Agenda

- What is machine learning (ML)?
- Machine learning at the edge
- Implementing machine learning
- Software, ecosystem, and resources
- Machine learning in silicon design
What is Machine Learning?

- Additional terms
- Location
  - **Cloud** – processing done in data farms
  - **Edge** – processing done in local devices (growing much faster than Cloud ML)
- Key components of machine learning
  - **Model** – a mathematical approximation of a collection of input data
  - **Training** – in deep learning, datasets are used to create a “model”
  - **Inference** – in deep learning, a “model” is used to check against new data
For each piece of data used to train the model, millions of model parameters are adjusted. The process is repeated many times until the model delivers satisfactory performance.
When new data is presented to the trained model, large numbers of multiply-add operations are performed using the new data and the model parameters. The process is performed once.
Compute Requirements Differ for Training and Inference

Cloud side

Train the model | Throughput oriented | High performance | Large data sets

High-performance compute, servers, GPU

Device side

Apply the model | Low/mid performance | Low latency Efficiency | Security Privacy

Embedded systems, heterogeneous processing
Machine Learning at the Edge
Why ML is Moving to the Edge

Bandwidth | Power | Cost | Latency | Reliability | Security
Range of “Edge” Applications

- **Increasing power and cost (silicon):**
  - Keyword detection (~1-10mW)
  - Pattern training (~1W)
  - Object detection (~10W)
  - Voice & image recognition (~100W)

- **Increasing performance (Ops/second):**
  - Autonomous drive (~100W)
  - Image enhancement (~10W)
  - Voice & image recognition (~1W)
  - Object detection (~10W)
  - Pattern training (~1W)
  - Keyword detection (~1-10mW)
Machine Learning Innovation

Hive View

Nauto for Safe Driving

BRAGI 'The Dash PRO'

DriveCore Platform

EZVIZ C5Si Camera

ReindeerCam
Increasing performance (ops/second)

- Keyword detection
- Pattern training
- Object detection
- Voice and image recognition
- Image enhancement
- Autonomous driving
- Data center

Increasing power and cost (silicon)

Flexible, Scalable ML Solutions
Targeting Multiple Markets with Scalable Architecture

Typical Market Performance Requirements

- **IoT**: ~20 GOP/s
- **Mobile**: 1~3 TOP/s
- **Industrial**: ~20-50 TOP/s
- **Automotive**: > 70 TOP/s
- **Networking**: > 70 TOP/s
- **Server**: > 70 TOP/s

Scalable Machine Learning Processor “Architecture”

- **Sensors (2 GOP/s)**: Scalable
- **Mobile (1~3 TOP/s)**: Compatible
- **Industrial (~20-50 TOP/s)**: Programmable
- **Automotive (> 70 TOP/s)**: Programmable
- **Networking (> 70 TOP/s)**: Programmable
- **Server (> 70 TOP/s)**: Programmable
Implementing Machine Learning
Machine Learning Tradeoffs in Embedded Systems

Considerations

- Machine learning
- Software
- Hardware
- System

Neural Networks

- Number of outputs
- Number of inputs
- Number of hidden layers
- Data type
- Transfer function
- Accuracy
- Neurons per hidden layer
- HW/SW
- Memory

System Considerations

- Latency
- Communication bandwidth
- Energy efficiency
- Privacy
- Security
- Scalability

Optimization of system and machine learning requirements
Heterogeneous Systems Enable Efficient ML

Sophisticated end node based on a heterogeneous architecture

Efficient “always-on” subsystem filters and processes data using ML

Advanced ML using Applications processor

Applications Subsystem

Always-on subsystem Subsystem

Shared L2

Interconnect

DMC

DDR

AHB/AXI

Local memory

AHB interconnect

Sensor

Timer

SRAM

Compute performance

Time
Software, Ecosystem, and Resources Support
Machine Learning Needs a Strong Ecosystem

Apps and Services

Frameworks and Models

Hardware Platforms
Frameworks simplifying ML development

Arm NN software translates existing NN frameworks:
- TensorFlow, Caffe, Android NNAPI, MXNet etc.
Developers maintain existing workflow and tools
Reduces overall development time
Abstracts away the complexities of underlying hardware

CMSIS-NN 5x
better efficiency and performance for NN functions

Compute Library 15x
faster than other open-source software (OSS)
Design Challenges That Are Faced Today

- **EDA wrestles with many data-driven problems**
  - Need to provide answers quickly, in-design solutions
  - Heavy degree of automation is needed
  - What-if analysis to explore alternative solutions faster

**Scale**
- Larger designs
- More rules and restrictions
- More data (simulation, extraction, shapes, techfiles)

**Productivity**
- More uncertainty leads to re-design and missed schedules
- Limited number of capable, trained design and layout engineers
- Larger complexity and scale... more activities in same schedule

**Complexity**
- More complicated silicon technologies (FinFET)
- More complicated design and electrical rules
- More interactions between chip, package and board
- Thermal related impact between devices and wires

One Impacts the others
Data-Driven Solutions = ML, Analytics, DM

Rarely is ML used in a Vacuum in Real Products and Systems

Creating an Intelligent and Adaptive Product Requires More than Machine Learning

Steps in Producing Intelligent Solutions

Prepare Data
- Data Preparation
  - Filter
  - Label
  - Augment
  - Reduce
  - Partition

Build Models
- Model-Based Inference
  - Select
  - Train
  - Test
  - Verify

Adapt to Uncertainty
- Adaptation
  - Calibrate
  - Learn
  - Optimize

We often say “Machine Learning,” but that implies some combination of analytics, ML, optimization, and distributed processing
Machine Learning at Cadence

Inside
• Better PPA, faster engines
• Improved testing/diagnostics

Outside
• Automated design flow
• Productivity improvement

Enablement
• Hardware/software co-design
Intelligent Design and Optimization

- Capture more abstractions of design intent and preferences
- Massively Parallel Distribution
- Electrically Driven Optimization
- Set of Design Alternatives Aligned with Intent
- Intent Captured As Design Constraints
- Cost Function
- Device Extraction
- Device and Interconnect Extraction
- Electrallv Driven Routing
- Electrallv Driven Placement

Key Technologies for Intelligent Solutions:
- Create knowledgebase that can be mined and statistically analyzed in future planning and design
- Local Session
  - Fast extraction and electrically aware assistance
- CPU
- GPU
- Massive Parallelization (Cloud-Enabled)
- Solutions
- Machine Learning
- Analytics and DM
- Optimization

Analytics

Layout Creation In Virtuoso Layout EAD

Observe
Recommend
Action or Decision
Machine Learning for Library Characterization

Challenges
- Many cells, many process corners
- Many new effects
- Millions of simulations for each new standard cell library

Machine learning improves throughput:
- Learn from previous process corners
- Smart interpolation by extracting critical measurements
- Critical corner identification
Summary

• Machine Learning is already enabling innovation

• ML at the edge is enhancing intelligence in even the smallest of devices

• Strong ecosystem of ML solutions

• ML is already enhancing silicon design

• Talk to Arm and Cadence about ML solutions
Thank You
Danke
Merci
谢谢
ありがとうございます
Gracias
Kiitos
감사합니다
धन्यवाद
תודה
The Arm trademarks featured in this presentation are registered trademarks or trademarks of Arm Limited (or its subsidiaries) in the US and/or elsewhere. All rights reserved. All other marks featured may be trademarks of their respective owners.

www.arm.com/company/policies/trademarks