AI 100
CBINSIGHTS

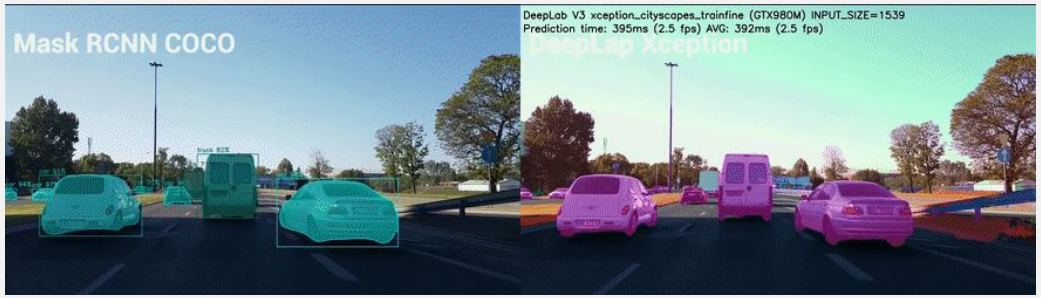
Deeplite

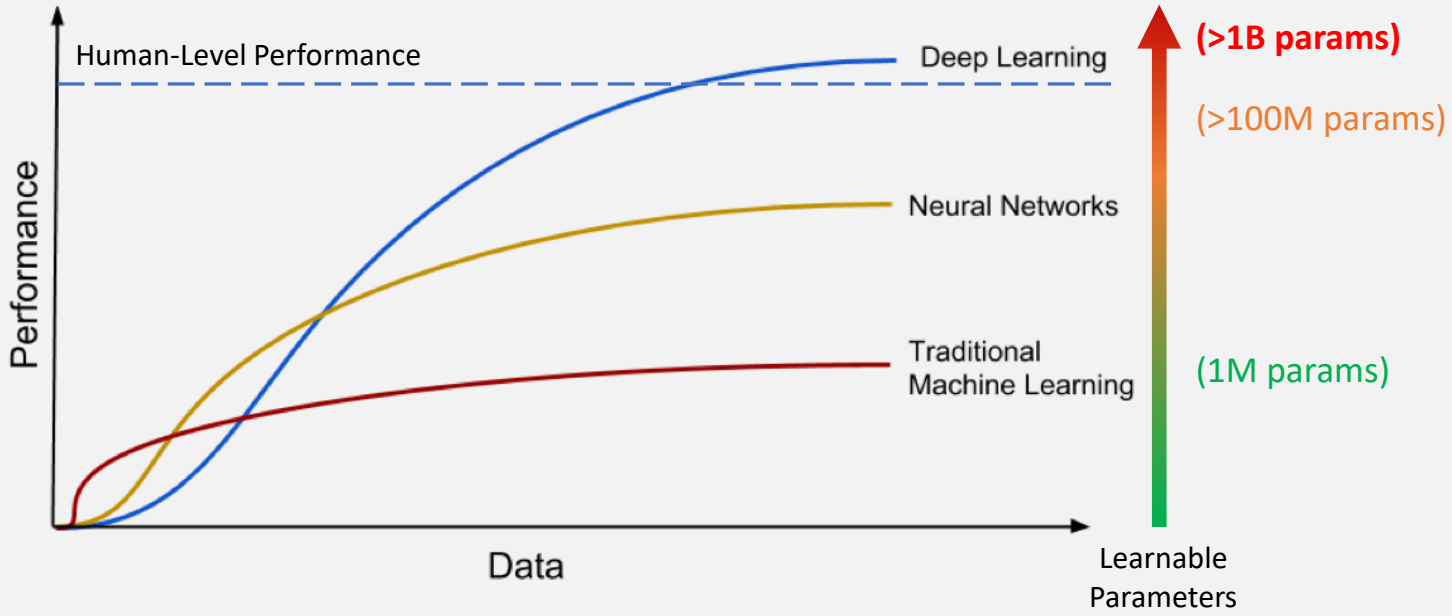
Lightweight Intelligence™ for everyday life

Ehsan Saboori, Ph.D.
Co-founder & CTO

Deep Learning Drives AI

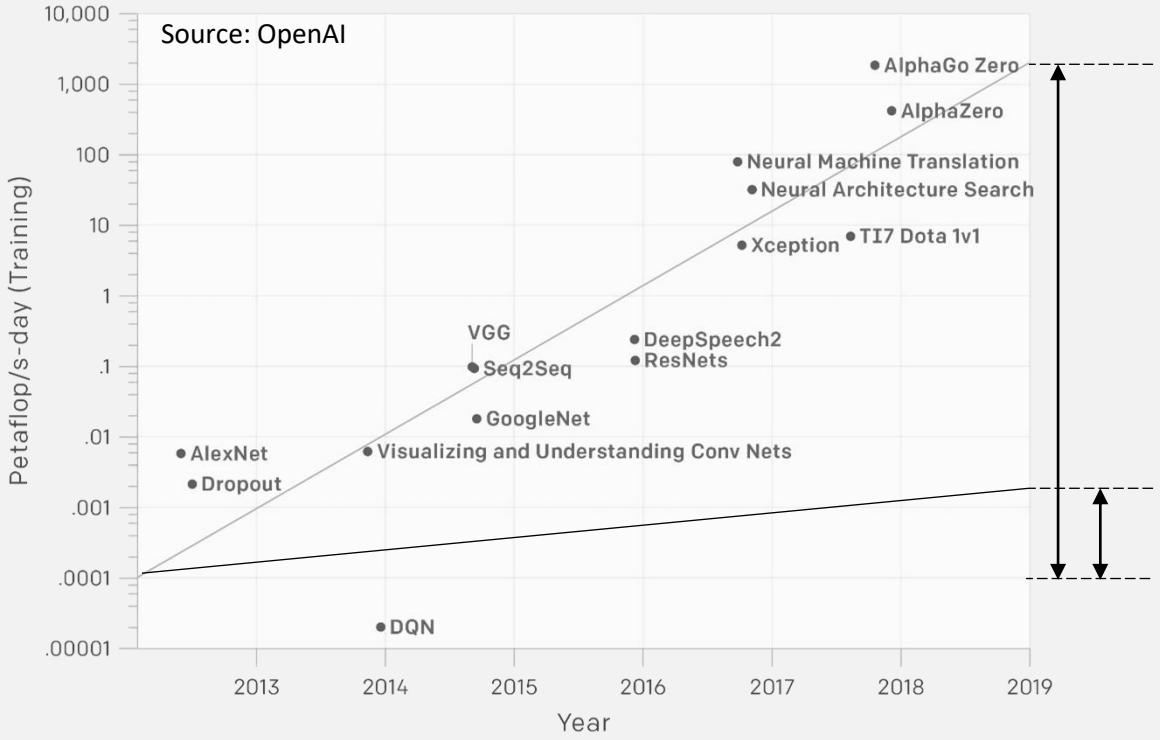
- Why Deep Learning?
  - Automated feature extraction
  - Scale improves performance





Deep learning models are growing rapidly

- Deep learning outperforms humans, but comes with **huge compute cost**
- **Deeper** neural network, **better** accuracy, **more** compute required



AlexNet to AlphaGo Zero:
A **300,000X** Increase in Compute

Moore's law
A 1**0X** Increases in # of Transistors

Edge Computing Challenges

High Computational Complexity

Millions of expensive floating-point operations for each input classification are needed.

Memory Footprint

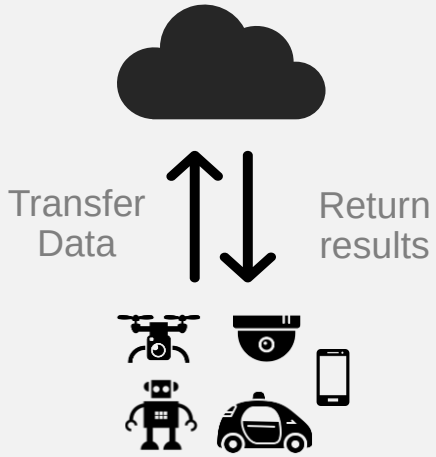
Huge amounts of weights and activations with limited on-chip memory and bandwidth.

Power consumption

Deep learning requires significant power and can easily consume battery life

These demands force AI to the cloud

- **Expensive hardware** required for deep learning
- **Huge power consumption** for cloud AI hardware
- **Real-time critical AI** cannot rely on the internet connection



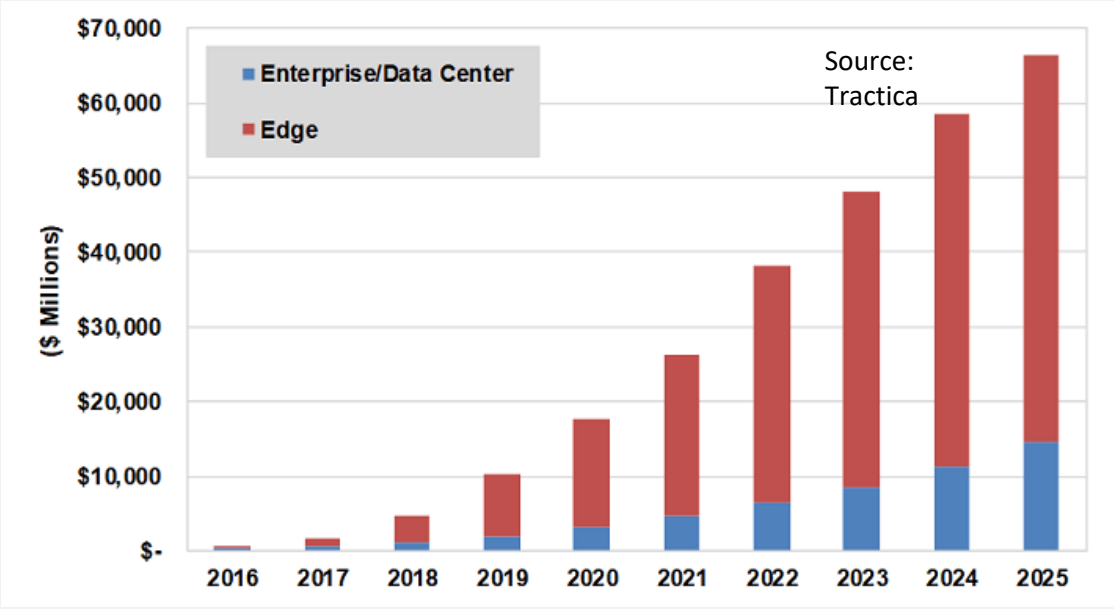
Typical Edge AI application workflow

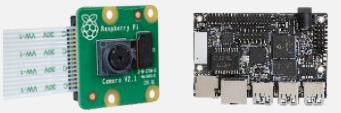
Memory Footprint	~>10G
Power Consumption	>~300w
Computational Complexity	> 100 TOPs
Cost (ASP)	> $5,000

Typical Cloud HW

Time to deploy AI on edge devices

- Massive value unlocked by making AI applicable for cost-effective hardware
- **AI inference must meet strict power, speed, cost and resource constraints**



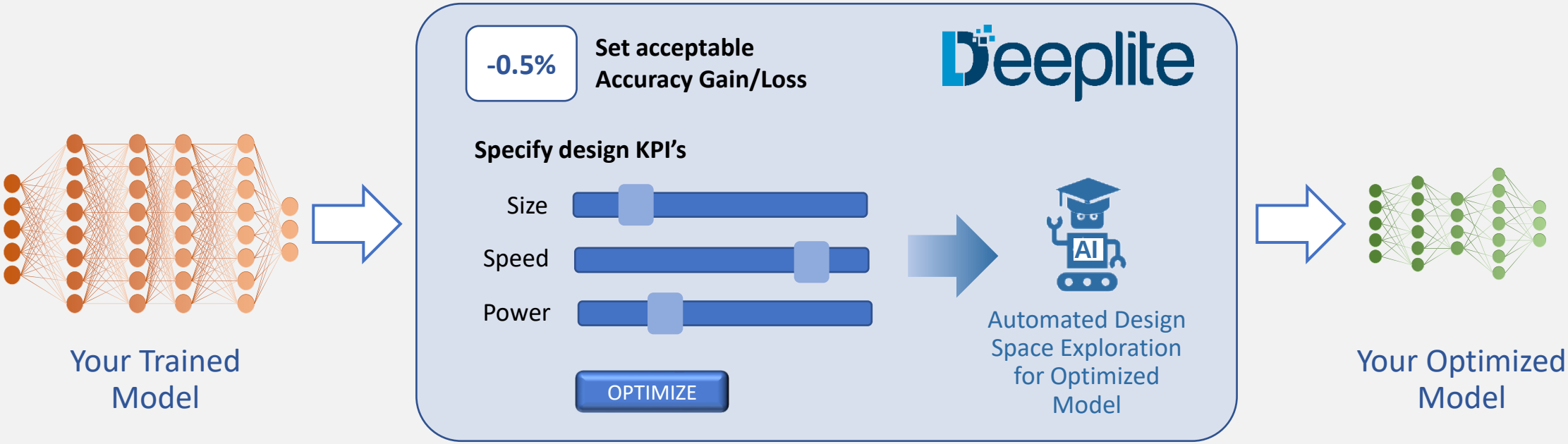


Memory Footprint	~<1M
Power Consumption	~<10w
Computational Complexity	~<10 TOPs
Cost (ASP)	~$10

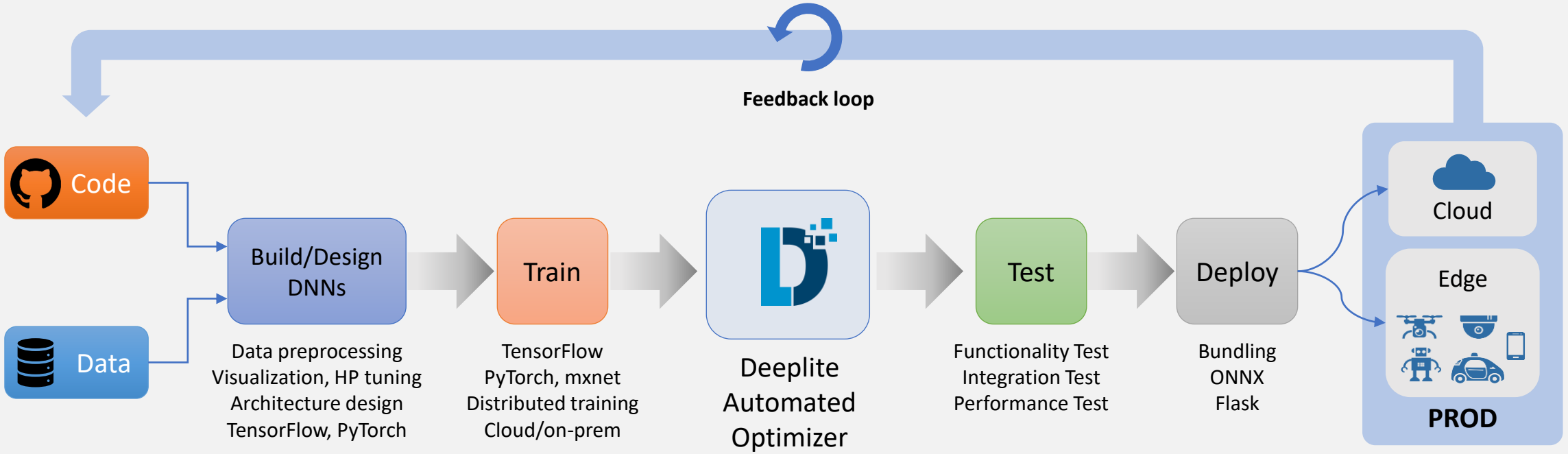
Typical Edge HW

Lightweight Intelligence™ for Edge Devices

Deeplite provides an automated optimizer for AI engineers to automatically create faster, smaller & more efficient model architectures for production edge devices.

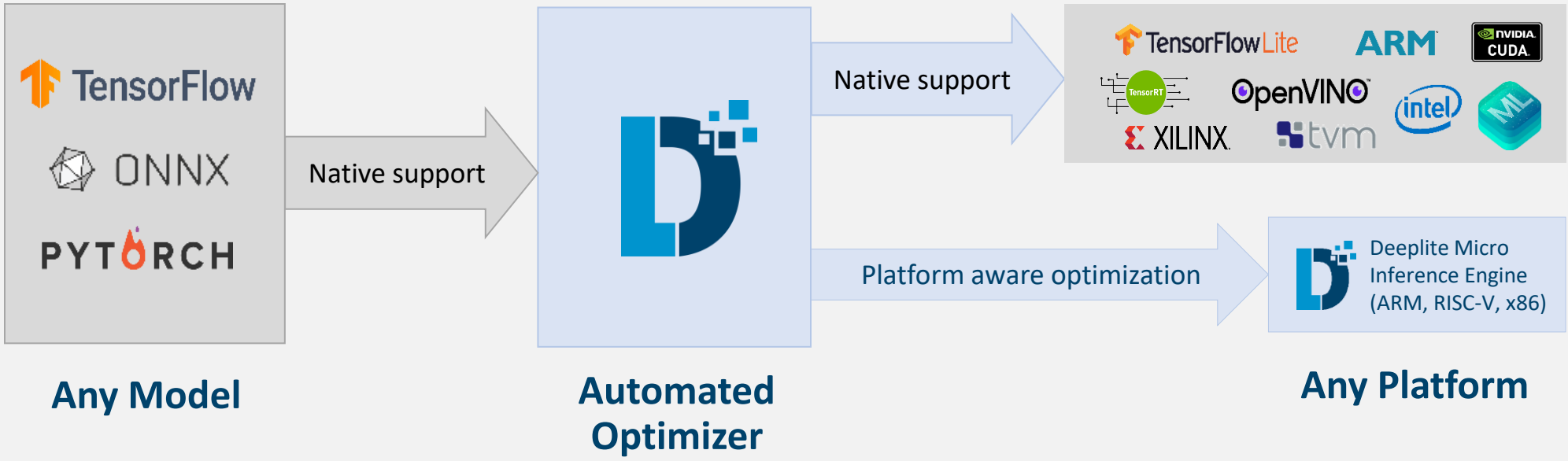


Where we fit in a ML/AI Workflow

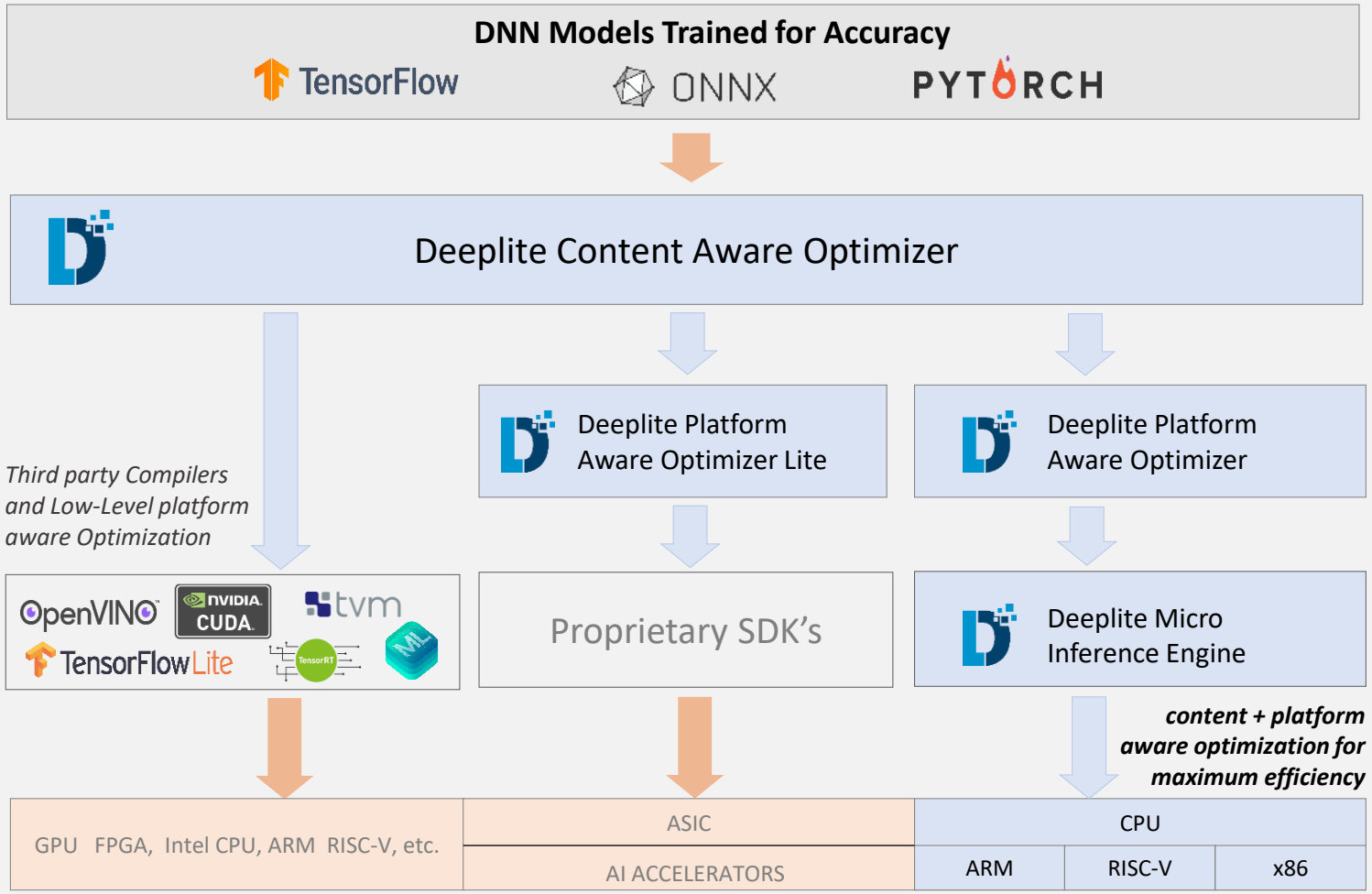


Agnostic Auto Optimizer

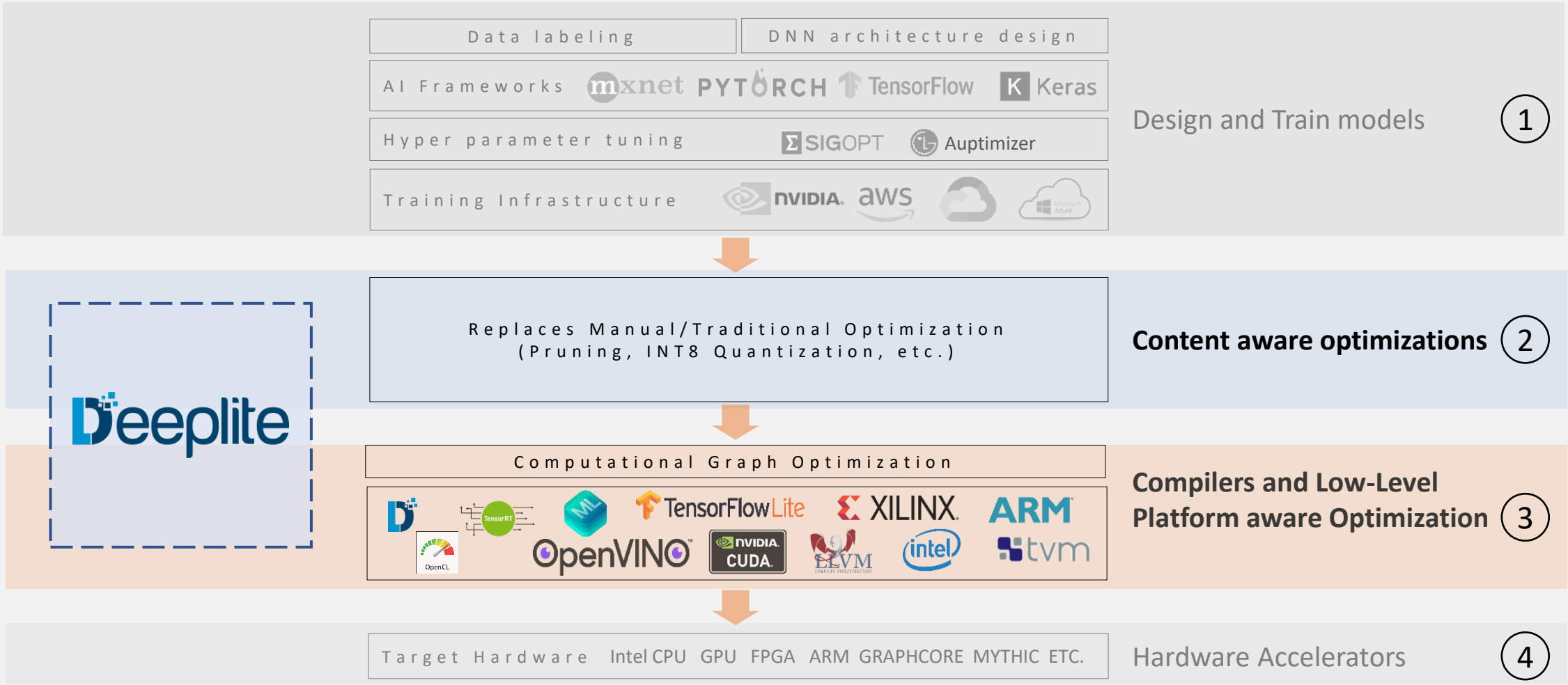
- Support different AI frameworks
- Support different hardware backend



Deeplite Solution Stack



Levels of Optimization

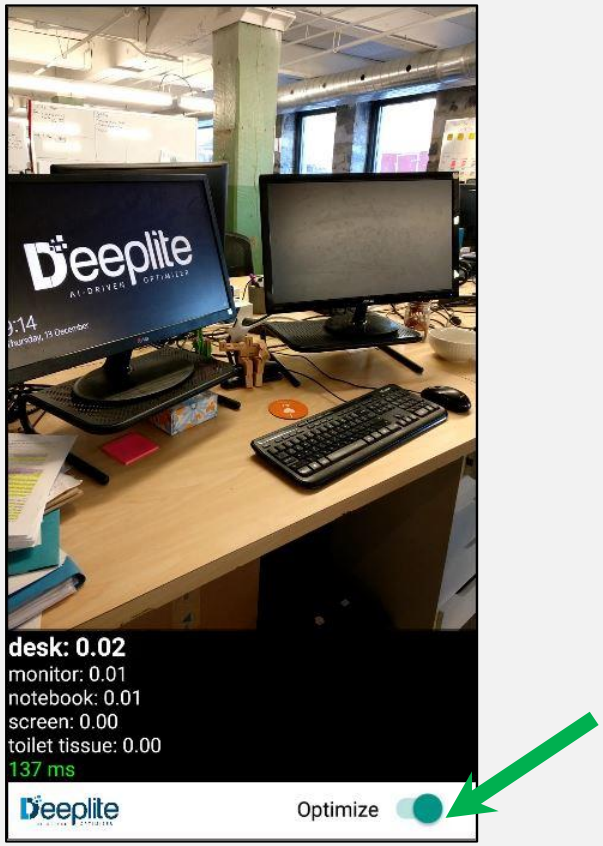


Computer Vision Results

Model	Compression[4]	FLOPs Reduction	Accuracy Drop (%)	Dataset
VGG19	**x84**	**x10**	<1%	CIFAR100
Resnet50	**x30**	**x6**	<1%	CIFAR100
Resnet18	**x33**	**x8**	~0%	CIFAR100
Mobilenet-v1.0	**x22**	**x5**	<1.5%	VWW
Industry use case[1]	**x60**	**x8**	<2%	Subset of Imagenet
Industry use case[2]	**x55**	**x100**	~0%	Custom dataset
SSD300[3]	**x8**	**x6**	~0%	Subset of COCO2017

[1] Based on Resnet18 architecture
[2] Custom architecture
[3] Resnet-50 backbone
[4] No quantization used for these experiments. An additional 4x gain available by using basic quantization techniques.

10x Speedup on mobile CPU



Optimized vs. Unoptimized model on Android phone

Deeplite Joint Project with Industry

- Optimize SSD300 object detection model with ResNet-50 backbone for autonomous vehicle deployment

Original Model:
- SSD300 **(56.7MB)**

Platform:
- NVIDIA Xavier GPU

Requirements:
- Accuracy drop < **1.0%**

Deeplite Results:
- Size = **4.8MB (12 times smaller)**
- Power reduction ~ **3x**
- Speed up ~ **3x**
- Accuracy drop ~ **0%**

Cost: $700
Power: 10-30w

Andes & Deeplite Joint Project

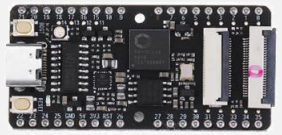
- Embedded solution where a home assistant “wakes up” when it detects a person via a small camera (link to press release)

Original Model:
- Mobilenet-v1.0 **(12.8MB)**

Platform:
- Andes RISC-V CPU cores

Requirements:
- Size < **256KB**
- Accuracy drop < **2.0%**

Deeplite Results:
- Size = **188KB (69 times smaller)**
- Accuracy drop ~ **1.0%**



Cost: $40
Power: 3w

Andes Technology, a leading Asia-based supplier of high-performance low-power compact 32/64-bit RISC-V CPU cores and a founding Platinum member of the RISC-V Foundation

Thank you!

For more information about Deeplite please contact:

info@deeplite.ai

Deeplite