70% of the world’s population uses Arm technology
The architects of global possibilities

The global leader in the development of licensable technology
- R&D outsourcing for semiconductor companies

Focused on freedom and flexibility to innovate
- Technology reused across multiple applications

With a partnership based culture & business model
- Licensees take advantage of learnings from a uniquely collaborative ecosystem

>1,400 licenses, growing by >100 every year

17.7bn Arm-based chips shipped in FY2016

>460 licensees
Industry leaders and high-growth start-ups; chip companies and OEMs
The road ahead is exciting

100 billion chips shipped

1991

26 years

2017

100 billion chips shipped

2021

4 years!
Distributing intelligence from edge to cloud

- On-device learning for enhanced user privacy
- Compute performance to deliver a hi-fidelity world
- Real-time inference for autonomous systems
- Security and privacy for your data
- 4k, HDR and 5G for more human-like interfaces
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, data-sets 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.
Neural Networks (NN) Can Now Outperform Humans

Deep learning introduced in 2012, resulting in big improvements.

Error rates have now stabilized at ~3%
The Smartphone is the World’s Most Popular AI Device
90% of AI today runs on smartphones* and 95% of the world’s smartphones run on Arm.

- Speech recognition
- Predictive text
- Face tracking camera
- Digital assistant
- Augmented reality
- Fingerprint identity

*IDC research

100x compute increase since 2009
The Smartphone is the World’s Most Popular AI Device

90% of AI today runs on smartphones* and 95% of the world’s smartphones run on Arm

- Speech recognition
- Digital assistant
- Predictive text
- Augmented reality
- Face tracking camera
- Fingerprint identity

*IDC research
Why is ML Deployed at the Edge

- Bandwidth
- Power
- Cost
- Latency
- Reliability
- Security
Significant Opportunity Ahead

- VR/MR
- Robotics
- Drones
- Shipping & Logistics
- Automotive
- IoT
- Home, Surveillance & Analytics
- Medical
- Mobile
- Servers
A suite of Arm ML IP: designed for unmatched versatility and scalability:

The recently announced products
- Machine Learning (ML) processor
- Object Detection (OD) processor
- Neural Network (NN) software libraries

Add to the existing ML capabilities of
- Cortex-A and Cortex-M CPUs
- Mali GPUs

Market growth in units (today to 2028):
- Mobile - 1.7Bn to 2.2Bn
  (source: Strategy Analytics and Arm forecast)
- Smart IP Cameras - 160M to 1.3Bn
  (source: Gartner and Arm forecast)
- AI-enabled devices - 300M to 3.2Bn
  (source: IDC WW Embedded and Intelligent Systems Forecast, 2017-2022 and Arm forecast)
Project Trillium: Arm’s ML Computing Platform

Ecosystem

AI/ML Applications, Algorithms and Frameworks
- TensorFlow
- Caffe
- Caffe2
- mxnet
- Android NNAPI

Software Products

Software Libraries Optimized for Arm Hardware
- Arm NN
  - CMSIS-NN
  - Compute Library
  - Object Detection Libraries

Hardware Products

Arm Hardware IP for AI/ML
- CPU
  - arm CORTEX-A
  - arm NEON
  - arm CORTEX-M
  - arm DynamIQ
- GPU
  - arm Mali
- ML and OD Processors
  - Machine Learning
  - Object Detection
- Partner IP
  - DSPs, FPGAs, Accelerators
Flexible, Scalable ML Solutions

- Only Arm can enable ML everywhere
- 90% of the AI-enabled units shipped today are based on Arm
  (source: IDC WW Embedded and Intelligent Systems Forecast, 2017-2022 and Arm forecast)

- Increasing performance (ops/second)
- Increasing power and cost (Silicon)

- Keyword detection
- Pattern training
- Voice and image recognition
- Object detection
- Image enhancement
- Autonomous driving
- ML and OD processors
- Mali GPUs
- Cortex-M/A CPUs
- Data center

Only Arm can enable ML everywhere
90% of the AI-enabled units shipped today are based on Arm
(source: IDC WW Embedded and Intelligent Systems Forecast, 2017-2022 and Arm forecast)
Targeting multiple markets with scalable architecture

IoT: ~20 GOPs
Mobile: 1~3 TOPs
Industrial: ~20-50 TOPs
Automotive: > 70 TOPs
Networking: 
Server: 

Scalable Machine Learning Processor “Architecture”

Scalable
Compatible
Programmable

Sensors (2 GOPs)

Phase 1

Servers 147 TOPs
Arm’s Scalable ML Processor Architecture

Architecture enables solutions from Sensors to Servers (2 GOPs to 150 TOPs)
Provide current state of the art and innovation for future industry developments
Project Trillium: Arm ML and OD Processors

- Ground-up design for high performance and efficiency
- Massive uplift from CPUs, GPUs, DSPs and accelerators
- Enabled by open-source software
- First-generation ML processor targets Mobile market
Trillions of Ops/s for Mobile

ML processor is built on a highly versatile and scalable architecture

First generation targets Mobile market for Inference at the Edge:

- Highest performance per mm$^2$ in the market
  - Typical mobile performance of >4.6 TOPs
  - Optimizations provide a further uplift of 2x to 4x in real-world use

- Unmatched performance in thermal- and cost-constrained environments
  - Efficiency of 3 TOPs/W$^1$

- First IP available to Partners mid 2018

$^1$Based on 7nm implementation
Arm ML Processor

**Network control unit**
- Overall programmability and high level control flow

**Onboard Memory**
- Central storage for weights and feature maps

**DMA**
- Move data in and out of main memory

**Fixed Function engines**
- Main fixed-function compute engines

**Programmable layer engines**
- Programmable engines for future proofing
Industry-leading Object Detection

OD processor:
- ✦ Second-generation OD processor
- ✦ Detects in real time with Full HD @ 60fps
- ✦ Object sizes from 50x60 pixels upwards
- ✦ Virtually unlimited objects per frame

Provides object detection and rich characterization:
- ✦ Direction people are facing
- ✦ Trajectory through robust inter-frame tracking
- ✦ Gesture and pose

First-generation OD processor powers the Hive security camera

UP TO 80x faster than DSP equivalent
OD plus ML Processors: Better User Experience

Combined Arm solution:
- Better user experience with high resolution, real-time face recognition
- OD processor isolates areas of interest in real time with Full HD @ 60fps
- ML processor analyzes fewer pixels for faster, fine-grain object recognition
- Leads to a new class of smart camera and other vision-based devices
Mobile Experiences: Insights From Advanced Compute

- ML and OD processors enable smartphone linking to any screen for awareness/protection (e.g. sunglasses, ski goggles, dive masks)

**Blue shark**
- Anti-shark suit electric current active
- Dive boat alerted

**Sea anemone**
- Poisonous, only touch with gloves

**Beware hole**
- Could be Moray eel hideaway

**Bigeye snapper**
- Not protected
Living: Interpreting Data for Smart City Planning

- ML and OD processors embedded in city camera systems for real-time information and control
- Detecting pedestrian impedance, congestion, safety issues (e.g. abandoned bag)
- Road-obstruction recognition, linked to GPS network to look for specific information (e.g. lost child)

By 2030, a projected 27 percent of people worldwide will be concentrated in cities with at least 1 million inhabitants. “
- United Nations
ML Support in Cortex CPUs & Mali GPUs

Cortex-A
- 10x SIMD performance improvement in two generations
- Cortex-A v8.2 instruction set with efficient FP16 and 8-bit dot product operation
- Future SVE ISA for general ML performance expansion

Cortex-M
- Optimized Compute Library and CMSIS-NN to improve ML compute
- Small area and power profile with enhanced compute capability for embedded devices

Mali GPU
- Parallel architecture with large compute processing capacity for higher ML performance
- Further improvements for ML planned
Optimum ML Performance on Arm for Any Application

- 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)
Arm NN for Android & Linux: Overview

**APK with Tensorflow Lite or Caffe 2**
- Android NNAPI
  - Runtime and Scheduler
- AndroidNN HAL
- Arm NN
  - (Inference Engine)
- Arm and Google Optimized Routines
- Cortex-A CPU
  - Compute Library
  - GPU
  - Mali GPU
  - Arm ML processor
  - NNX

**Application**
- TensorFlow or Caffe
- Others to follow
- Support for Cortex-M in development
- Support for ML Processor available on release

**Arm NN providing support for Cortex-A CPUs and Mali GPUs under embedded Linux**

**Arm NN providing support for Mali GPUs under Android NNAPI**

**Arm NN SDK**
- Runtime and Scheduler
- CMSIS-NN
- Compute Library
- Partner IP driver and SW functions
- Cortex-M CPU
- Cortex-A CPU
- Mali GPU
- ML processor
- NNX
- 3rd party IP
Compute Library

Optimized low-level functions for CPU and GPU

- Most popular CV and ML functions
- Supports common ML frameworks
- Over 80 functions in all
- Quarterly releases
- CMSIS-NN separately targets Cortex-M

Enable faster deployment of CV and ML

- Targeting CPU (NEON) and GPU (OpenCL)
- Significant performance uplift compared to OSS alternatives (up to 15x)

Publicly available now (no fee, MIT license)

https://developer.arm.com/technologies/compute-library
**CMSIS-NN**

**Optimized low-level NN functions for Cortex-M CPUs**

A collection of efficient neural network kernels developed to maximize the performance and minimize the memory footprint of neural networks on Cortex-M processor cores.

**Publicly available now** (no fee, Apache 2.0 license)

---

**Table:**

<table>
<thead>
<tr>
<th></th>
<th>Relative throughput</th>
<th>Energy efficiency improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv</td>
<td>4.6x</td>
<td>4.9x</td>
</tr>
<tr>
<td>Pooling</td>
<td>4.6x</td>
<td>4.9x</td>
</tr>
<tr>
<td>ReLU</td>
<td>4.6x</td>
<td>4.9x</td>
</tr>
<tr>
<td>Total</td>
<td>4.6x</td>
<td>4.9x</td>
</tr>
</tbody>
</table>

**Charts:**

- NN Runtime improvement: 4.6x higher perf.
- Energy efficiency improvement: 4.9x higher eff.
ML Developer Community

Summary of Arm & ecosystem software available for ML

Several how-to guides for ML use cases

White papers & research papers from ML team

Explore ecosystem partners & get closer to deploying ML solutions!

http://developer.arm.com/mlcommunity
Software Resources

Arm ML Developer Resources: http://developer.arm.com/mlcommunity

Arm Software Repositories: https://github.com/ARM-software

ArmNN: https://github.com/ARM-software/armnn

Arm Compute Library: https://github.com/ARM-software/ComputeLibrary

CMSIS-NN: https://github.com/ARM-software/CMSIS_5
Project Trillium: Arm ML and OD Processors

- Ground-up design for high performance and efficiency
- Massive uplift from CPUs, GPUs, DSPs and accelerators
- Enabled by open-source software
- First-generation ML processor targets Mobile market
Project Trillium Summary: Unleashing Innovation for ML on Arm

- ML processor delivers performance of >4.6 TOPs with efficiency of 3 TOPs/W
- OD processor provides object detection and rich characterization in real time with Full HD @ 60fps
- Full suite of Arm NN software supports leading NN frameworks
- Targets mobile and smart camera markets first and scaling to all devices
Thank you!

Vrajesh Bhavsar
Senior Ecosystem Manager, ML
vrajesh.bhavsar@arm.com