arm

Autonomous Driving and Smart Transportation

Dr. Peter Hsieh

Arm Ltd

June 6, 2021

Table of Contents

Edge Computing: Convergence of 5G, AI, and Applications

Key Observations

- Autonomous driving is an edge application. Autonomous driving requires 100-1000 TOPS at the edge. High performance, low power, edge AI is the industry barrier.

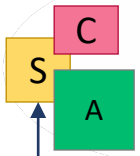
- 5G will connect edge devices and the cloud. With SDN & NFV, Network Slicing will provide dedicated & secured networks for targeted applications to meet throughput and latency requirements.

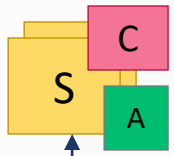
- AI computing needs domain optimized algorithms and data flow architecture. Moore's law is ending. Driving performance based only on processing technologies without the right algorithms and architectures will not achieve the desired results.

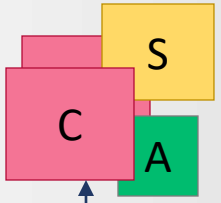
- IoT markets are highly diversified and fragmented. Dedicated AI models, data, and SoC for dedicated IoT application. Application platform supporting software defined functionalities is the right approach for the IoT markets.

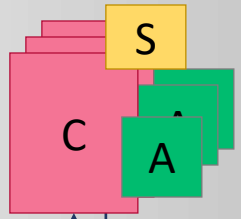
Intelligent Edge & Cloud

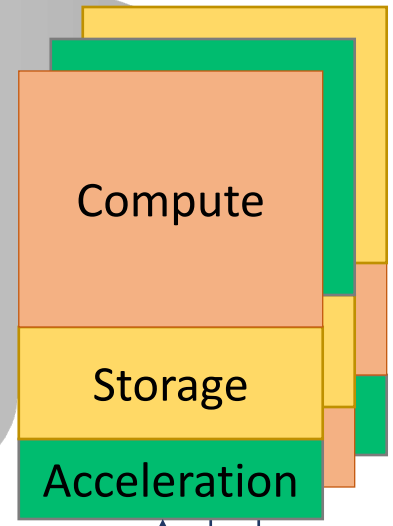
- Applications run where the data is, independent of the network node
- Heterogeneous Compute is distributed into the network
- Networks and Compute resources are both managed and configured using standard IT technologies



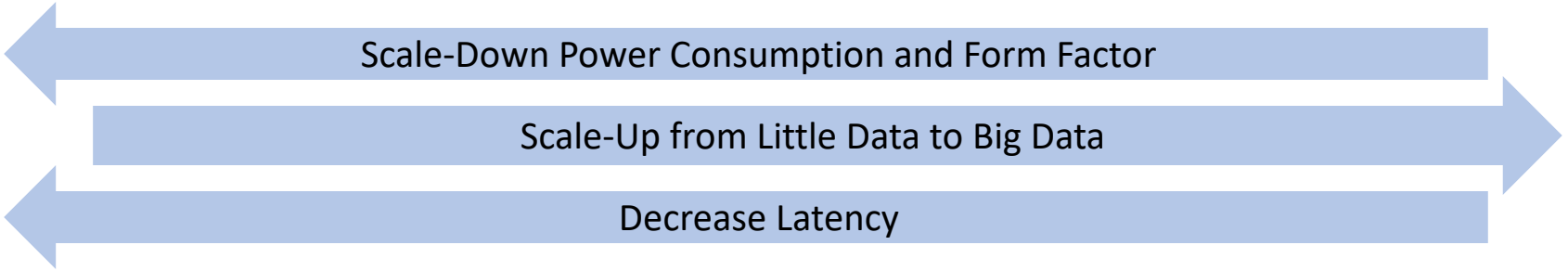




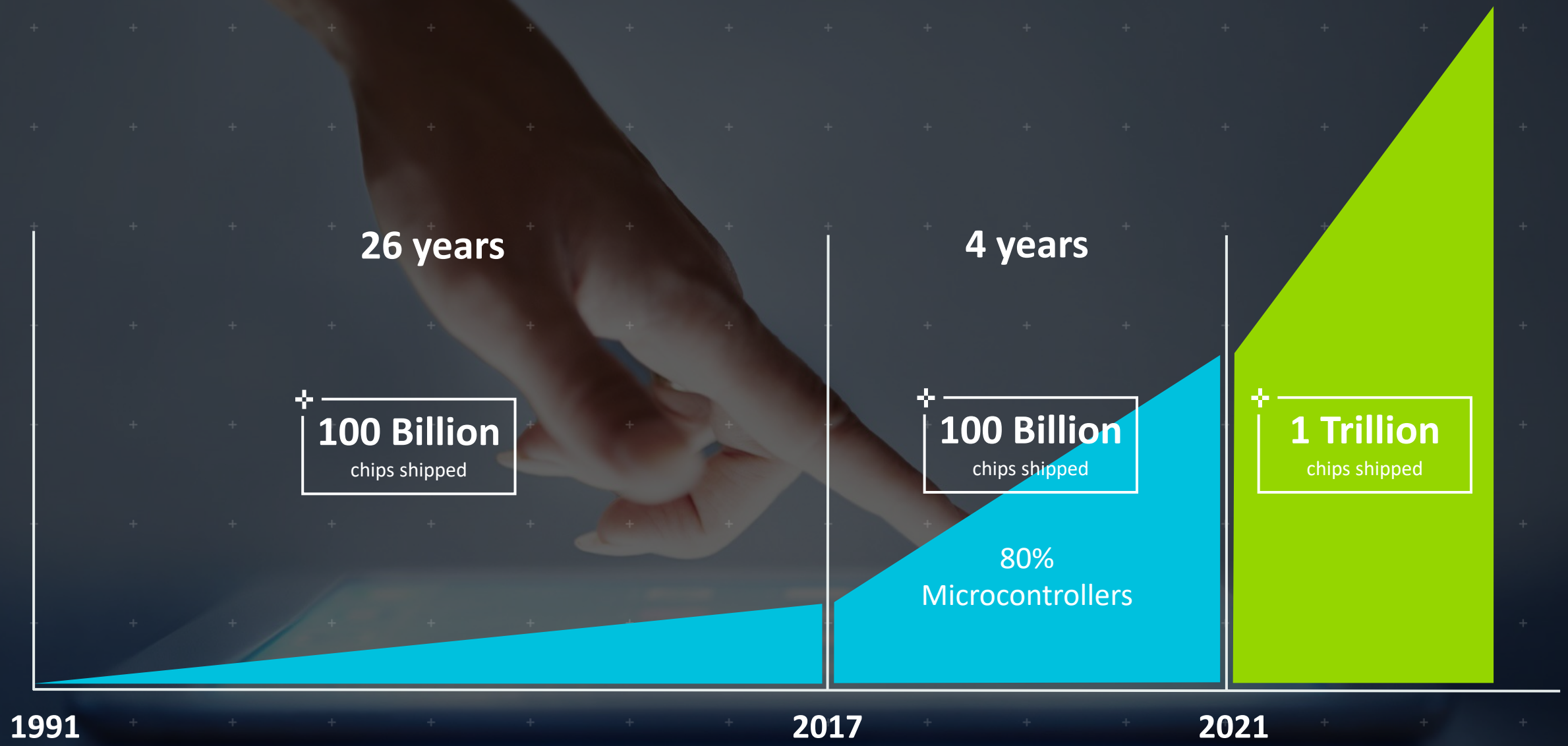




Devices Access Edge Aggregation Core Data Center



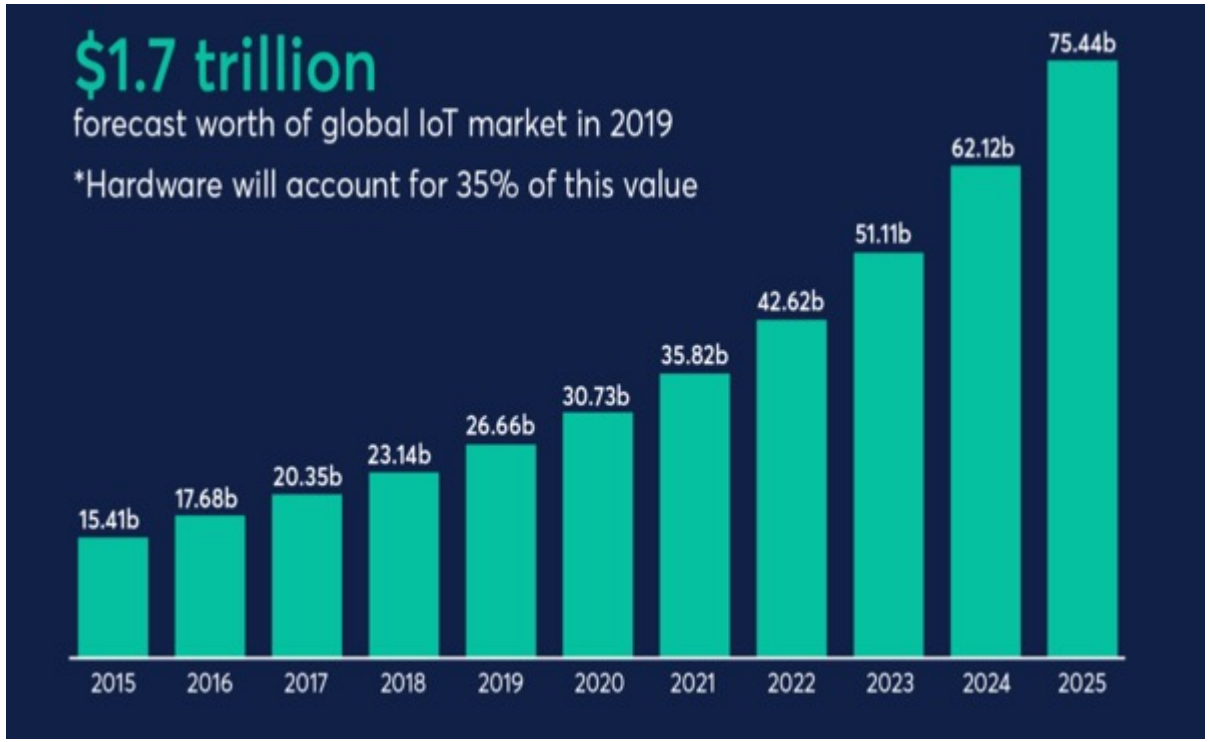
A world of 1 trillion connected devices



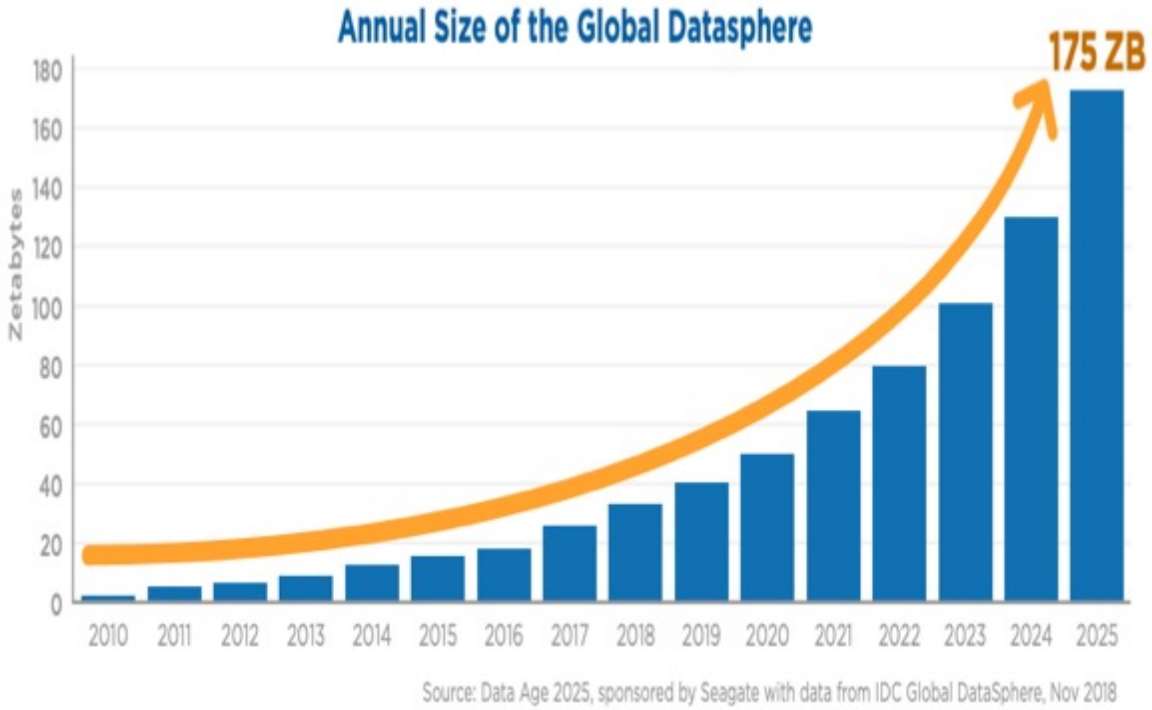
arm

Huge amount of data will be generated from IoT devices.

The number of connected(IoT) device expects to triple its current level by 2025



IDC predicts that the Global Data will grow from 33 Zettabytes (ZB) in 2018 to 175 ZB by 2025



Source: IDC, ipropertymanagement

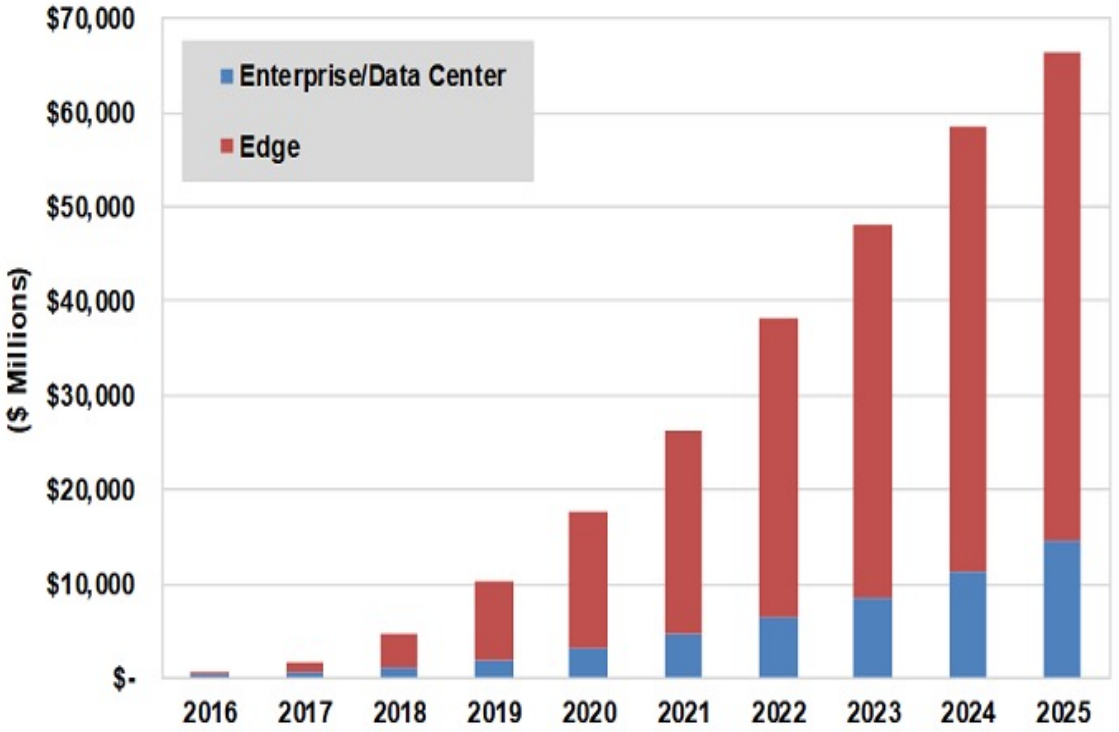
Back to the future: edge dominates future computing

“the end of cloud computing”

- A16Z

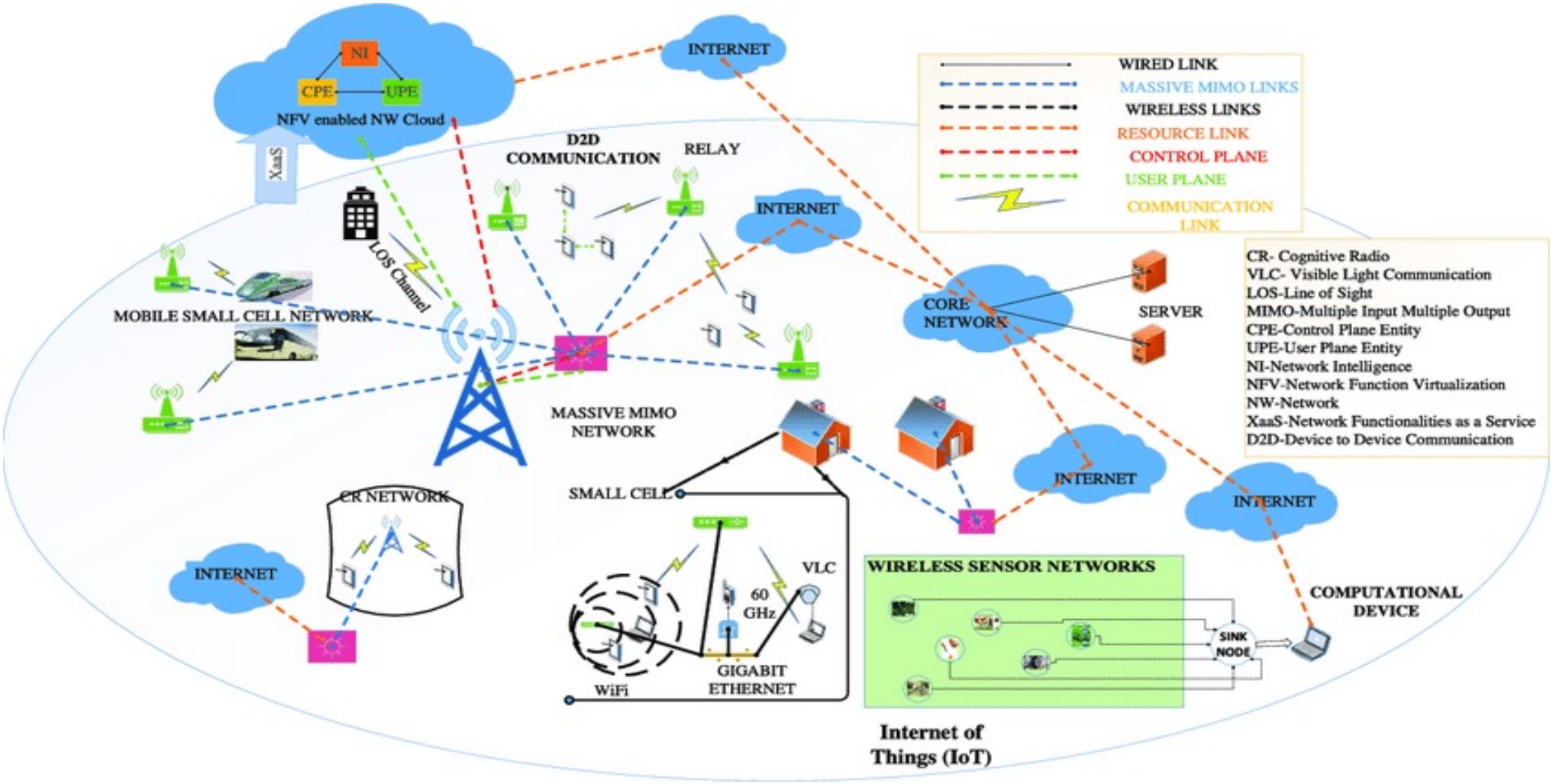
Edge computing becoming the mainstream again

The market for edge computing is surging.

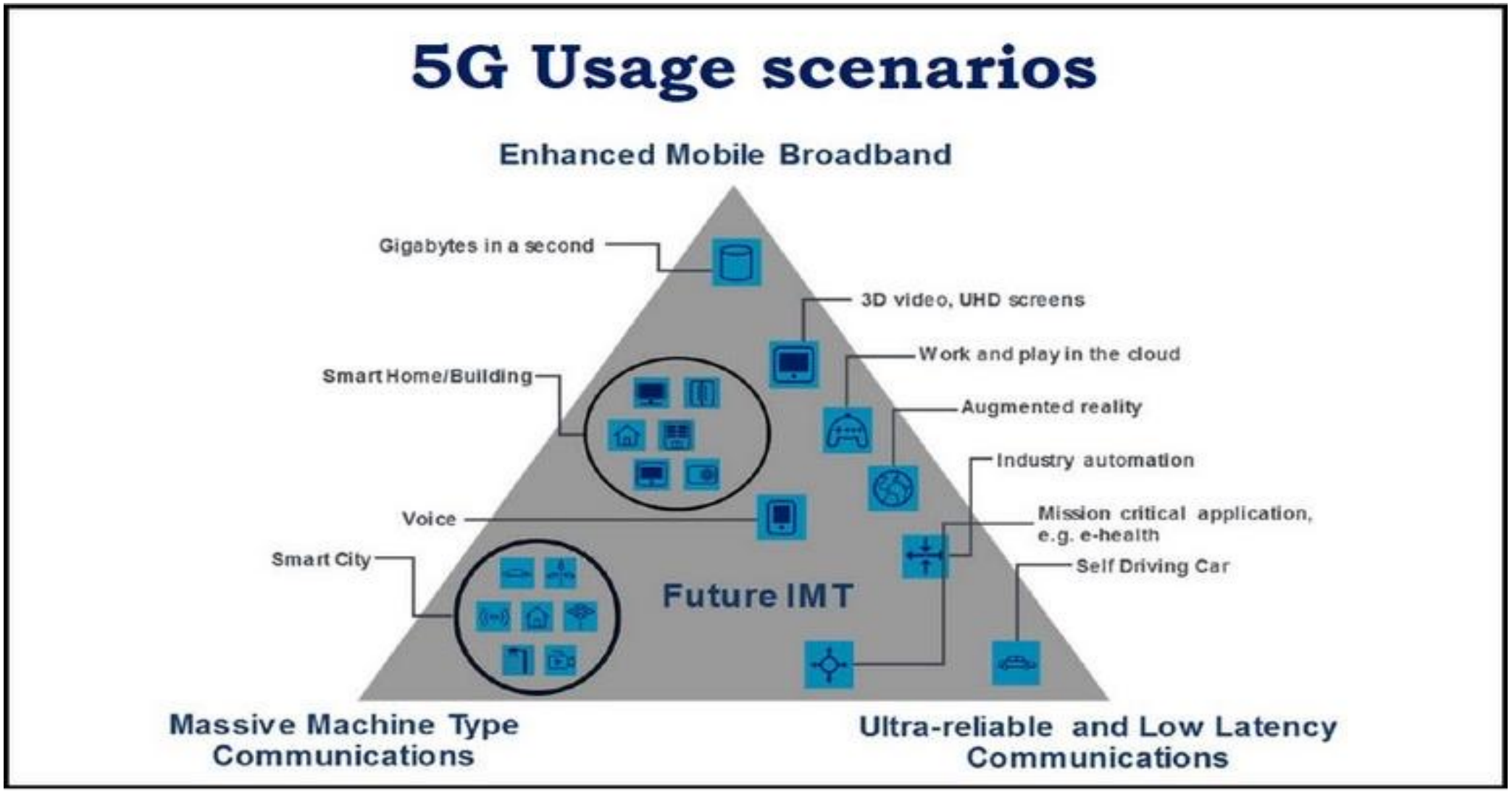


*Source: IDC

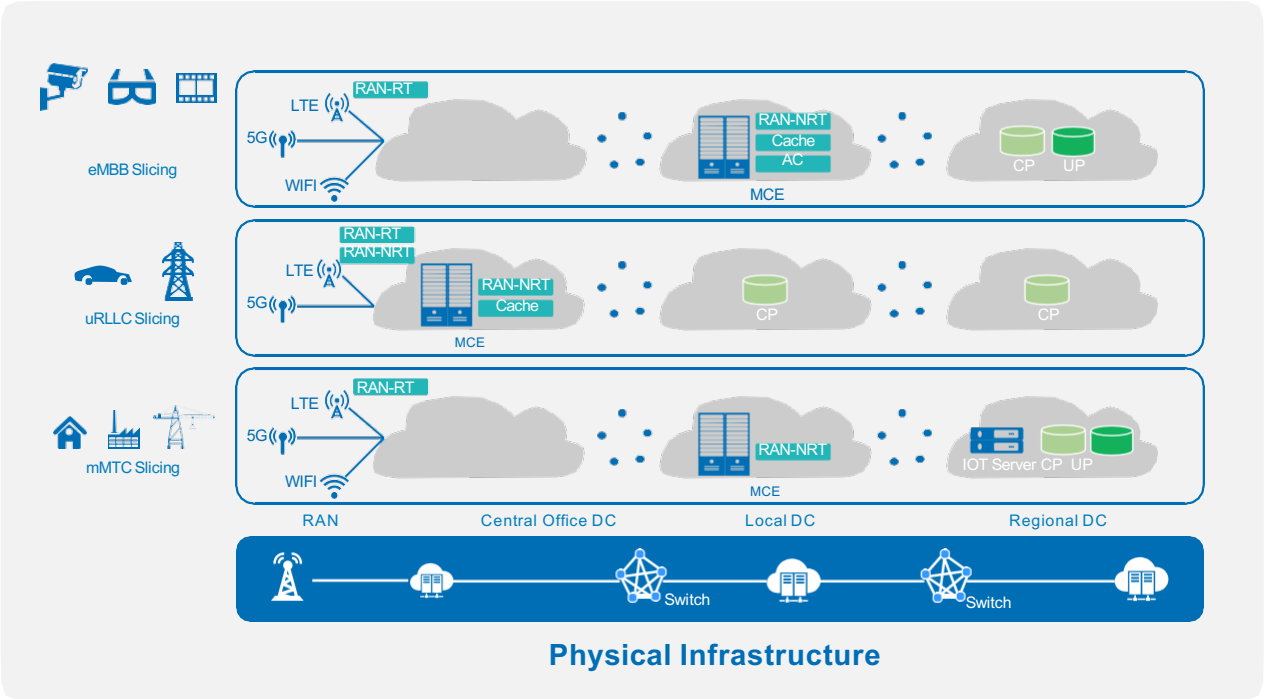
A General 5G Network Architecture



Vertical Applications & 5G



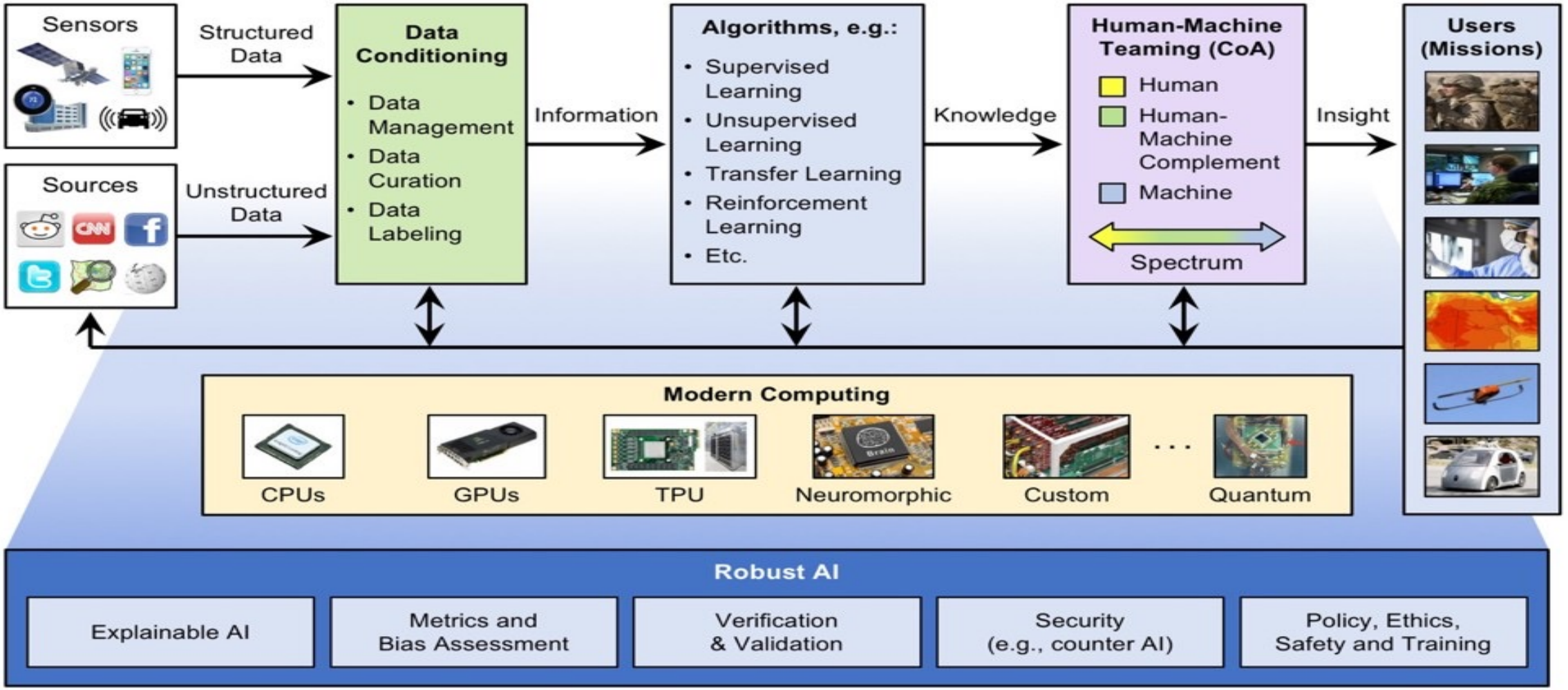
Service-Driven 5G Physical Infrastructure & Network Slicing



E2E network slicing is a foundation to support diversified 5G services and is a key to 5G network architecture evolution.

Future Computing Platforms

Vertical Applications, AI Algorithms & Architectures



CoA = Courses of Action GPU = Graph Processing Unit TPU = Tensor Processing Unit

Von Neumann AI Architectures

- **Harvard University** introduced ParaDNN, a parameterized deep learning benchmark suite, which is a systematic and scientific cross platform benchmarking tools which not only compare performance of different platforms running a broad range of different deep learning models, but also support deeper analysis of the interactions across model attributes, hardware design, and software support.

- **TPU** – **TPU** is highly optimized for large batches for CNNs and DNNs has the highest training throughput

- **GPU – GPU** shows similar performance like TPU but with better flexibility and programmability for irregular computations such as small batches and non-MatMul computations.

- **CPU – CPU** achieves the highest FLOPS utilization for RNNs and supports the largest model because of large memory capacity.

Non-Von Neumann AI Architectures

- **Compute in Memory (CIM)** – CIM arrays based on SRAM, NAND flash as well as emerging memories such as ReRAM, CeRAM, MRAM are being considered as possible reconfigurable, reprogrammable accelerators for neural network computations. Advantages: High performance, density, low power, and latency. Current challenge: Readout bit line analog signal sensing and ADC for specialty RAM processing technologies.

- **Neuromorphic Computing** – Neuromorphic computing will extend AI into areas that correspond to human cognition, such as interpretation and autonomous adaptation. Next generation AI must be able to address novel situations and abstraction to automate ordinary human activities.

- **Quantum Computing** - In quantum computing, the smallest unit of data is the qubit, based on the spin of a magnetic field. Based on quantum entanglement it allows for more than 2 states and entanglement speed is extremely fast. (e.g. Google Sycamore, Quantum Supremay, 53 Qbits, 1.5 trillion times faster, completed a task that takes classical computer 10,000 years to do in 200 seconds). Current challenge: the rate of errors and decoherence in noisy intermediate-scale quantum (NISQ) computers

- **Quantum Neuromorphic** – Quantum neuromorphic computing physically implements neural networks in brain-inspired quantum hardware to speed up their computation.

Vertical Applications & Edge AI

- **Edge AI will dominate future computing.** AI is the technology that will enable future horizontal & vertical applications.

- Horizontal AI applications solved a broad range of problems across many different industries (e.g. **computer vision** & speech recognition). Vertical AI is applied to a specific industry that is highly optimized for that industry (e.g. **HD mapping, localization, and navigation for autonomous driving** & genome sequencing variant calling for cancer treatment).

- With deep domain know-how, efficient AI models & algorithms could speed up computation 10-100000X. This is the most intriguing and important of future AI for autonomous driving.

- All vertical application solutions require multi-level AI models for multi-tasks.

AI Models & Algorithms

- **DNN** is the backbone of AI. Today's DNN uses a learning formalism called backpropagation. Today's DNN are slow to train, static after training and sometimes impractical to change in real applications.

- **Transfer learning** is an approach where a previously developed DNN is "recycled" as the starting point from which DNN learns a second task. With transfer learning, no changes to DNN methodology except one can train a DNN model with less data.

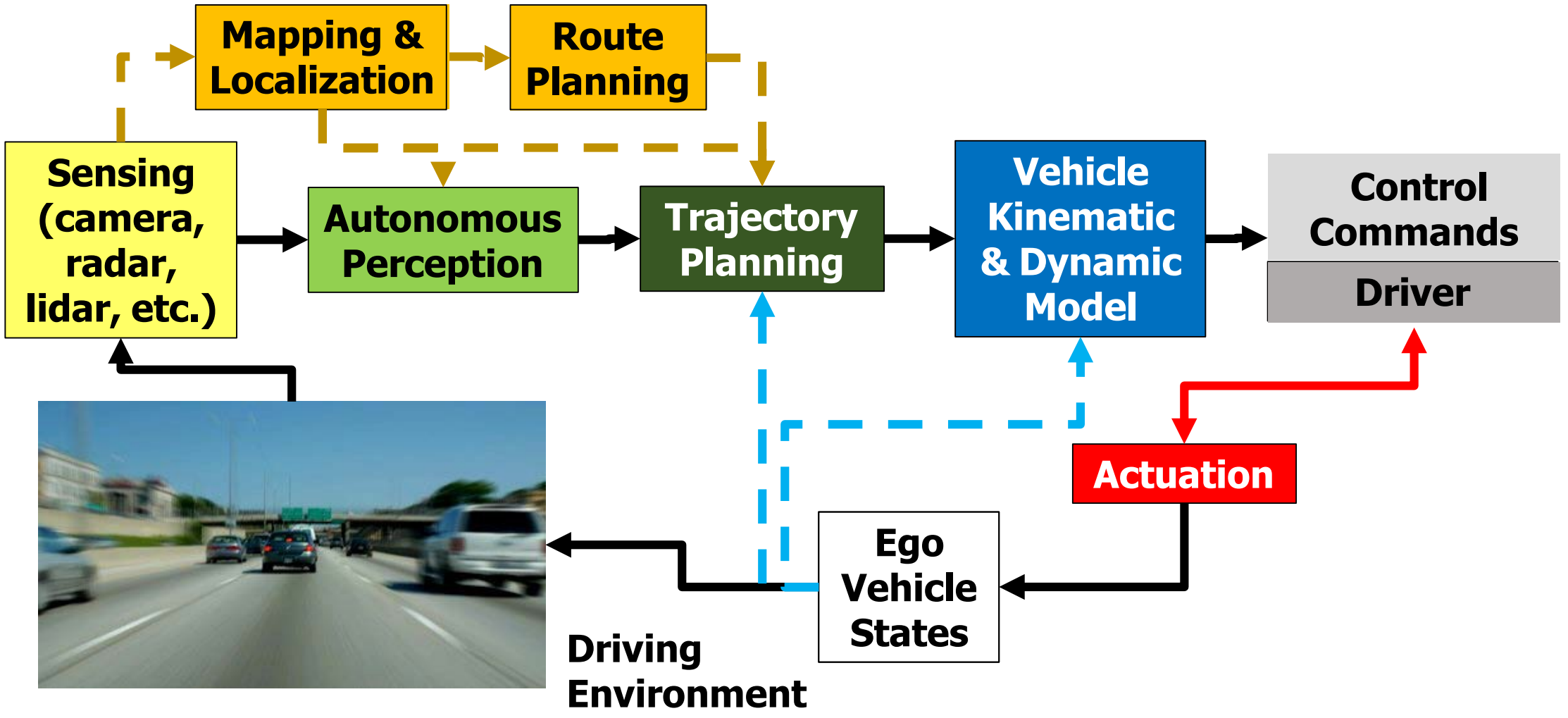
- **Continual (Lifelong) learning** is the ability to continually learn over time by accommodating new knowledge while retaining previously learned experiences. For instance, an autonomous agent interacting with the environment is required to learn from its own experiences and must be capable of progressively acquiring, fine-tuning, and transferring knowledge over long time spans.

- **Reinforcement Continual Learning (RCL)** searches for the best neural architecture for each coming new task via sophisticatedly designed reinforcement learning strategies. RCL method not only has good performance on preventing catastrophic forgetting but also fits new tasks well.

arm

Autonomous Driving & Smart Transportation

Automated Driving Systems (ADS) - Functional Block Diagram



Autonomous Driving Technology Breakthroughs Required

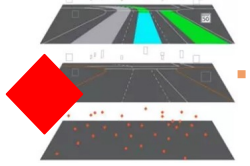
- Precision Localization and Navigation at Edge – Light Weight, Precision Fingerprint based Localization and Navigation.
- Critical Real Time Response – 20-30 MS like human brain
- Eliminate Blind Spots – V2X, V2I, DSRC, 5G
- Commercially Scalable - Low Power & Low Cost

Autonomous Driving Technologies

- Lightweight map data
- Feature extraction algorithm
- Localization algorithm through feature-map
- Updated automatically

Perception

Feature-Map



HD-Map

Enhancement
- Prediction algorithm
- Decision modeling algorithm

Object-based
Driving Policy
Training & Inferencing

Properties

Designed to automotive standards	Optimized for power and thermal efficiency	Low latency sustained in real-time environments
ASIL B-D compliance	ISO 26262	AI ML/DL-based

Sensing

- Sensing
- Camera
- Lidar
- Radar
- IMU
- GNSS
- V2x/ITS Message
- Wayside Assists

Mapping & Localization

Detection

Object Tracking

Perception & Recognition

Route Planning

Prediction

Behavior Planning

Trajectory Planning

Control

Vehicle Kinematic & Dynamic Model

- Kinematic model ML/DL

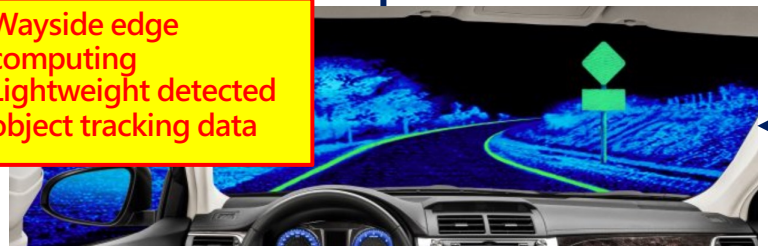
Control Commands

Driver

Actuation

Vehicle states & Trajectory tracking

- Wayside edge computing
- Lightweight detected object tracking data



Driving Environment

The Arm ADAS/ADS Ecosystem

Commercial

DEEPSCALE

Brodmann17

mapbox

Civil Maps

Green Hills SOFTWARE

WIND™

rti

Mentor A Siemens Business

ETAS

Elektrobit

BlackBerry | QNX

OPENSYNERGY

COREAVI

escrypt Embedded Security by ETAS

Sensor Fusion | Perception | Localization

Planning | Maps | Control

Middleware

Operating System/Hypervisor

Security | Platform | Communication

Ubiquitous compute architecture

Open Source

apollo

AutoWare

ROS

AUTOMOTIVE GRADE LINUX

Linaro

arm

arm

Summary

Summary

- Autonomous Driving Requires Handling of Massive Data in HD Mapping, Localization & Perception of Environment All at the Edge in Critically Few Milliseconds.
- Intelligent and precision reduction of data in perception, localization, navigation, reinforcement interaction (driving policy) will allow ADS to shorten latency & respond to ever changing traffic conditions swiftly.
- Powerful, high performance edge AI is one of the key barriers.
- 5G Connection Supports Reliable MIMO Connectivity, Low Latency, High Bandwidth.
- 5G and Powerful Edge AI together with Innovations in HD Mapping, Localization, & Perception will Make Autonomous Driving a Reality.

arm

Thank You
Danke
Merci
谢谢
ありがとう
Gracias
Kiitos
감사합니다
धन्यवाद
تشكر