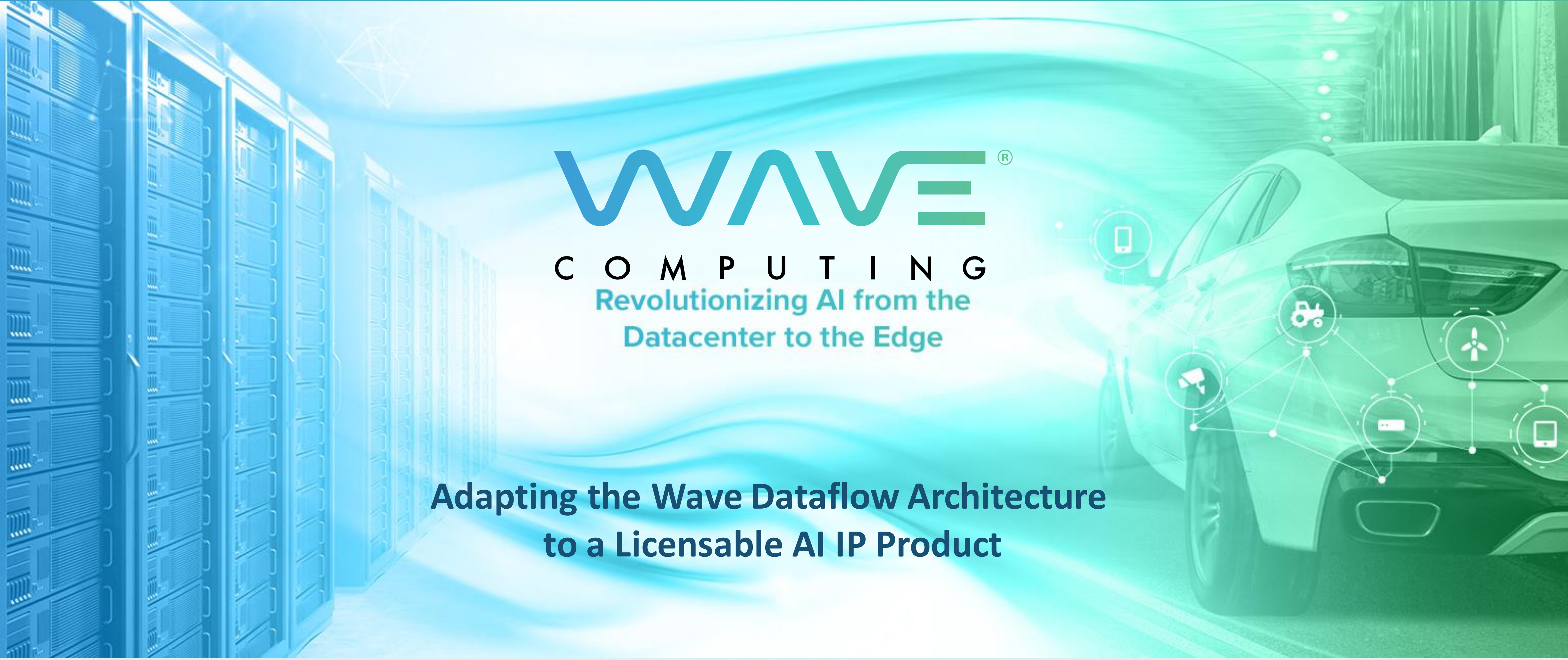


WAVE® COMPUTING

Revolutionizing AI from the Datacenter to the Edge

Adapting the Wave Dataflow Architecture to a Licensable AI IP Product

Presented by **Yuri Panchul**, MIPS Open Technical Lead

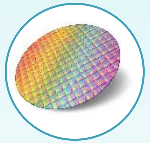
On SKOLKOVO Robotics & AI Conference. April 15-16, 2019

www.wavecomp.ai

WAVE COMPUTING

Wave + MIPS: A Powerful History of Innovation

2010 2011 2012 2013 2014 2015 2016 2017 2018 2019



2010

Wave founded by Dado Banatao as Wave Semiconductor, with a vision of ushering in a new era of AI computing

2011

Wave develops dataflow-based technology, providing higher performance and scalability for AI applications

MIPS introduces first Android-MIPS based Set top box at CES

Google selects MIPS architecture for use in its Android 3.0, "Honeycomb" mobile device

2012

Developed Coarse Grain Reconfigurable Array (CGRA) semiconductor architecture

2013

Wave expanded team to include architecture, silicon and software expertise

MIPS Technologies is sold to Imagination Technologies.

2014

Delivered 11GHz test chip at 28nm

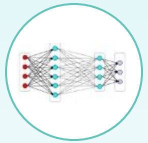
2015

Renamed the company to Wave Computing to better reflect focus on accelerating AI with dataflow-based solutions

MIPS Automotive MCUs & ADAS engagements with MobilEye, Microchip & MStar

MICROCHIP

MOBILEYE



2016

Announced Derek Meyer as CEO

Launched Early Access Program to enable data scientists to experiment with neural networks

2017

MIPS business is sold by Imagination Technologies to Tallwood Venture Capital as Tallwood MIPS Inc. for $65M

MediaTek selects MIPS for LTE modems

MEDIATEK

TALLWOOD Venture Capital

2018

Wave acquires MIPS to deliver on its vision for of revolutionizing AI from the datacenter to the edge

Announced partnership with Broadcom and Samsung to develop next-gen AI chip on 7nm node

Launched MIPS Open initiative

Closed Series D Round of funding at $86M, bringing total investment to $115M+

Expanded global footprint with offices in Beijing and Manila

2019

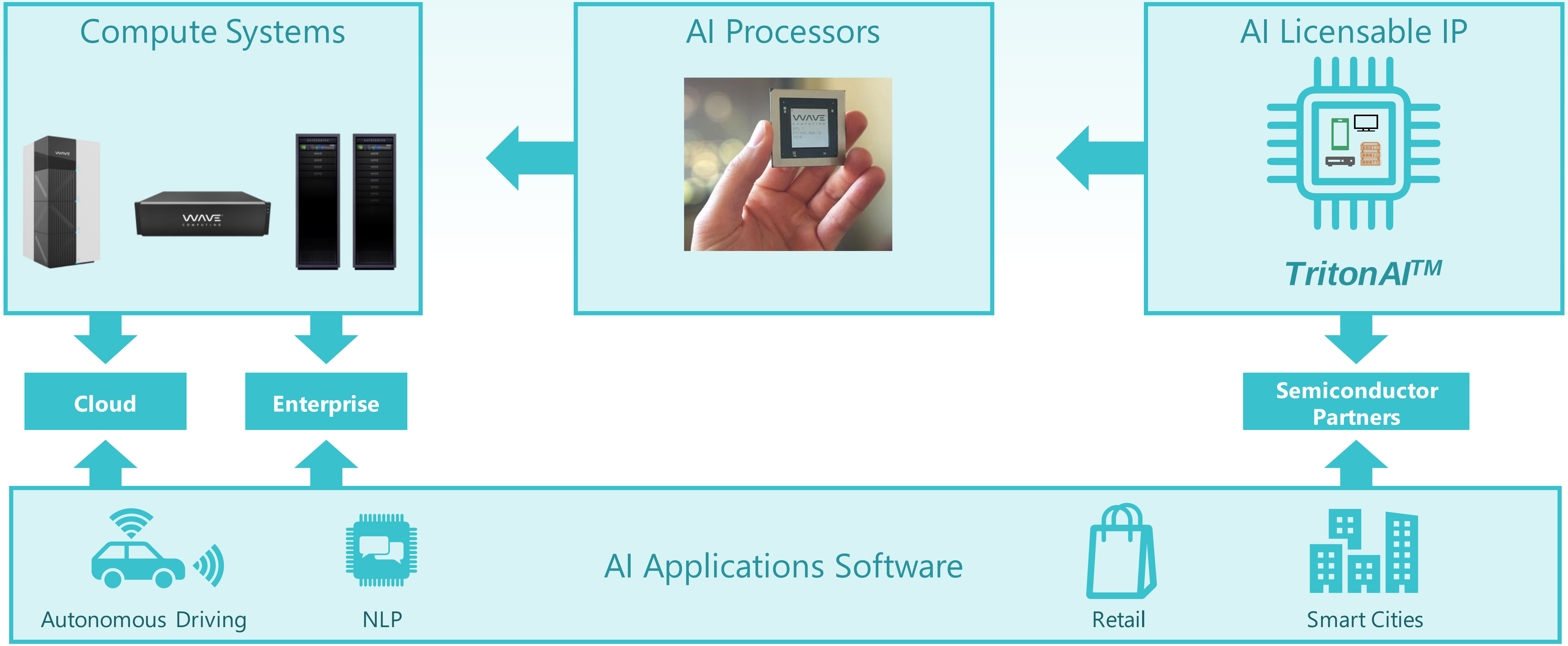
Created MIPS Open Advisory Board

MIPS Open

MIPS Powering 80% of ADAS-enabled automobiles

Wave Joins Amazon, Facebook, Google and Samsung to Support Advanced AI Research at UC Berkeley

Datacenter-to-Edge Needs Systems, Silicon and IP



Compute Systems

AI Processors

AI Licensable IP

TritonAI™

Cloud

Enterprise

Semiconductor Partners

AI Applications Software

Autonomous Driving

NLP

Retail

Smart Cities

WAVE® COMPUTING

What is Driving AI to the Edge?

AI was born in Datacenter

Revolutionizing AI from the Datacenter to the Edge

Market Drivers

Networking
Enterprise
Mobile
Industrial
Autonomous
IOT

AI Use Cases

Privacy
Security
Isolated
Low latency

Cost

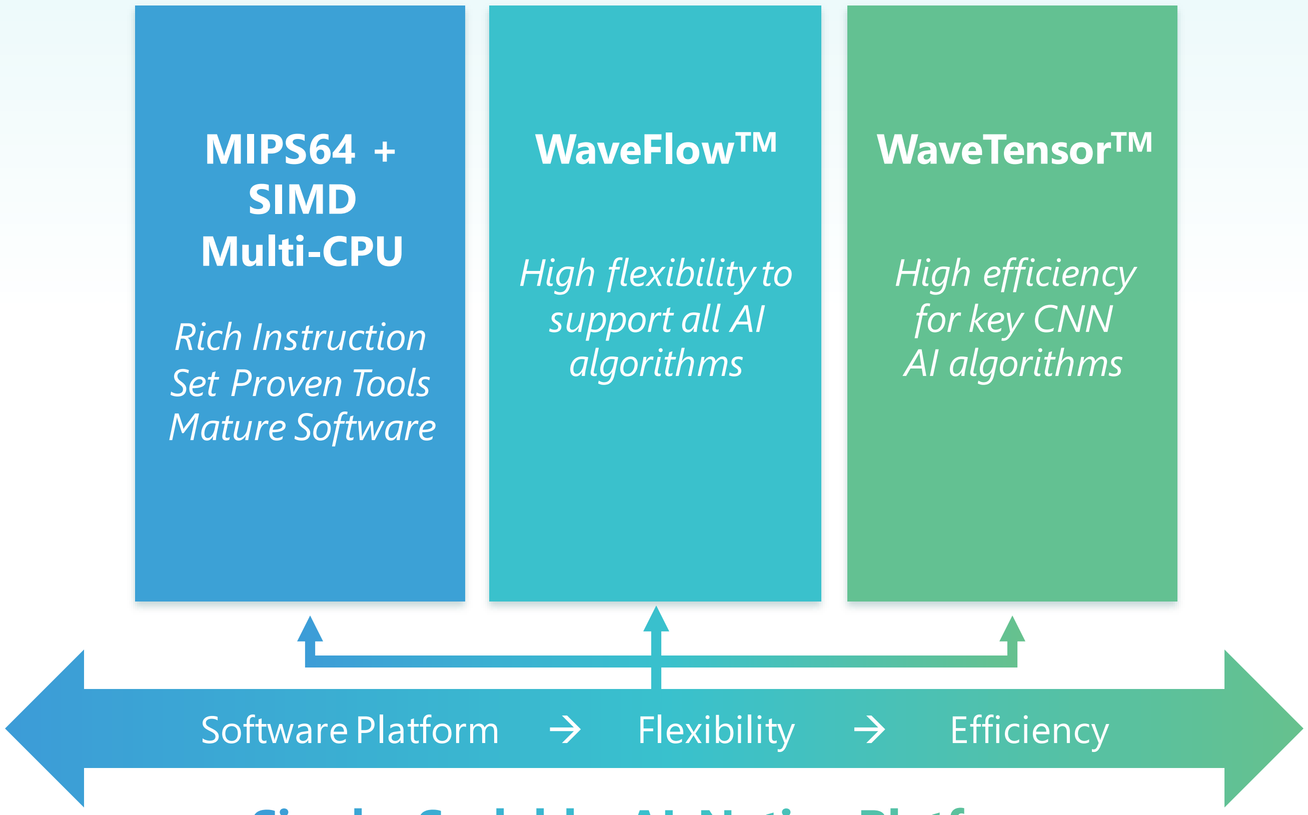
Bandwidth
Storage
Compute

Scalable AI Platform: Triton*AI*™ 64

Key Benefits:

- Highly Scalable to address broad AI use cases
- Supports **Inference** and **Training**
- High flexibility to support all AI algorithms
- High efficiency for key AI CNN algorithms
- Configurable to support AI use cases
- Mature Software Platform support

Triad of Technologies



MIPS64 + SIMD Multi-CPU

Rich Instruction Set Proven Tools Mature Software

WaveFlow™

High flexibility to support all AI algorithms

WaveTensor™

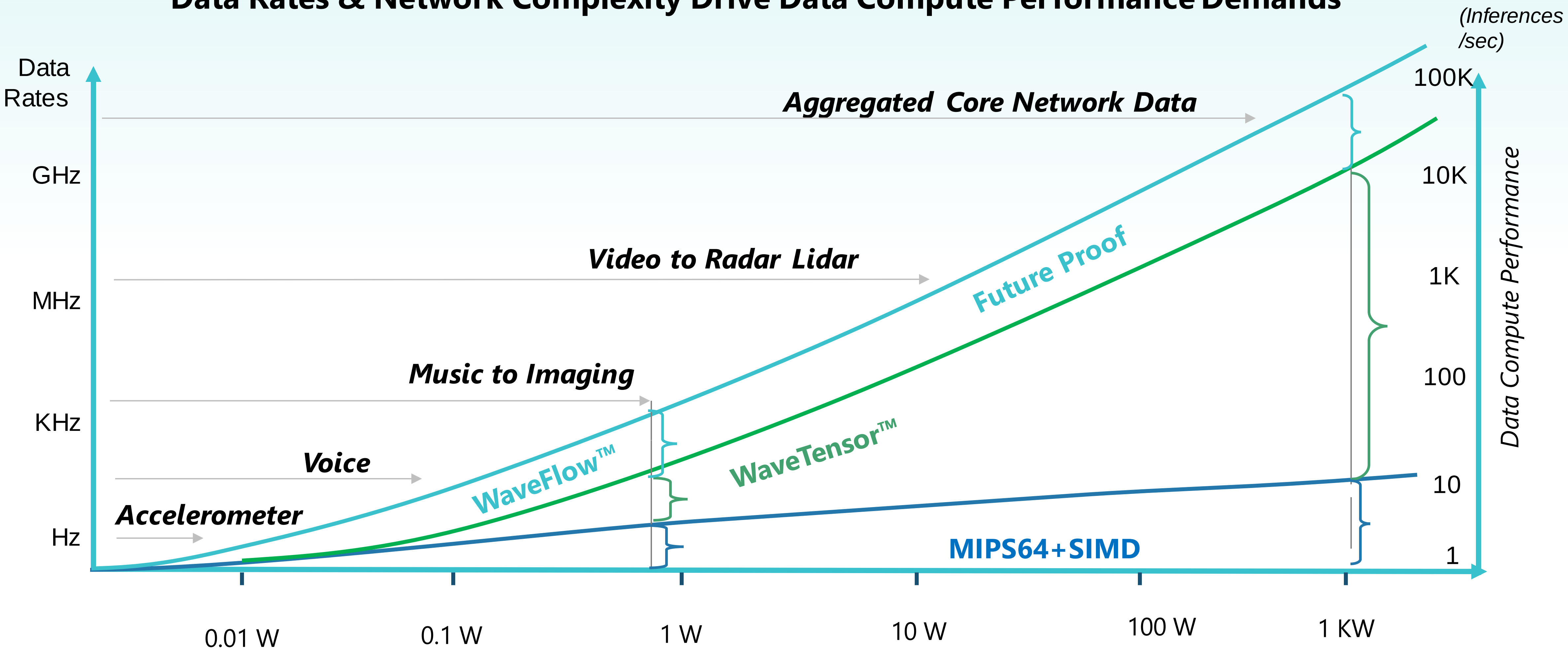
High efficiency for key CNN AI algorithms

Software Platform → Flexibility → Efficiency

Single, Scalable, AI-Native Platform

Data Rates and New AI Networks Requirements

Data Rates & Network Complexity Drive Data Compute Performance Demands



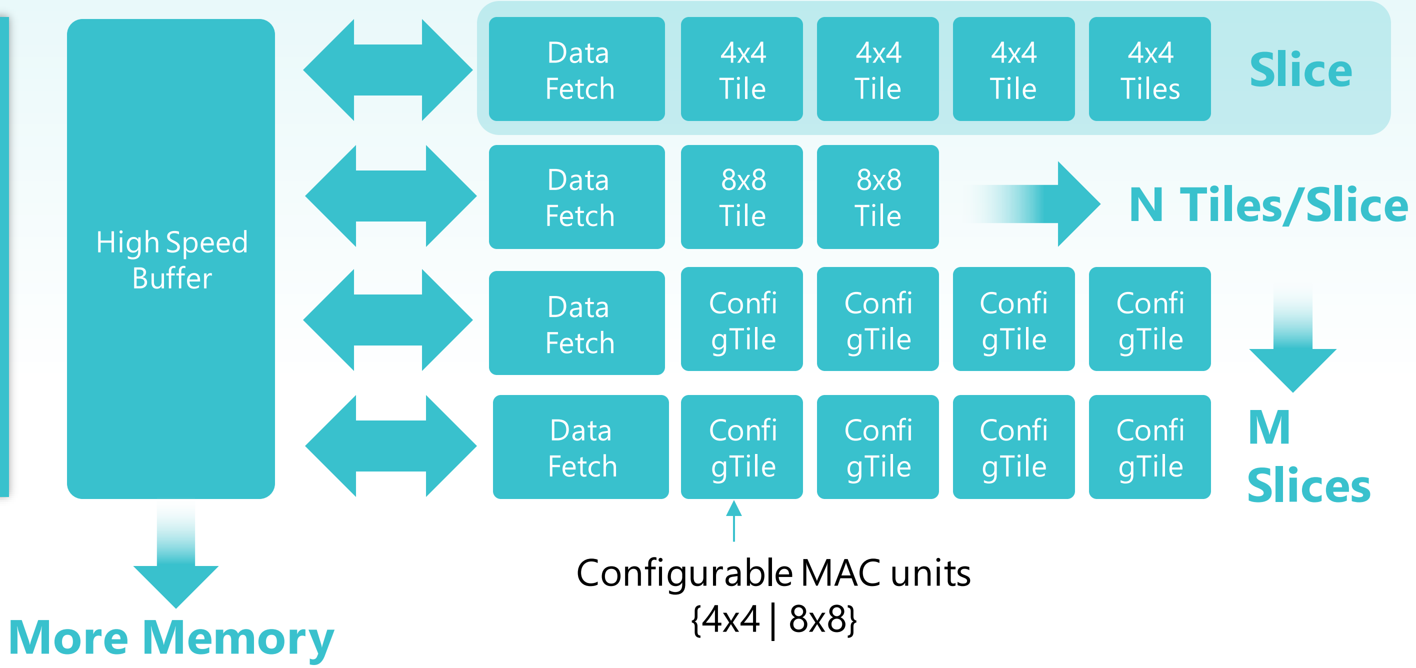
These curves represent the conceptional combination of these technologies, not actual independent performance.

WAVE COMPUTING

WaveTensor ™ Configurable Architecture

Configurable Architecture for Tensor Processing

- Configurable MACs, Accumulation and Array Size
- Overlap of Communication & Computation
- Compatible datatypes with WaveFlow™ Core
- Supports int8 for inferencing
- Roadmap to bfloat16, fp32 for training



ResNet-50	Inferences/Sec*
Compute Density	~1K/mm2
Compute Efficiency	~500/watt

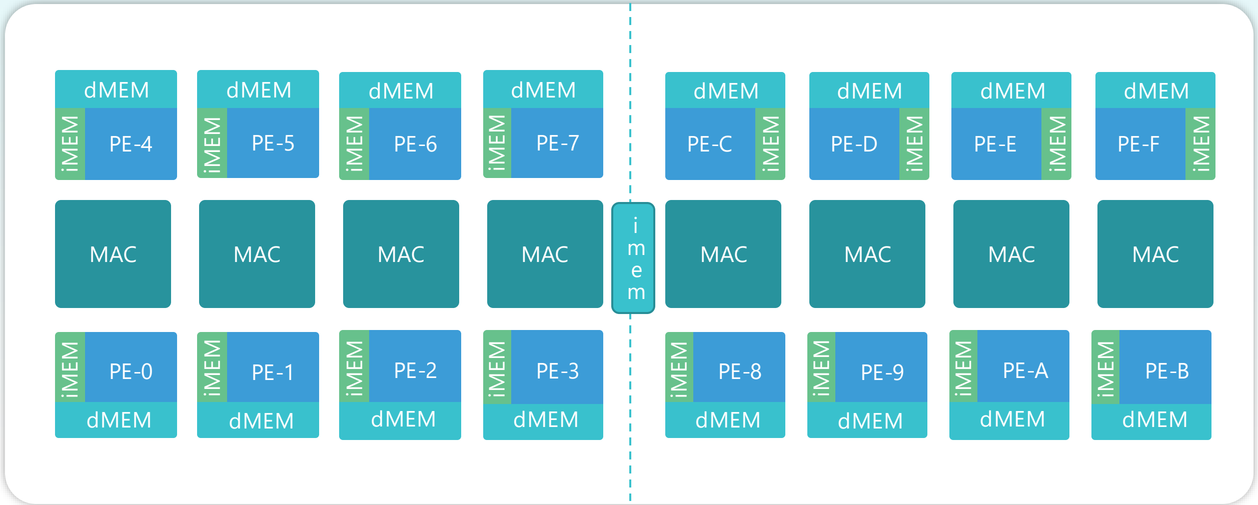
- Core, Int8, 7nm Fin-FET nominal process/Vdd/temp
- Batch=1, std model w/o pruning, performance and power vary with array size/configuration

MAX TOPs	TOPs/Watt	TOPs/mm^2
1024	8.3	10.1

- Core, Int8, 8x8 tile config, 7nm Fin-FET nominal process/Vdd/temp

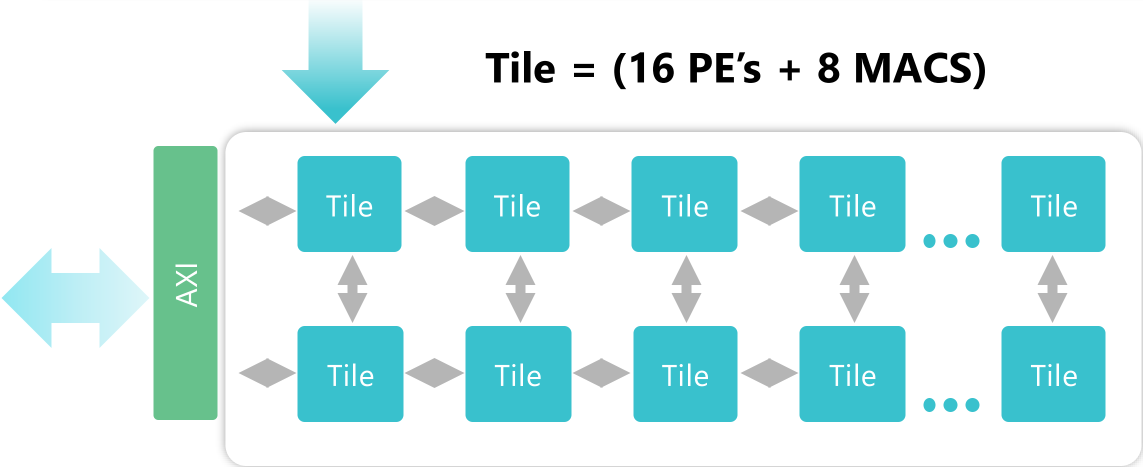
WaveFlow™ Reconfigurable Architecture

- Configurable IMEM and DMEM Sizes
- Overlap of communication & Computation
- Compatible datatypes with WaveTensor™
- Integer (Int8, Int16, Int32) for inference
- Roadmap (bfloat16, fp32) for training



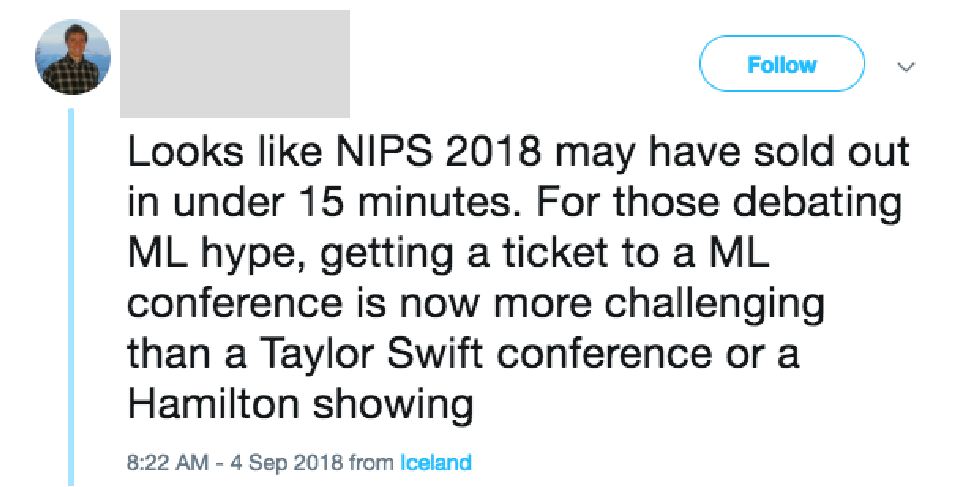
Tile = (16 PE's + 8 MACS)

- Wide range of scalable solutions (2-1K tiles)
- Future Proof all AI algorithms
- Flexible 2 dimensional tiling implementation
- Reconfigurable for dynamic networks
- Concurrent Network execution
- Supports signal and vision processing



WaveFlow™ = Wave Dataflow Array of Tiles

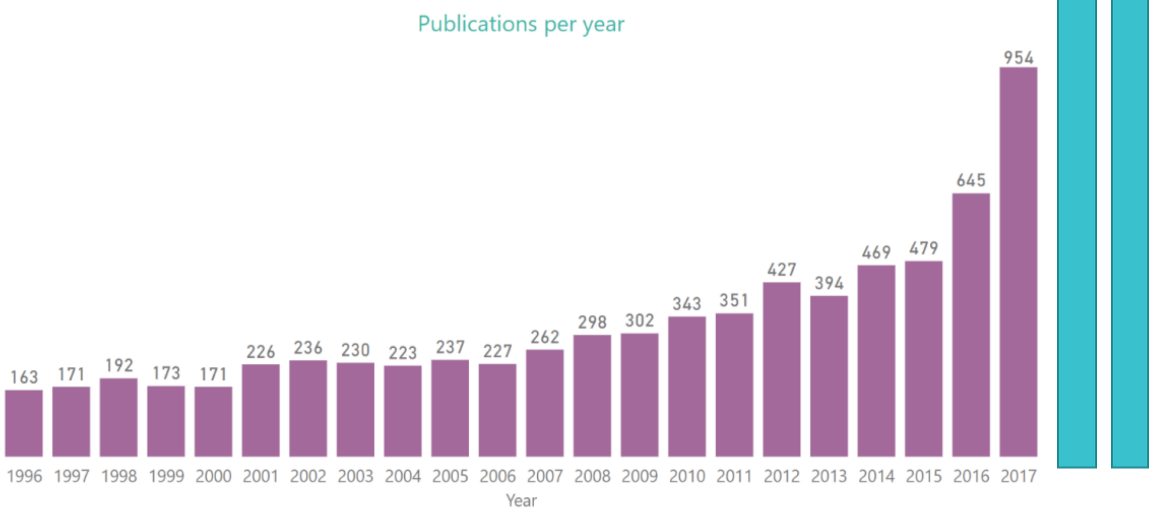
Future Proofing in an Rapidly Evolving Industry



[redacted]

Looks like NIPS 2018 may have sold out in under 15 minutes. For those debating ML hype, getting a ticket to a ML conference is now more challenging than a Taylor Swift conference or a Hamilton showing

8:22 AM - 4 Sep 2018 from Iceland



What is the likelihood that your DNN accelerator will run all these "yet to be invented" networks?

Wave's TritonAI™ 64 platform combines a reconfigurable processor with an efficient neural network accelerator.

Offers customers peace of mind and investment protection

Operations using WaveFlow™ Reconfigurable Processor

Future-proof your Silicon

CNN Layers

- Sparse Matrix-Vector Processing
- Stochastic pooling
- Median pooling (illumination estimation & color correction)

Activation functions

- Leaky rectified linear unit (Leaky ReLU) (used in Yolo3)
- Parametric rectified linear unit (PReLU)
- Randomized leaky rectified linear unit (RReLU)

Custom Operators (e.g.)

- Novel Loss Function
- New Softmax Implementation
- Image resize nearest neighbor

Data Preprocessing

- Scaling
- Aspect Ratio adjustment
- Normalizing

Other Functions

- Compression/Decompression
- Encryption/Decryption
- Sorting

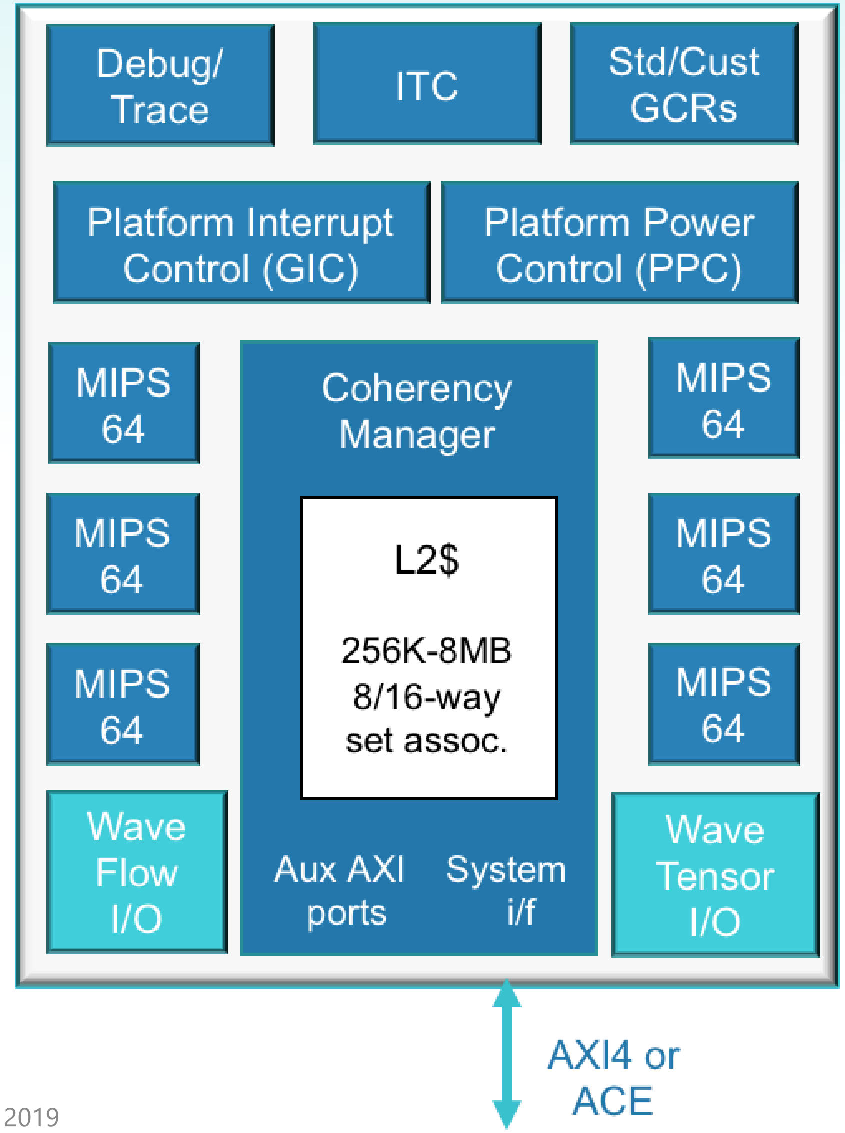
MIPS-64 Processor

MIPS-64:

- MIPS64r6 ISA
  - 128-bit SIMD/FPU for int/SP/DP ops
  - Virtualization extensions
- Superscalar 9-stage pipeline w/SMT
- Caches (32KB-64KB), DSPRAM (0-64KB)
- Advanced branch predict and MMU

Multi-Processor Cluster:

- 1-6 cores
- Integrated L2 cache (0-8MB, opt ECC)
- Power mgmt. (F/V gating, per CPU)
- Interrupt control with virtualization
- 256b native AXI4 or ACE interface



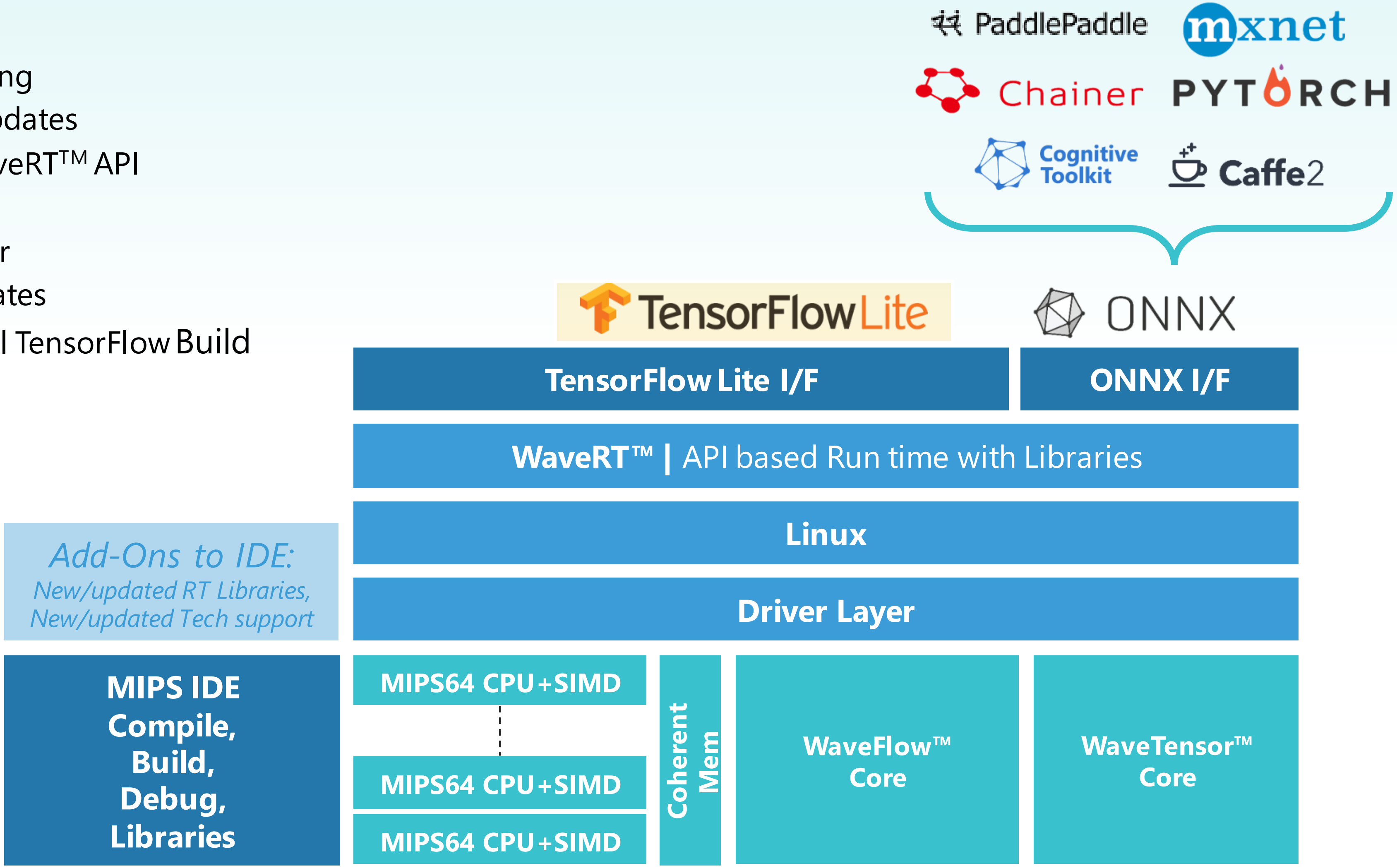
Wave's TritonAI™ 64 IP Software Platform

Software Platform:

- Mature IDE & Tools
- Driver Layer for Technology Mapping
- Linux Operating system support/updates
- Abstract AI Framework calls via WaveRT™ API
- Optimized AI Libraries for:
- CPU/SIMD/WaveFlow/WaveTensor
- TensorFlow-Lite Build support/updates
- Extensible to Edge Training with Full TensorFlow Build

Configurable Hardware Platform:

- MIPS64r6 ISA Cluster
  - 1- 6 cores
  - 1-4 threads/core
  - L1 I/D (32KB-64KB)
  - Unified L2 (256K to 8 Mbytes)
- WaveFlow Tile Array
  - 4 – N Tiles
- WaveTensor Slice Array
  - 1 – N Slices



Federated Learning:

The Next Frontier in Edge AI

Today's Approach: Centralized Machine Learning

Better ML comes at a cost of collecting data
Most training done in the cloud. i.e. Send your data to the cloud.

Diminished Privacy

- Where is your data?
- Who has access to your data?

Incompatible with Banks, Insurance, Military, Health sectors

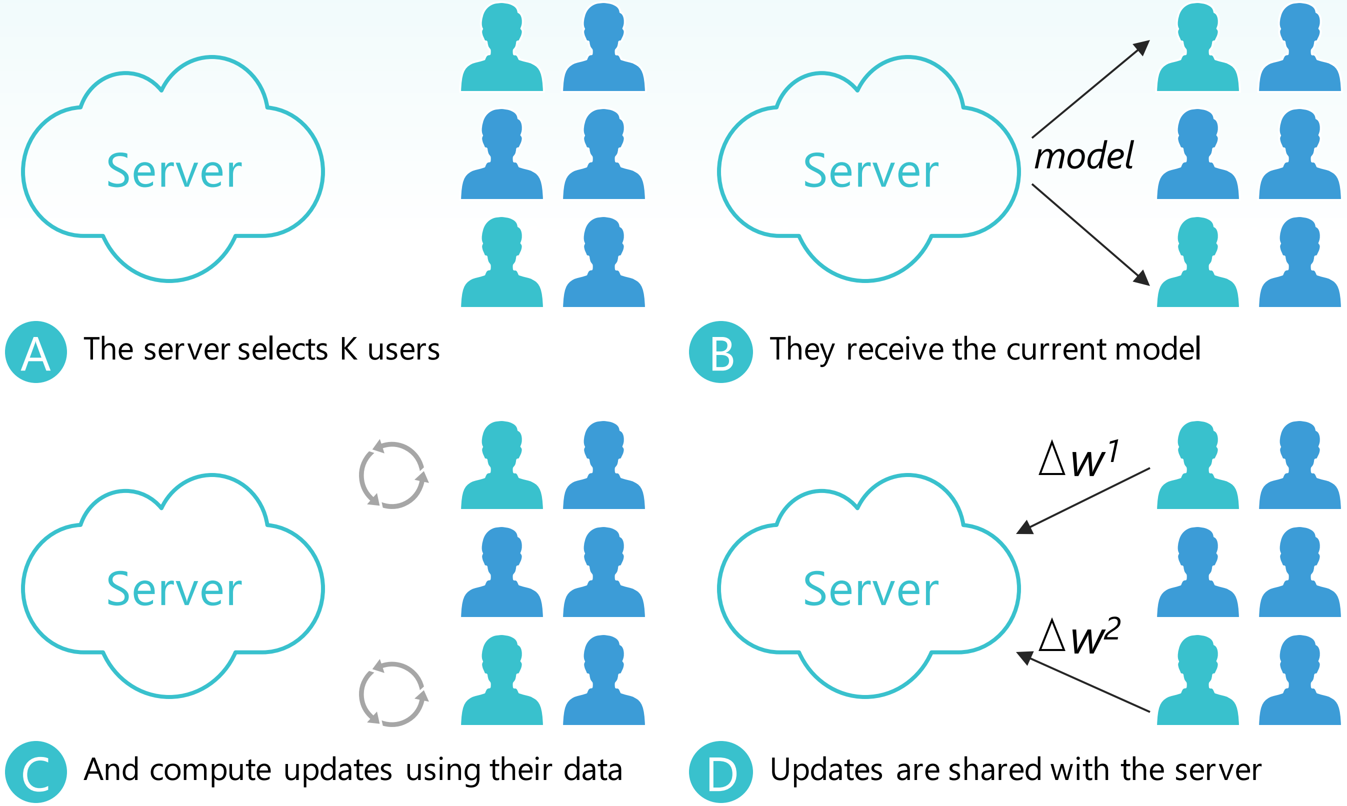
Latency Problems

- Most access technologies are asymmetric

High Communications Costs

Federated Learning: Edge AI with Data Privacy

Federated learning uses training at the edge to refine the global model



A The server selects K users

B They receive the current model

C And compute updates using their data

D Updates are shared with the server

1. Server selects a group of users
2. Users receive copy of central model
3. Users update model based on local data (“Training at the Edge”)
4. Updates are shared with the server

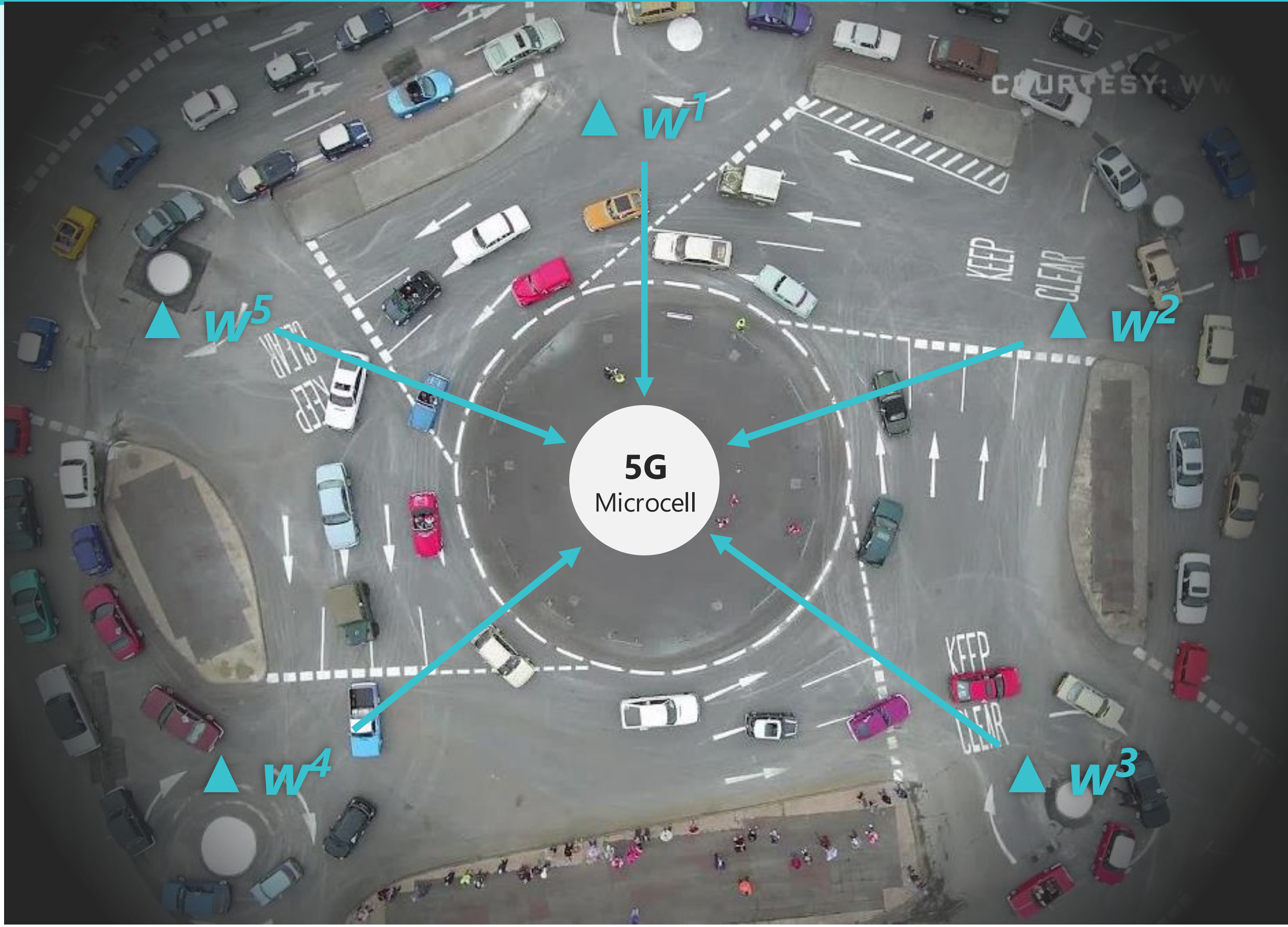
   (User data remains private)
5. Server aggregates the changes and updates the central model

Source: Florian Hartmann, Google AI

Federated Learning: Edge AI with Data Privacy

Benefits & Use Cases:

- Transfer learning using local data at edge
- Edge data remains private
- Social networking applications
- Intelligent transportation systems that help increase passenger & pedestrian safety + traffic flow

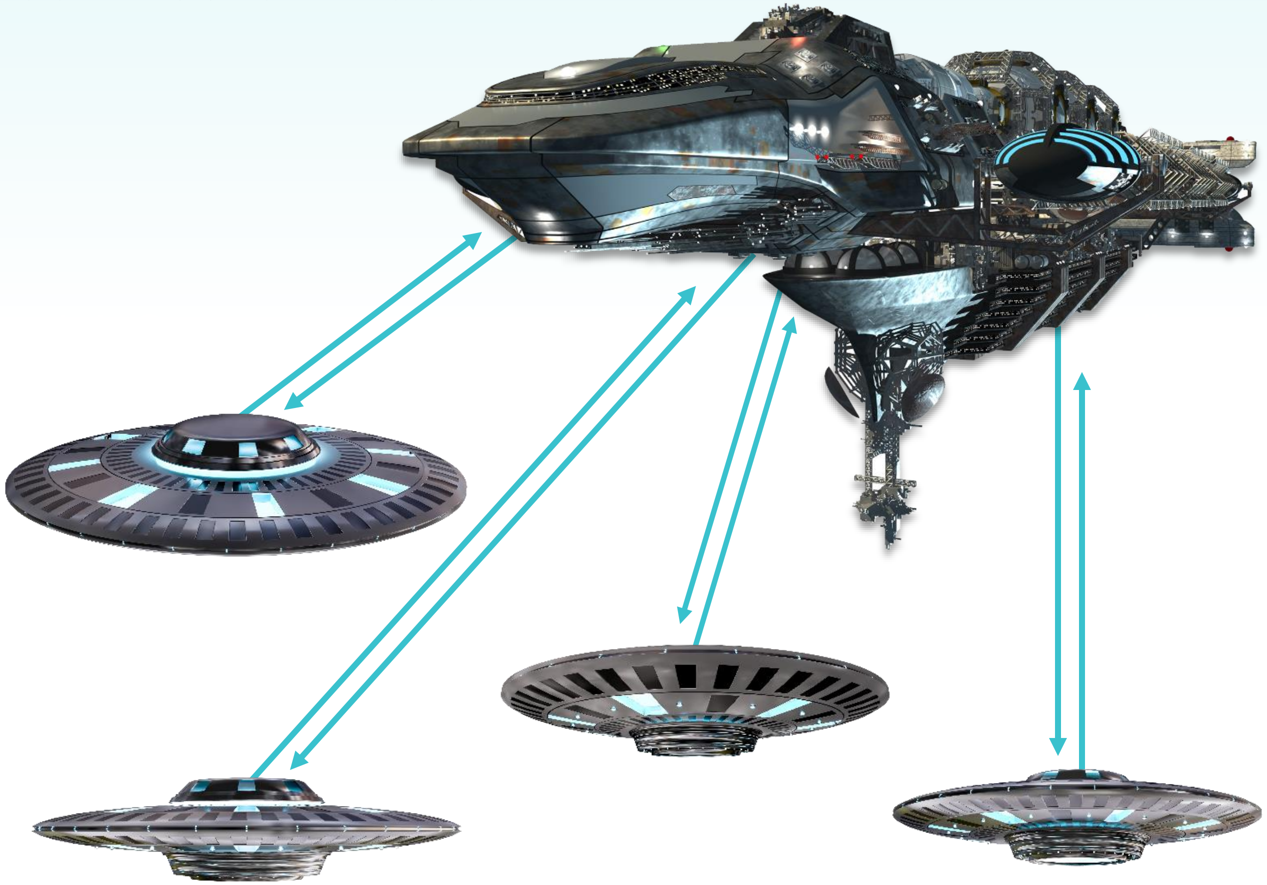


Federated Learning: Edge AI with Data Privacy

Federated learning uses training at the edge to refine a global master model

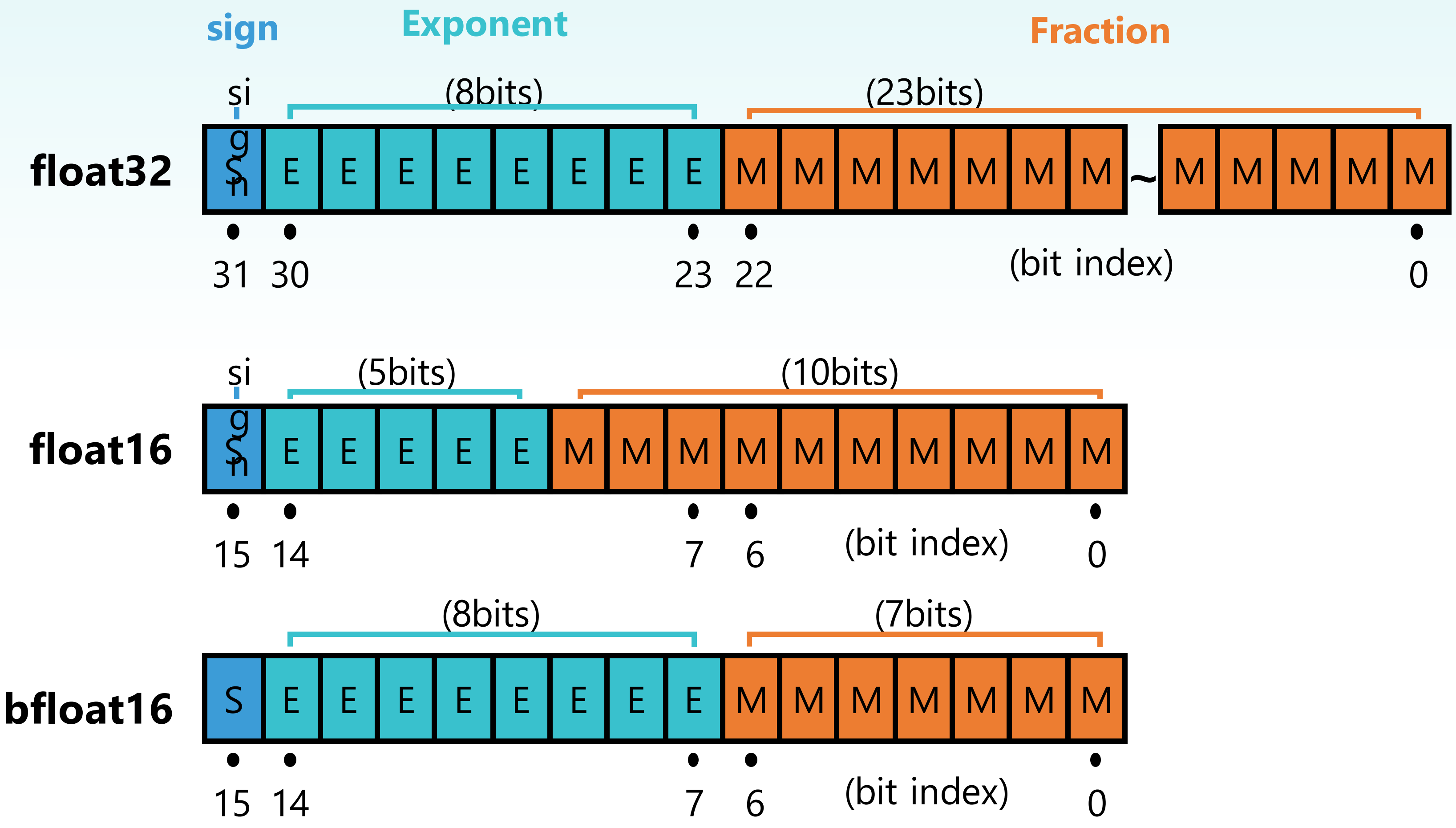
Benefits & Use Cases:

- Transfer learning using local data at edge
- Edge data remains private
- Social networking applications
- Intelligent transportation systems that help increase passenger & pedestrian safety + traffic flow



Wave is Driving Training at the Edge with bfloat16

Training into Two Camps:
Float16 or bfloat16 datatypes



sign Exponent Fraction

float32

si (8bits) (23bits)

sign E E E E E E E E M M M M M M M ~ M M M M M

31 30 23 22 (bit index) 0

float16

si (5bits) (10bits)

sign E E E E E M M M M M M M M M M

15 14 7 6 (bit index) 0

bfloat16

(8bits) (7bits)

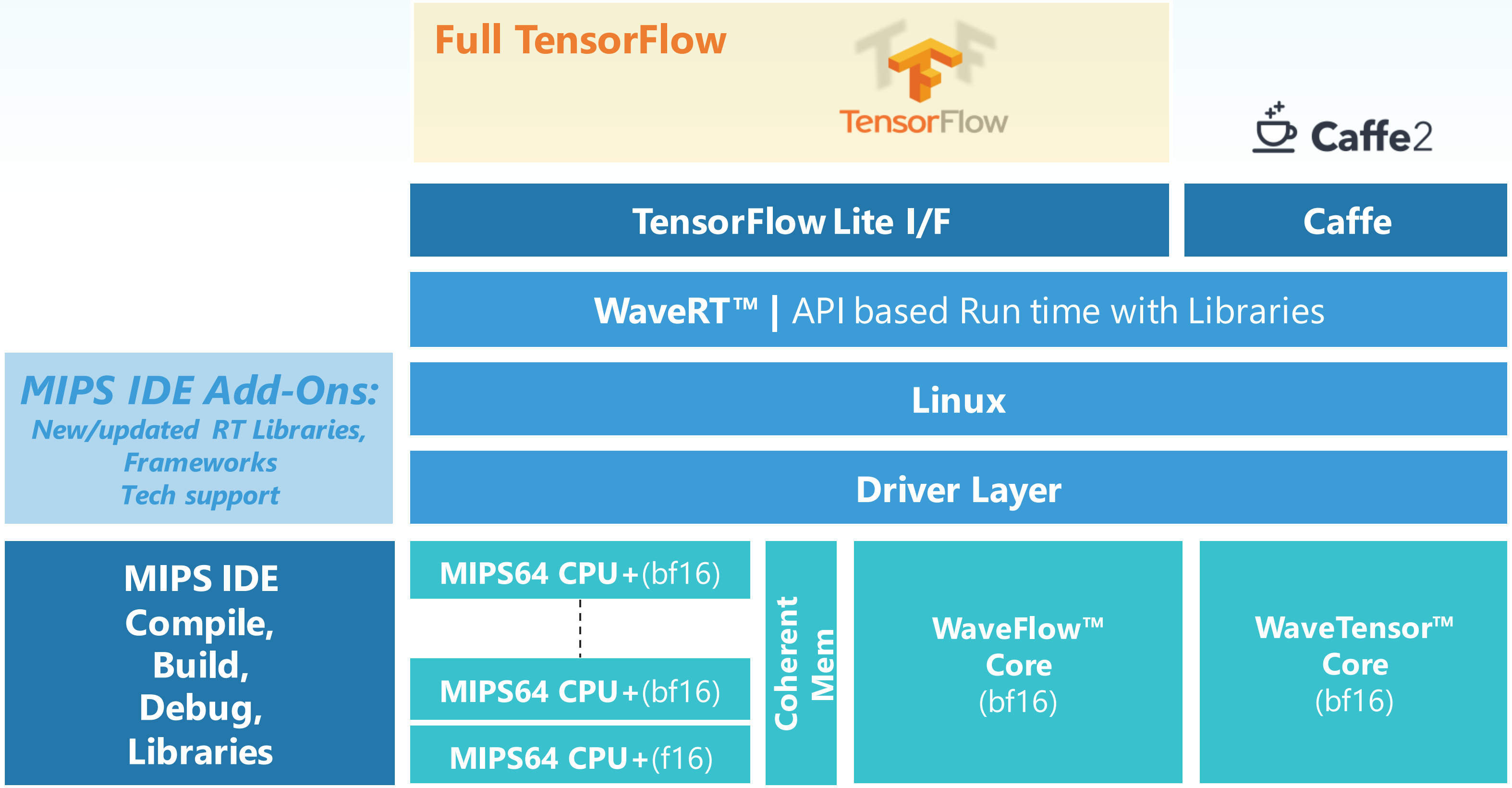
S E E E E E E E E M M M M M M M

15 14 7 6 (bit index) 0

The **TritonAI™** Edge Training **(Development)**

Edge Training Development

- Training Stacks for
  - Federated Learning at the Edge
  - Transfer Learning at the edge
  - Local or personalized models
- Full TensorFlow Build
  - WaveRT API Ext for Training
  - Optimized SIMD FP32 & bfloat16 eigen libraries
  - Deploy training at the edge



Summary

Wave's TritonAI™ Platform Drives Inferencing to the Edge

Wave's TritonAI™ Platform is a configurable, scalable & programmable offering customers' efficiency, flexibility and AI investment protection

Wave will enable "Training at the edge" with next-gen MIPS AI processor bfloat16 architectures

Thank You

If you have questions or would like more information,
visit www.wavecomp.ai

@wavecomputing

https://www.linkedin.com/company/wave-computing

https://www.facebook.com/WaveComp/