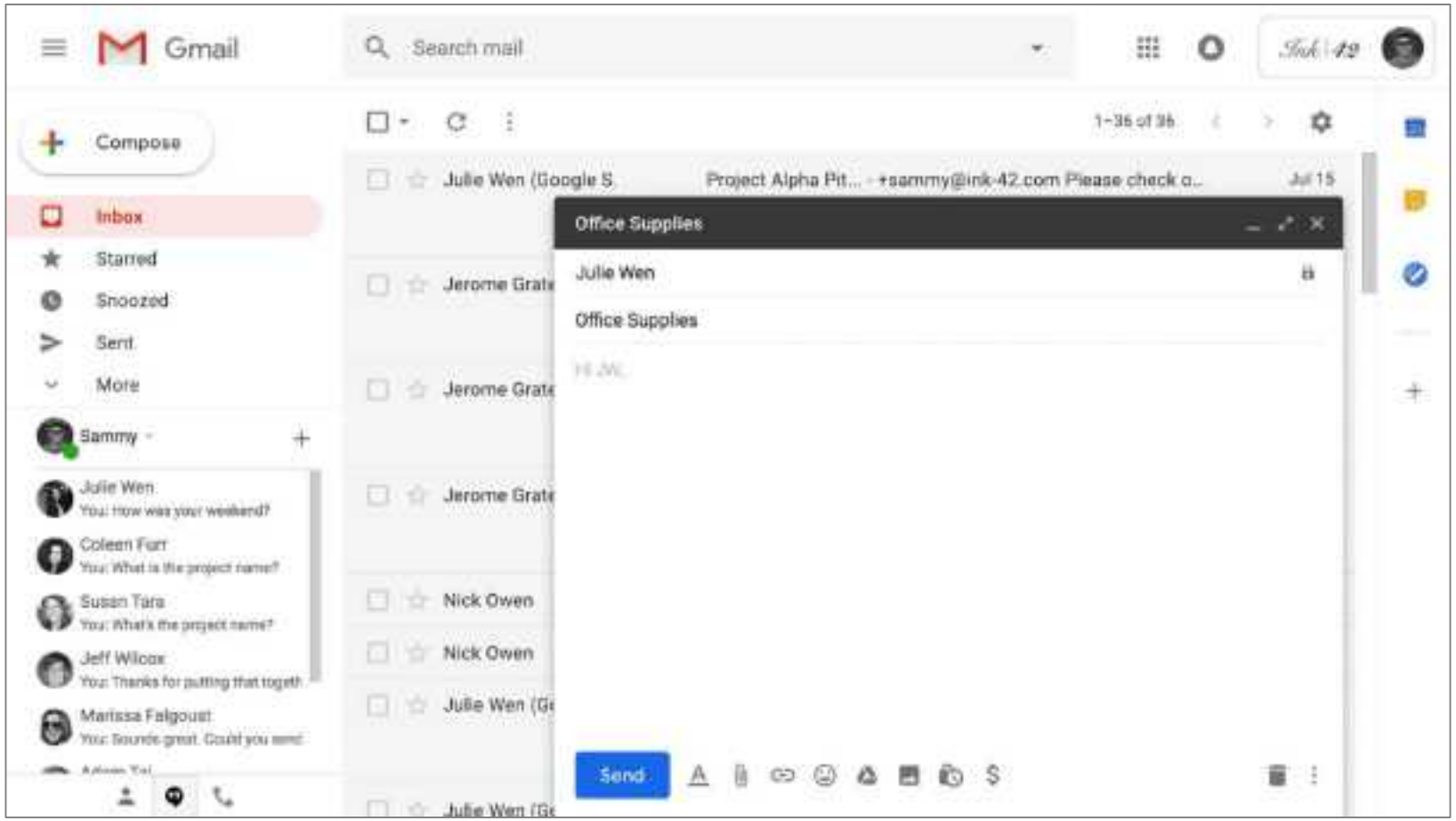
Today, training Chinese speech recognition models requires 20 exaflops of compute

By 2040, data storage needs will exceed 100 Zetta-Bytes/yr

By 2025, data centers will consume 1/5 of all the world's energy

Recent AI Advancements

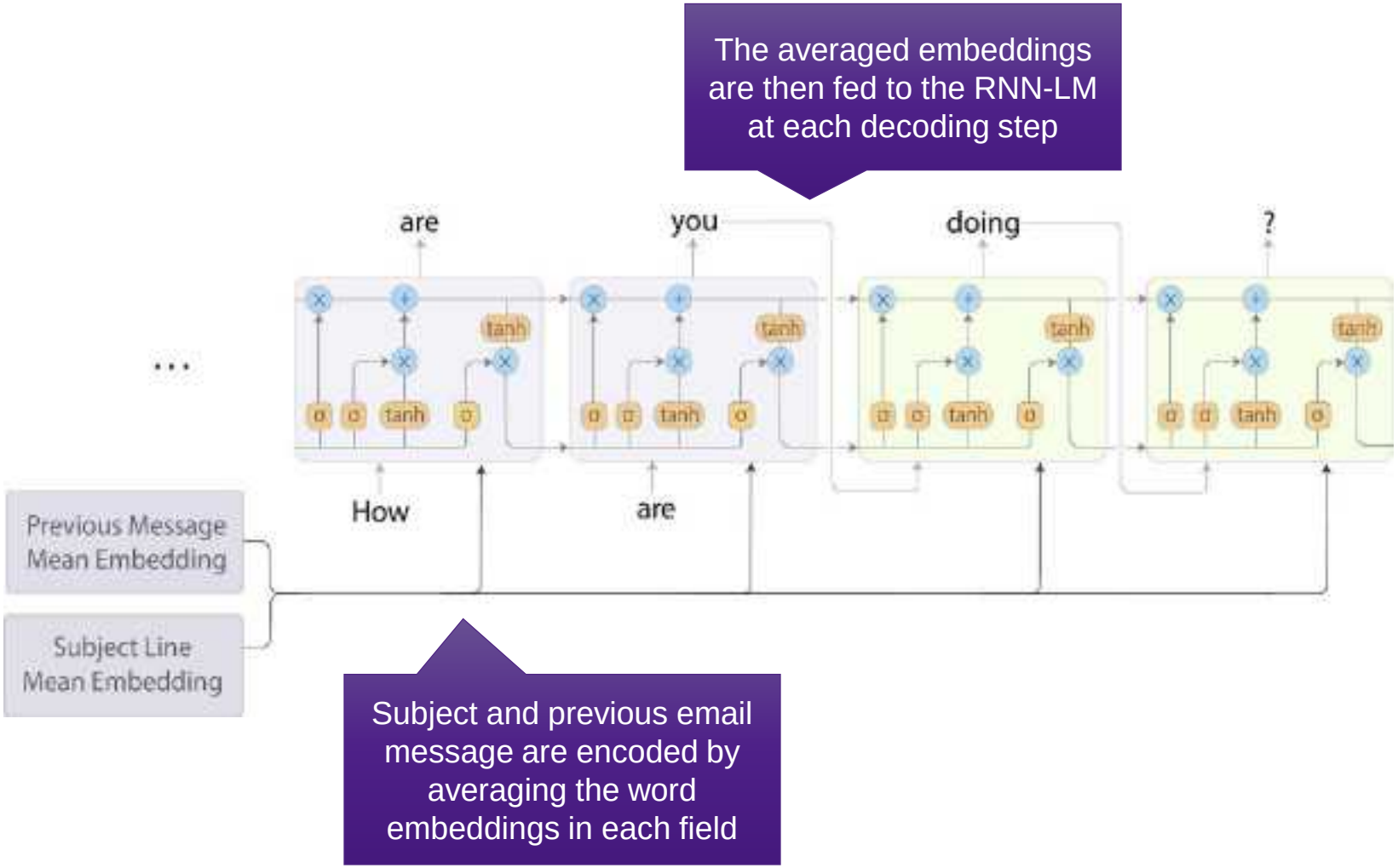
1 - Natural Language Processing



- Supervised learning
- RNN Language Modeling (LM) architecture
  - Predict next word in a sequence, within conversation context
- Model capacity for 1.4+ billion users
  - Make tailored suggestions
- 100ms response time
  - Running on TPU2

Recent AI Advancements

1 - Natural Language Processing



- Supervised learning
- RNN Language Modeling (LM) architecture
  - Predict next word in a sequence, within conversation context
- Model capacity for 1.4+ billion users
  - Make tailored suggestions
- 100ms response time
  - Running on TPU2

Recent AI Advancements

2 - Generative Adversarial Networks

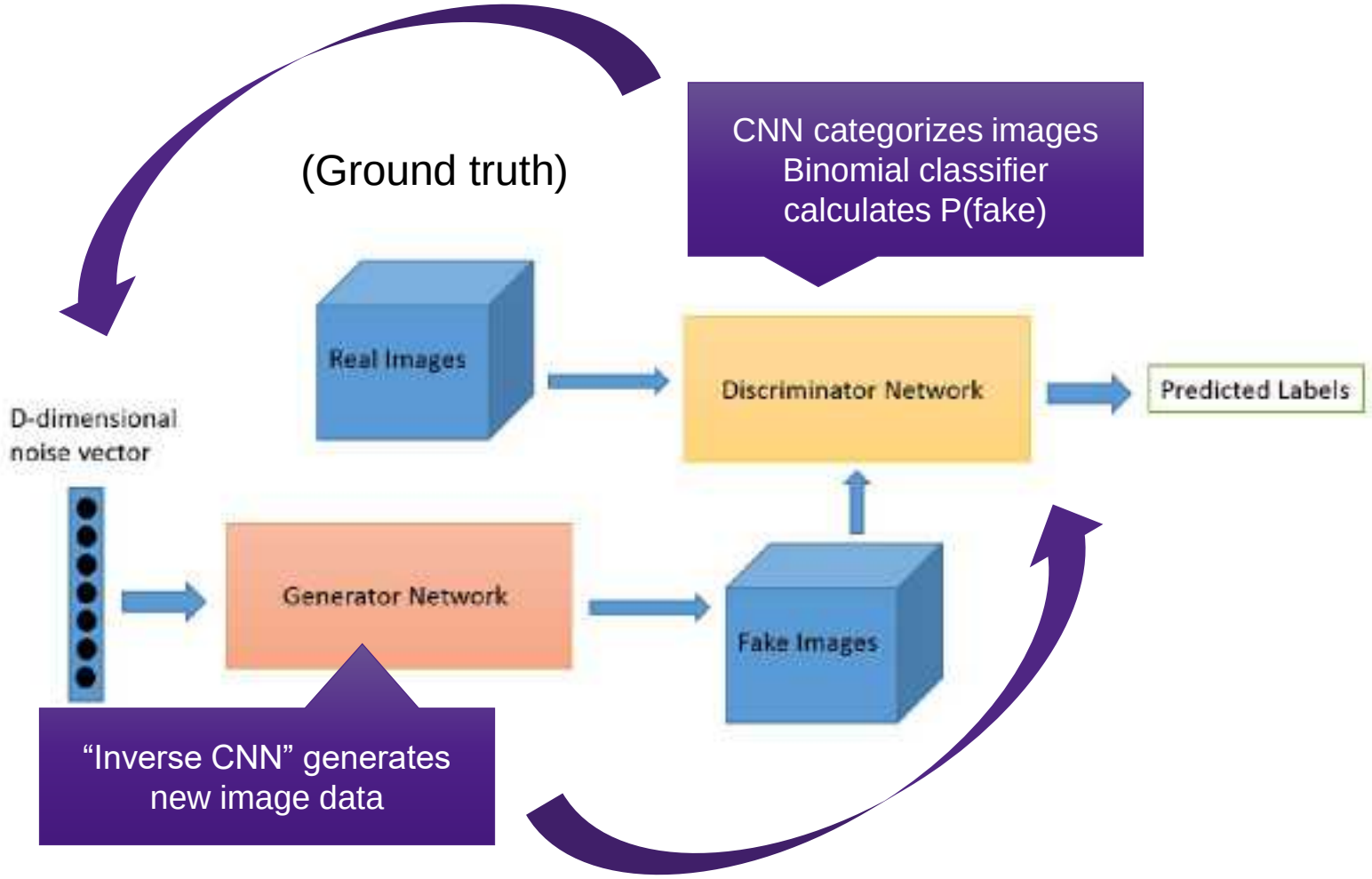
Which Face is Real?

https://skymind.ai/wiki/generative-adversarial-network-gan

- Unsupervised learning
- 2 neural nets competing
  - One generates images
  - One attempts to tell the difference
- Each network optimizes a different and opposing objective function
  - Zero-sum game

Recent AI Advancements

2 - Generative Adversarial Networks

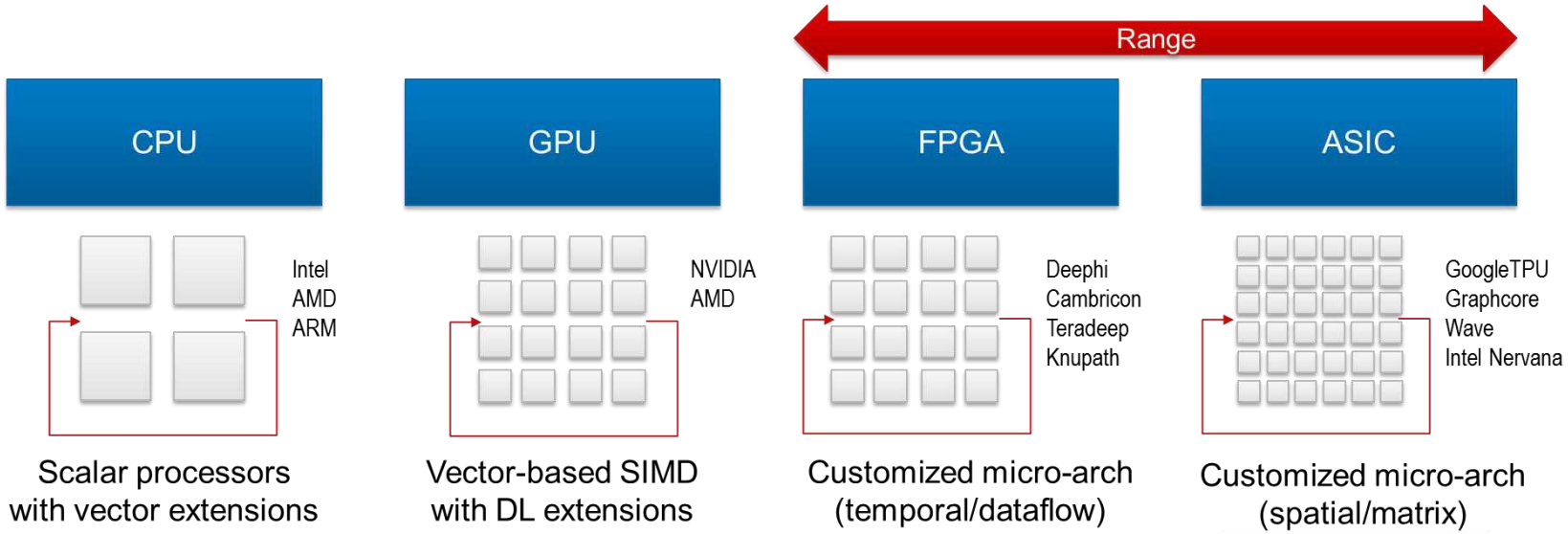


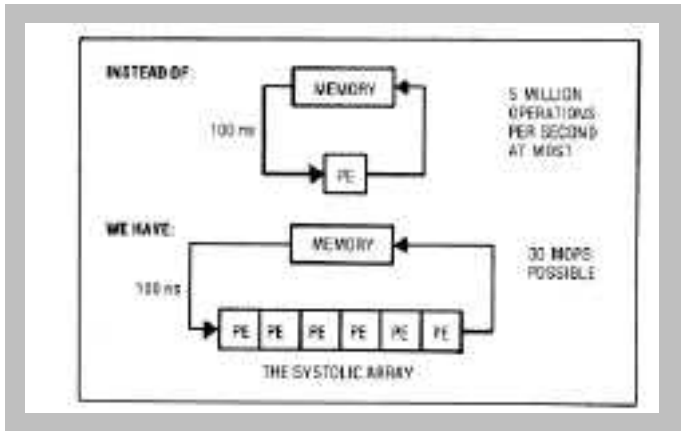
https://skymind.ai/wiki/generative-adversarial-network-gan

- Unsupervised learning
- 2 neural nets competing
  - One generates images
  - One attempts to tell the difference
- Each network optimizes a different and opposing objective function
  - Zero-sum game

Recent AI Advancements

3 - AI Accelerators

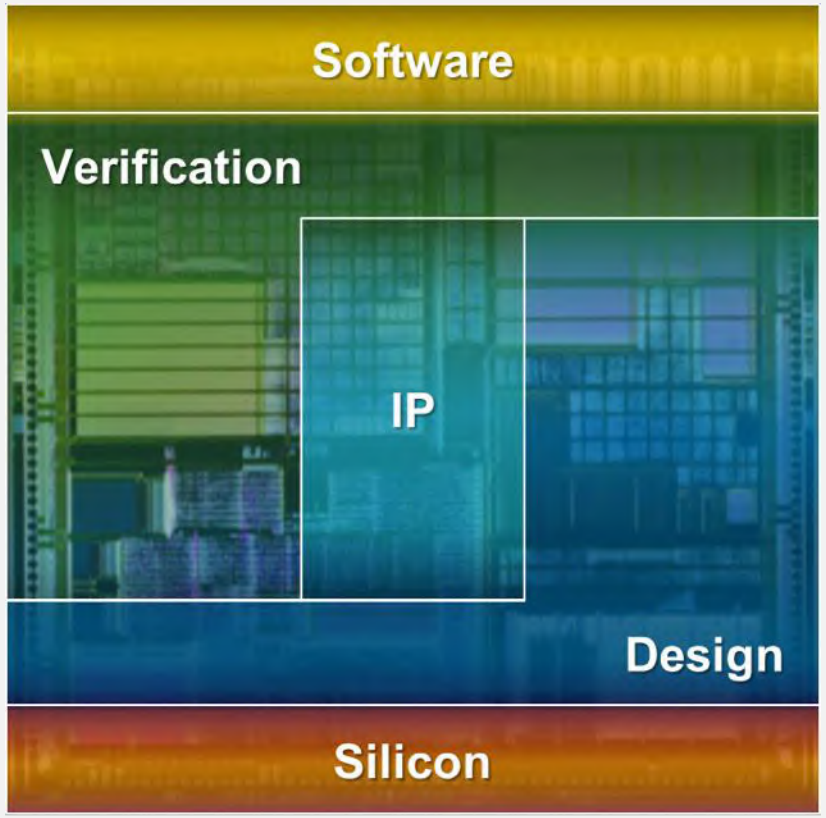




Architectures for Accelerating Deep Neural Networks, Xilinx, Hot Chips 2018

- Data center chips for deep learning training & inference
  - 20B+, 800mm^2 designs
  - Thousands of processing elements @ 100+ TOP/s
- Edge IP (primarily) for deep learning inference
  - Mixed scalar/vector/spatial compute
  - Ultra energy efficient: Several TOP/s/W

Synopsys: Silicon to Software



Software
- Application security testing & quality
- Leader in Gartner's Magic Quadrant

Verification
- Fastest engines & unified platform
- HW/SW verification & early SW bring-up

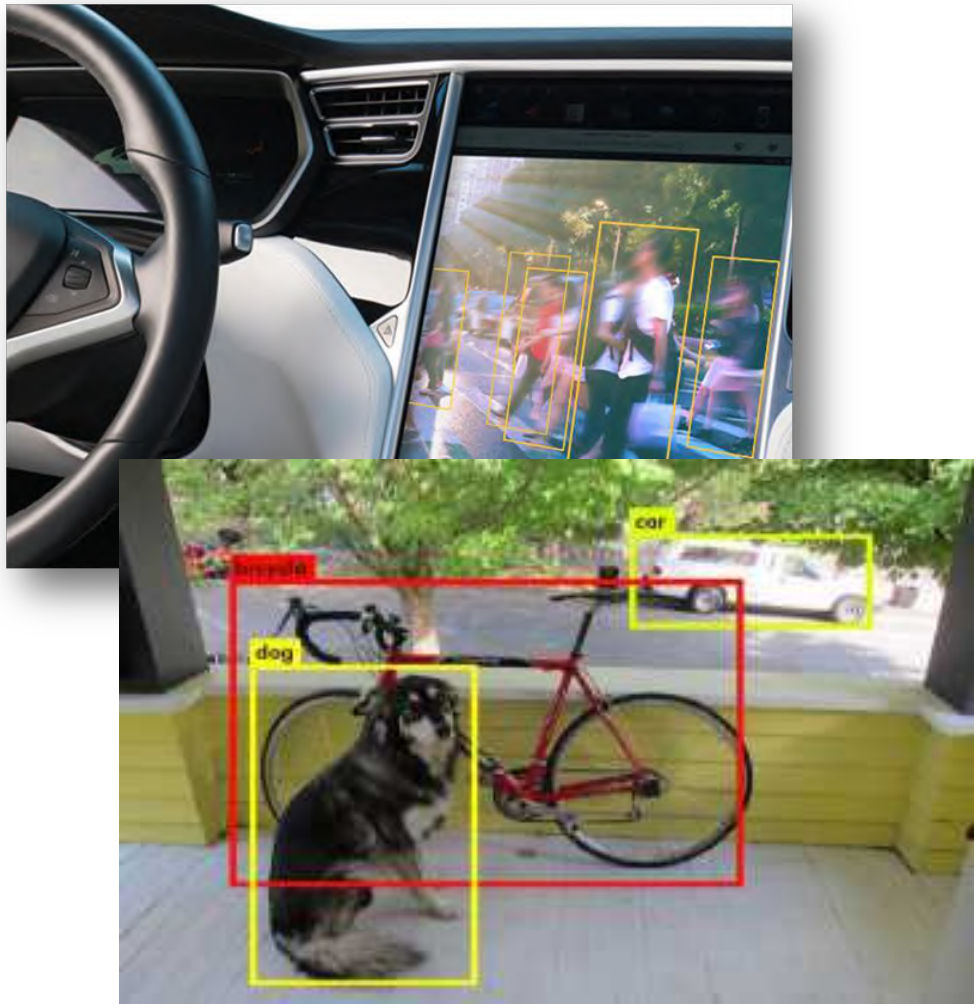
IP
- Broadest portfolio of silicon-proven IP
- #1 interface, analog, embedded mem. & phys. IP

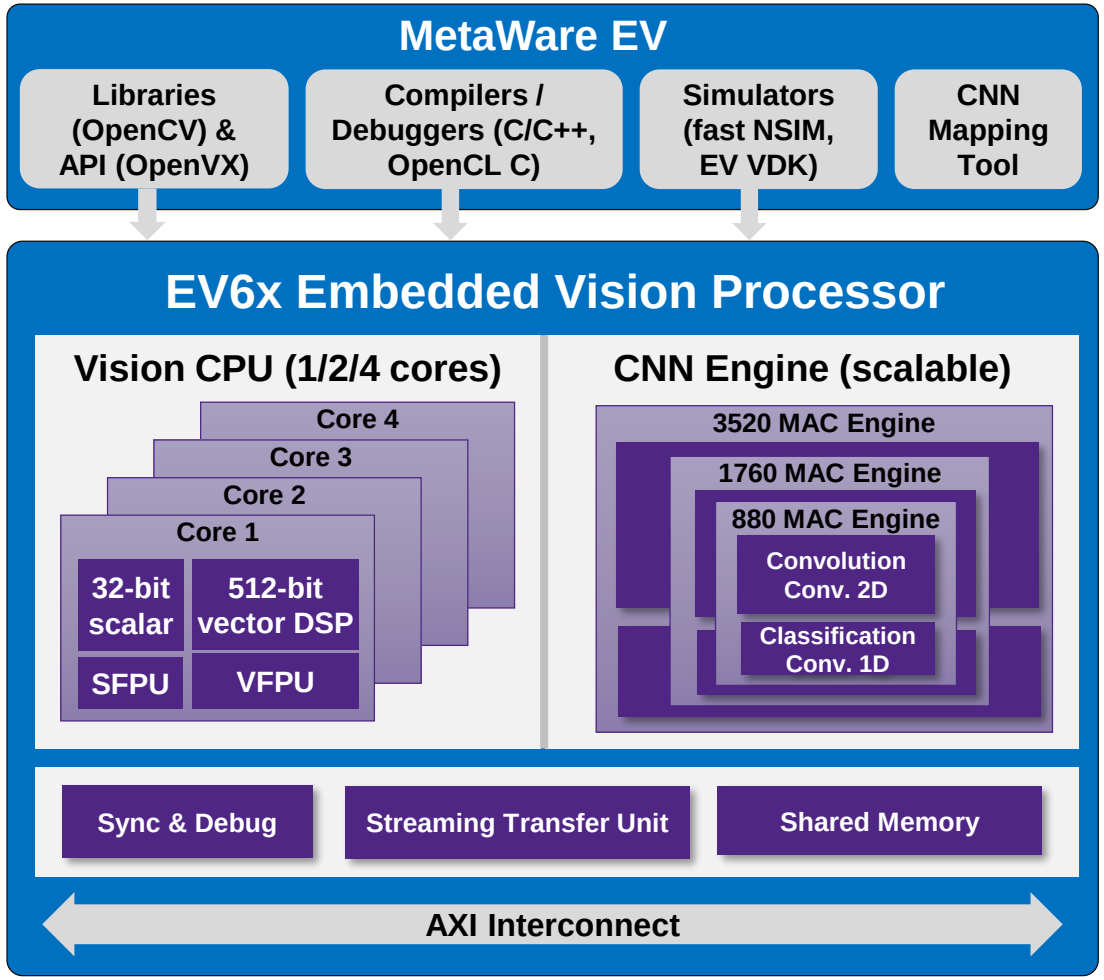
Design
- Digital & custom AMS platforms
- Best quality of results & highest productivity

Silicon
- TCAD, lithography tools & yield optimization
- Down to 5nm & below

On-Device Intelligence: AI Super-chips at the Edge

Example: Synopsys DesignWare EV6x Vision Processor





Enabling On-Device Intelligence at the Edge

Unique Requirements for Processing, Memory, Connectivity

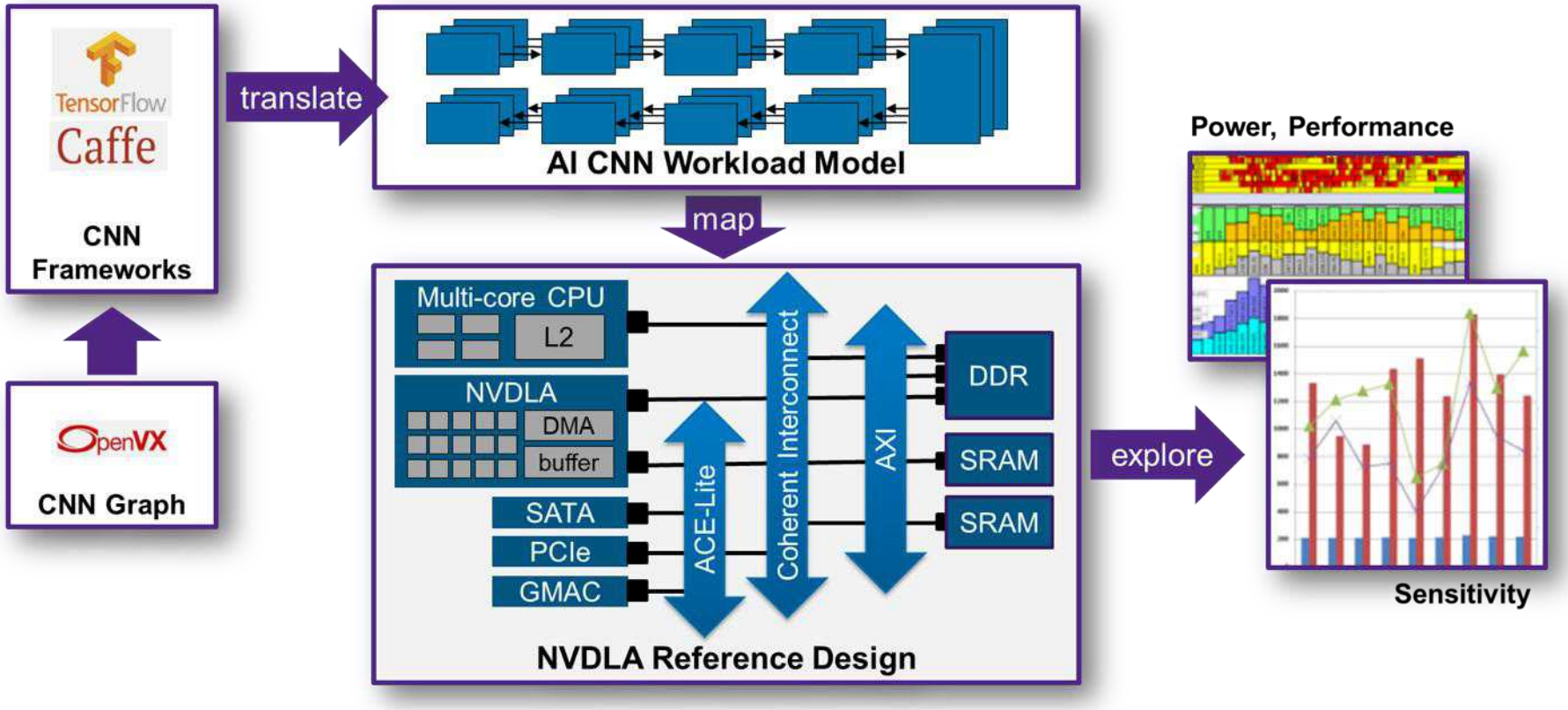
Specialized Processing

Memory Performance

Real-Time Connectivity

Early Architectural Exploration is Essential

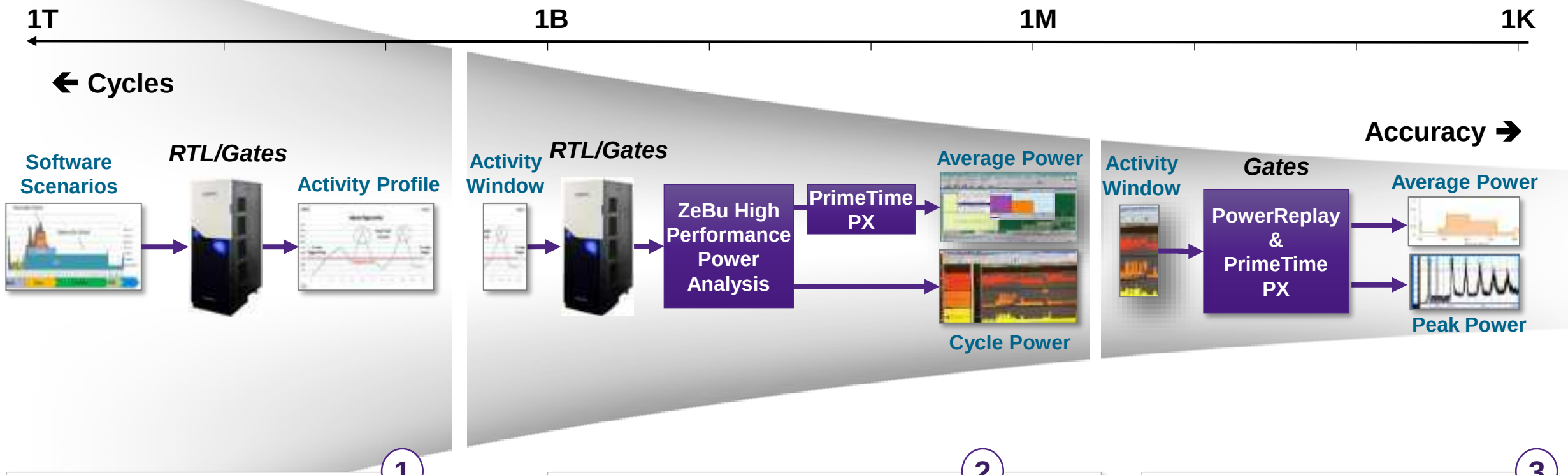
Intelligent Architectures



DLA Optimization Flow with Platform Architect Ultra

Software-Driven SoC Power Analysis

Intelligent Architectures



1

Mini-batch Activity Profile

High activity time windows over billions of cycles

2

Major Layers Average & Cycle Power

Average power and cycle-by-cycle power profile over millions of cycles

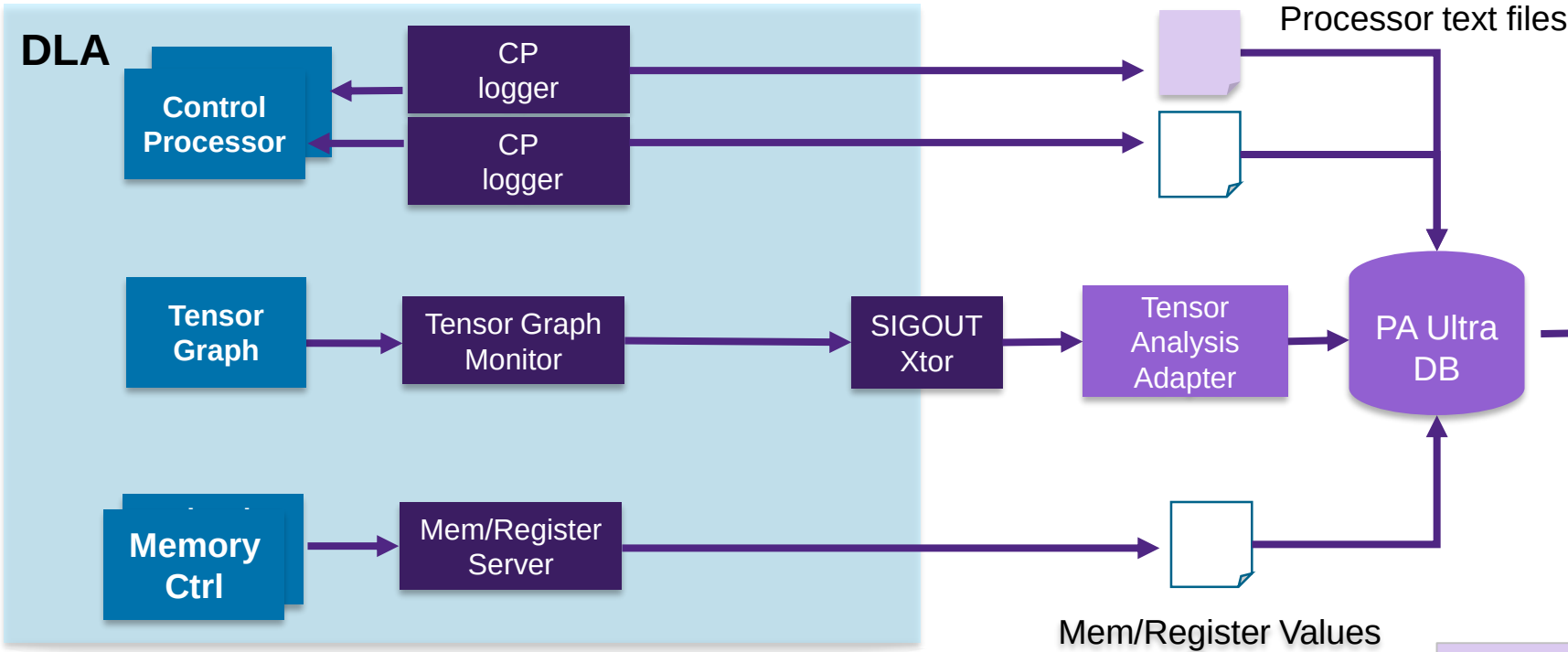
3

Power Signoff

Average and peak power signoff over thousands of cycles

Fast and Efficient Verification Infrastructure

Intelligent Architectures



ZeBu

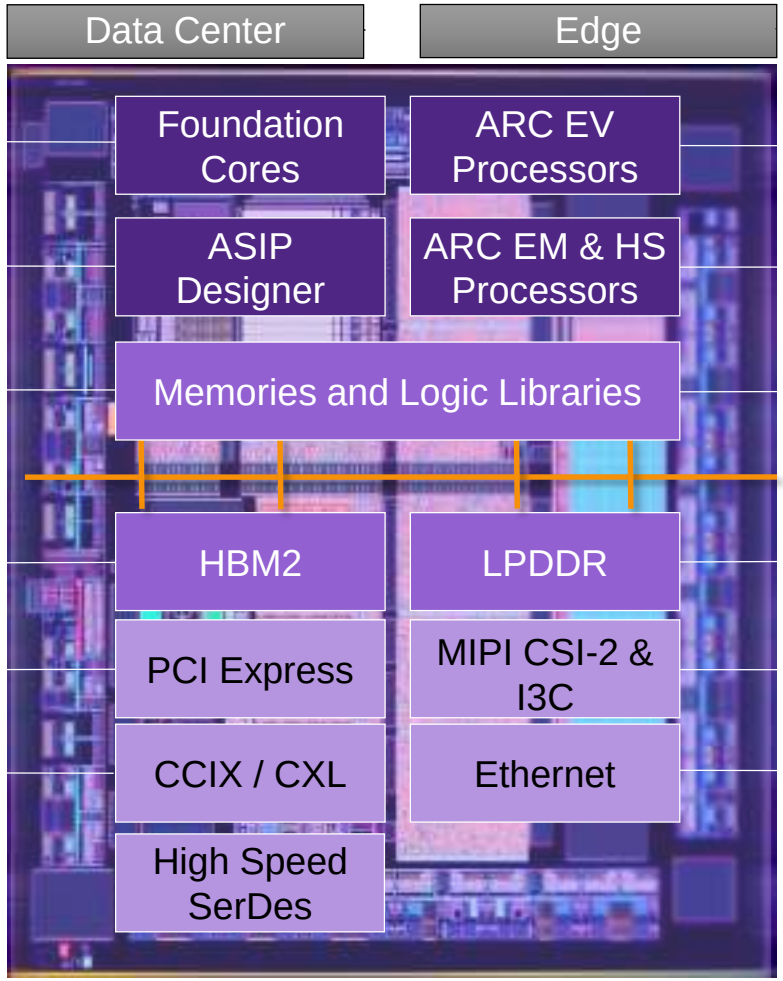
Profile Visualization

System Performance Visualized

- Runtime traceback
- Tensor Graph throughput
- Memory performance
- and correlation between these …

Critical Building Blocks for AI Acceleration

Intelligent Architectures



Flexible Deep Learning Math Primitives

Custom Deep Learning Processors

Multi-Port Memories

High Speed Bandwidth Low pJ/bit

Chip-to-chip bandwidth

Chip-to-Chip Cache Coherency

Complete Scalar, Vector, CNN engines

Pre & Post Processing & DL Acceleration

Near Threshold Libraries

Low Power DRAM

CMOS Sensor Interface

Cloud Connectivity

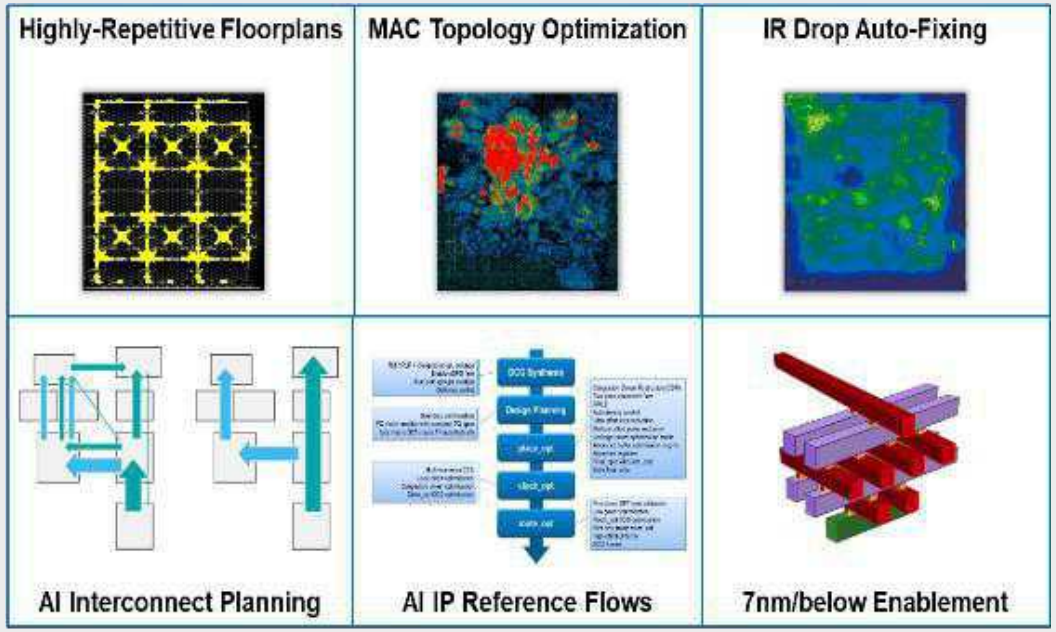
- ARC & EV Processors, ASIP Designer, and Foundation Cores for specialized processors
- HBM2, LPDDR, and Embedded Memories for optimized memory performance
- Portfolio of silicon-proven interface IP for real-time data connectivity

Achieving Best Performance-Power-Area

Intelligent Architectures

Fusion Design Platform

Technologies Enabling the Future of AI Accelerators



GRAPHCORE

The Synopsys Design Platform with Fusion Technology provides all of the capabilities we need to achieve superior processing performance for artificial intelligence and machine learning.

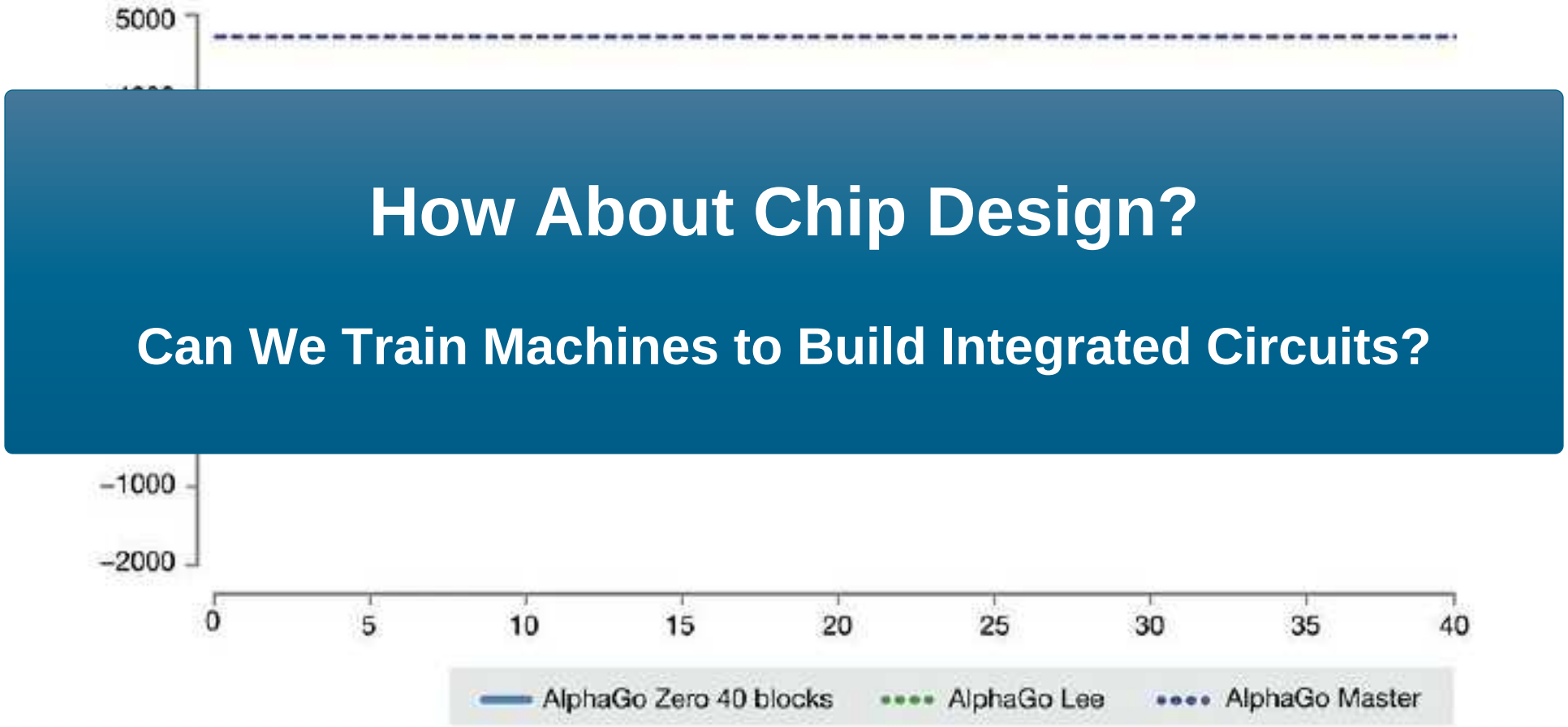
Phil Horsfield
Vice President of Silicon

Synopsys Fusion Technology Enables Superior PPA for AI Chip Design

Today's Design: A Deluge of New Challenges

Learning to play GO

DeepMind AlphaGo goes from zero to world champion in 40 days



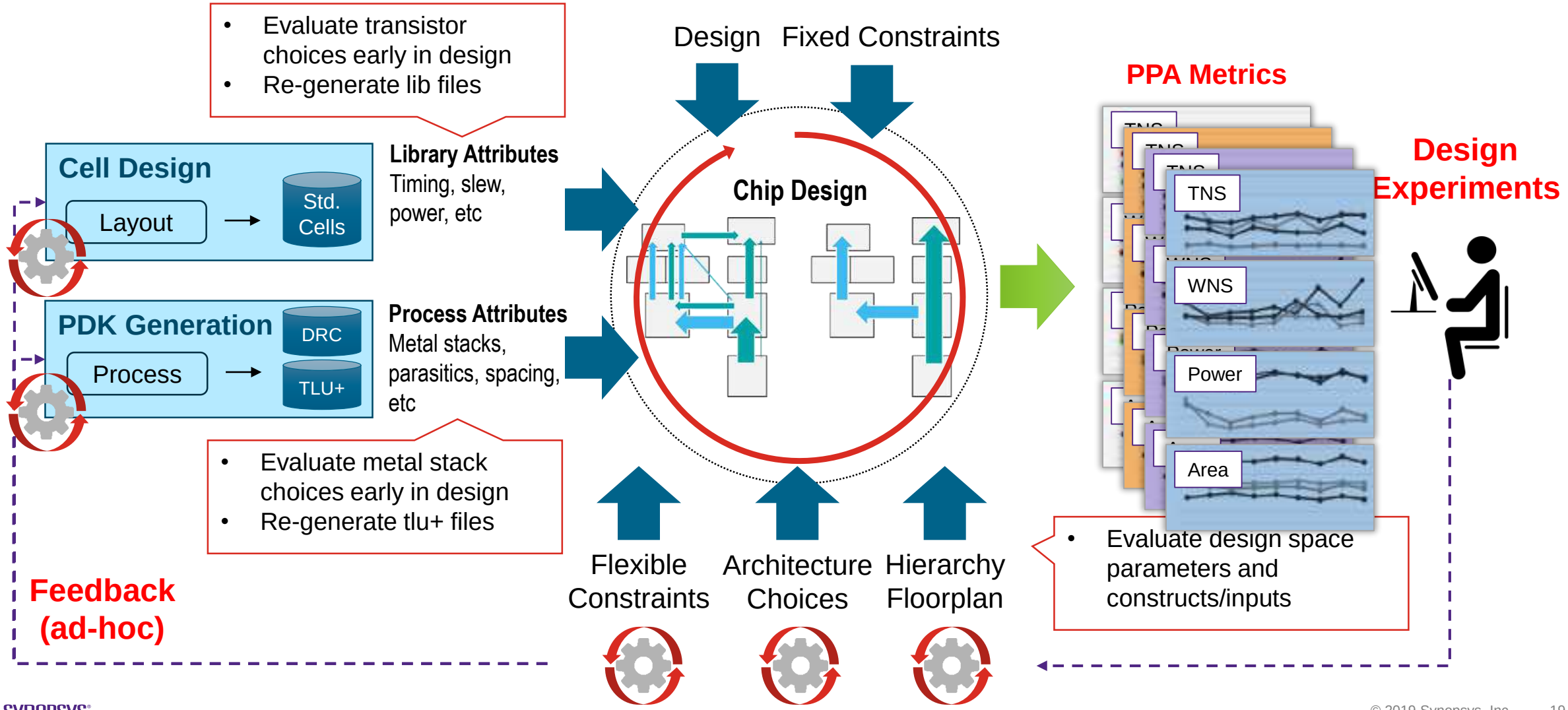
How About Chip Design?

Can We Train Machines to Build Integrated Circuits?

Example: https://deepmind.com/blog/alphago-zero-learning-scratch/

A Closer Look at the Chip Design Process Today

Intelligent design processes

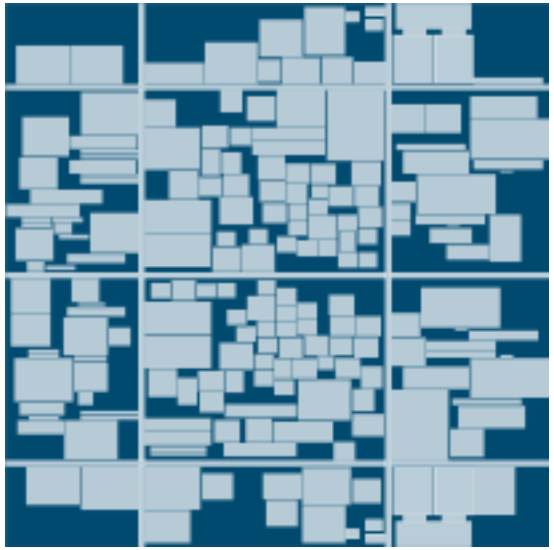


Highly Complex “Intelligent Search” Problems

Clear need to tackle the enormous problem spaces of physical design and verification

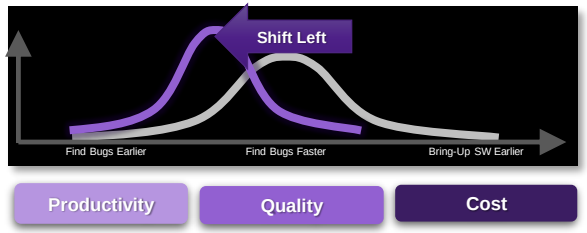
Design: Feasibility

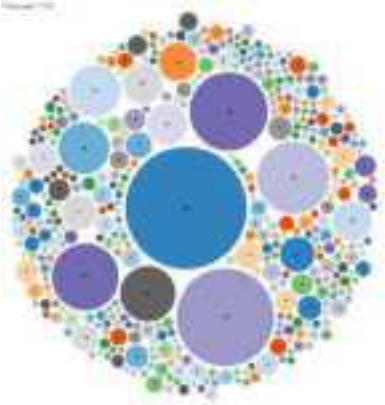
- 1,000s macros, millions of standard cells
- Complex spacing relationships
- Non differentiable response functions
- NP-complete problems



Verification: Shift Left with Exponential Growth

- Enormous volumes of verification data
- No use of results across time and designs
- Unguided and manual processes
- Computationally intractable problems

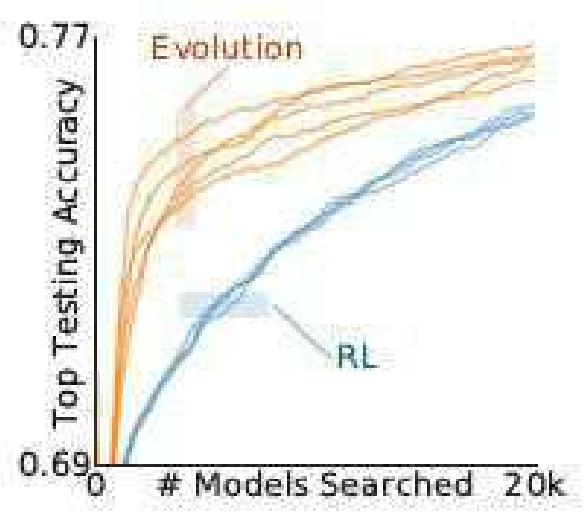




AI Becoming Pretty Good at Search

Recent success outside of EDA paving the way for design and verification innovation

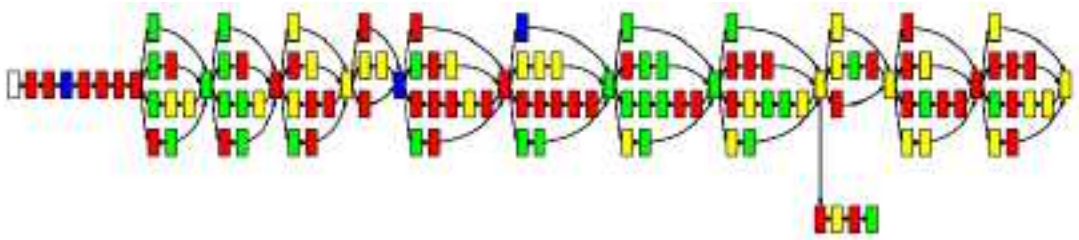
Example: Neural Architecture Search



- Automatically discovers image classifiers
- AmoebaNet-A surpasses hand crafted NNs
- 96.6% (#1) ImageNet accuracy, ~40B FLOPS

Regularized Evolution for Image Classifier Architecture Search
Esteban Real, Alok Aggarwal, Yanping Huang and Quoc V. Le, AAAI 2019, https://arxiv.org/abs/1802.01548v7

Example: Neural Acceleration Search



- Automatically discovers best scheduling for neural network computational graphs
- 20% faster runtime vs. experts, self-trains in hrs

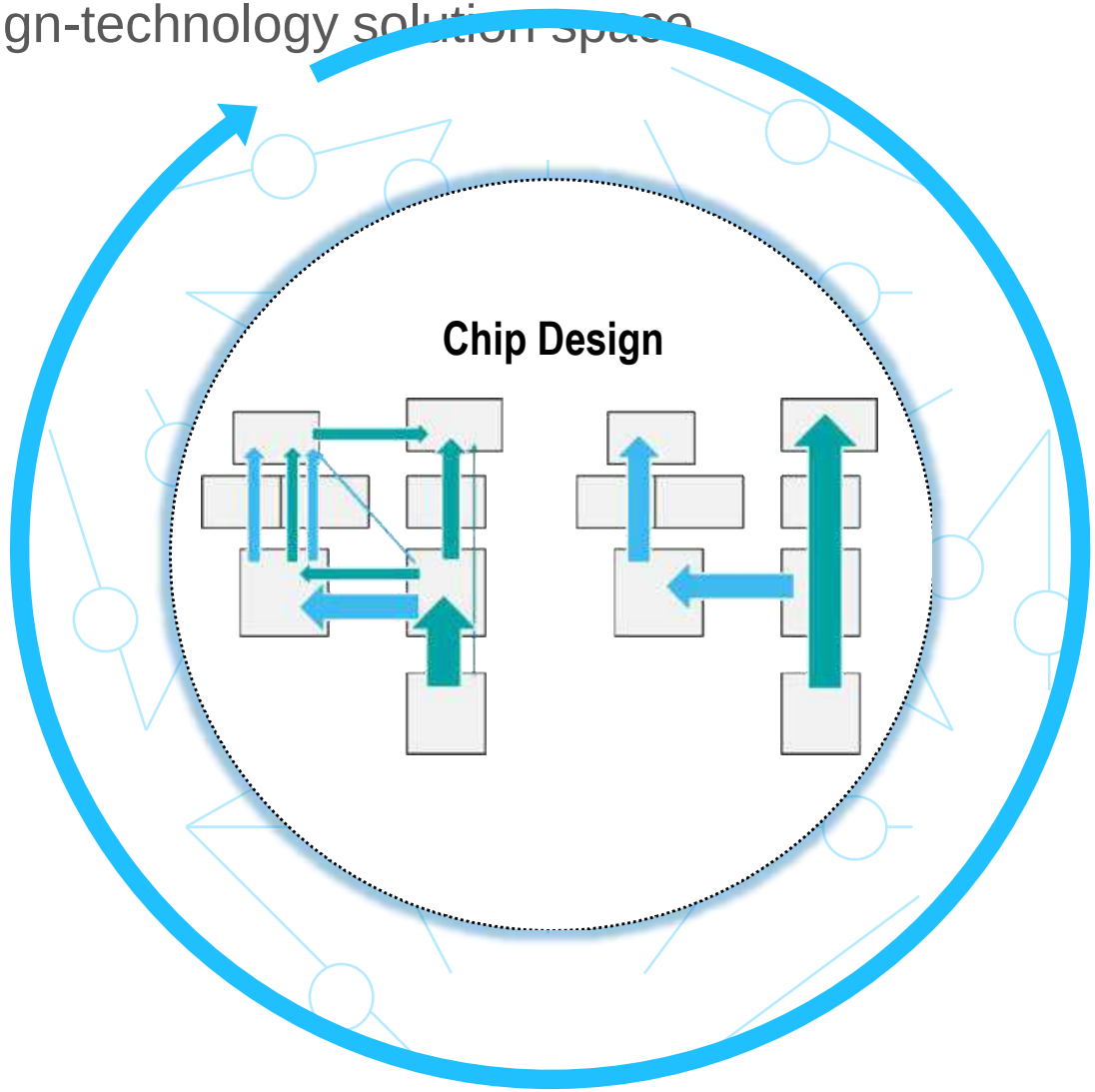
Device Placement Optimization with Reinforcement Learning,
Azalia Mirhoseini, Hieu Pham, Quoc Le, Mohammad Norouzi, Samy Bengio, Benoit Steiner, Yuefeng Zhou, Naveen Kumar, Rasmus Larsen, and Jeff Dean, ICML 2017, arxiv.org/abs/1706.04972

Towa lf-Optimizing Design Processes

Brea lesign-technology solution space

Architecture

Flow



Process

Constraints

Summary: Enabling On-Device Intelligence

- An explosion of abundant-data computing is creating disruptions across the semiconductor design space
- On-device intelligence at the edge is no playground; intelligent machines require intelligent architectures and design processes
- Machine-learning enabling new ways of thinking about design, breaking up silos across design and technology
- Synopsys working with AI pioneers to enable the future of AI hardware

www.synopsys.com/ai


SYNOPSYS®
Silicon to Software™