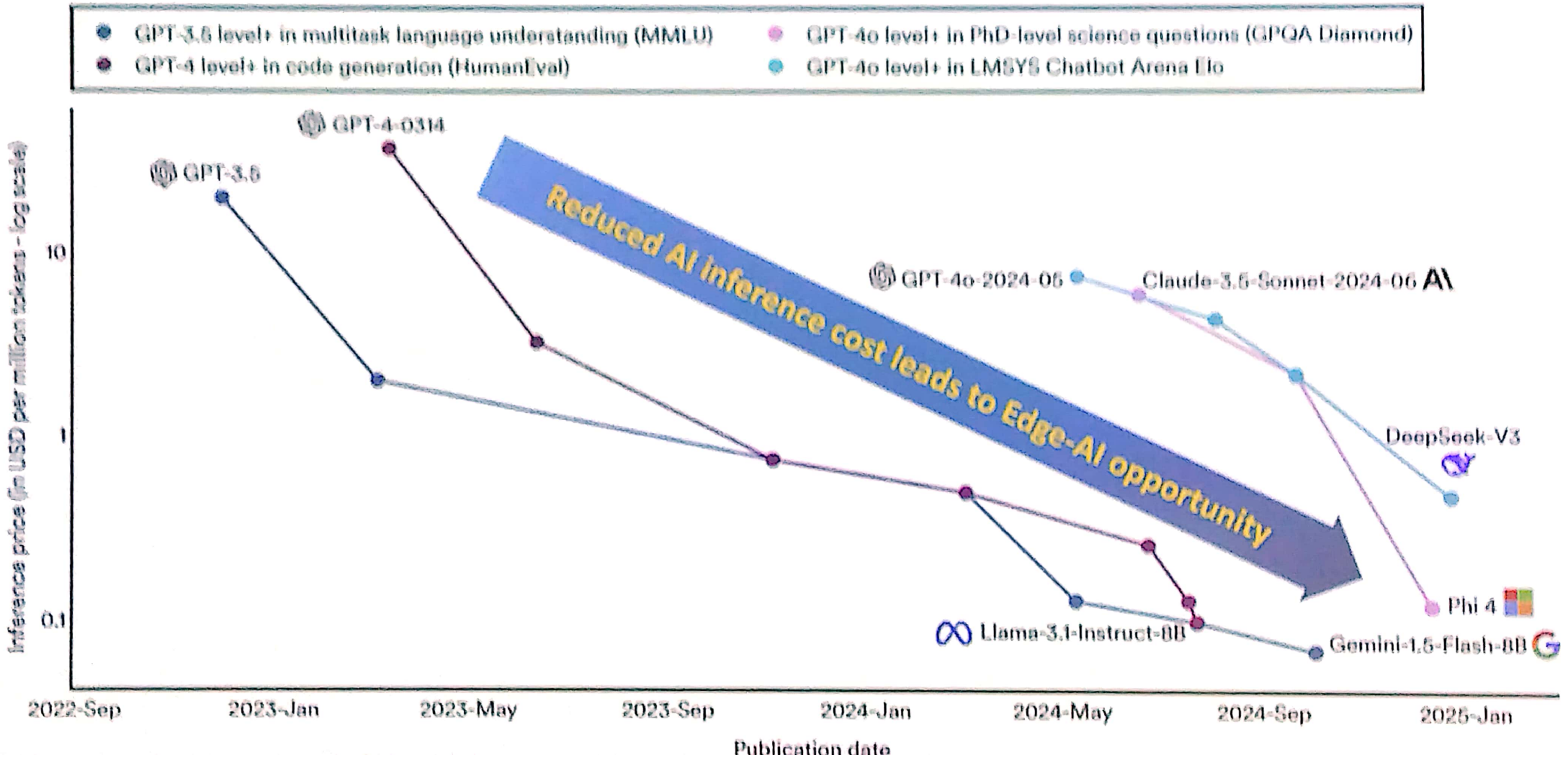
Inference price across select benchmarks, 2022–24

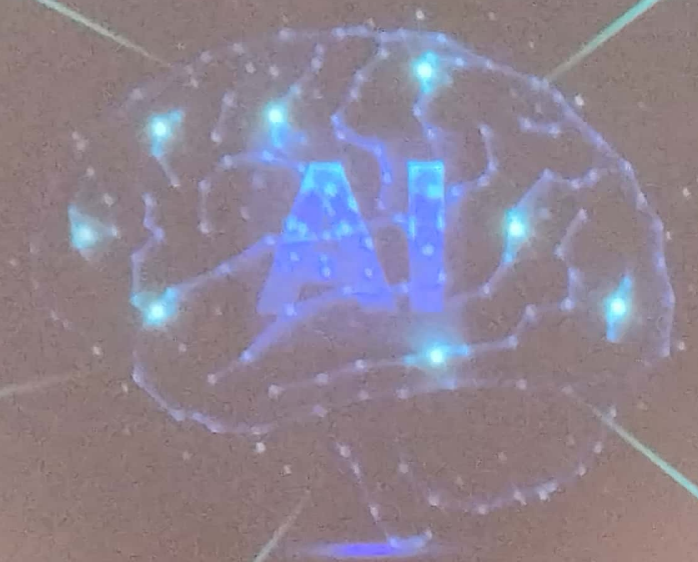
Source: Epoch AI, 2025; Artificial Analysis, 2025 | Chart: 2025 AI Index report



Challenges to enable ubiquitous AI

AI computing / storage offloading

- On-device processing or
- Server-side processing – requiring low latency, high data rate or Hrbrid/distributed computing



Device type & LLM scalability

- Ubiquitous AI experience for all device form-factors (e.g. smartphones, wearables, etc.)
- AI-assisted Modem as the key gateway to enable efficient large AI model operation end-to-end

AI power consumption

- AI data center power estimated to 85~134 TWh @ 2027
- Tokens/mW vs Tokens/s/Hz

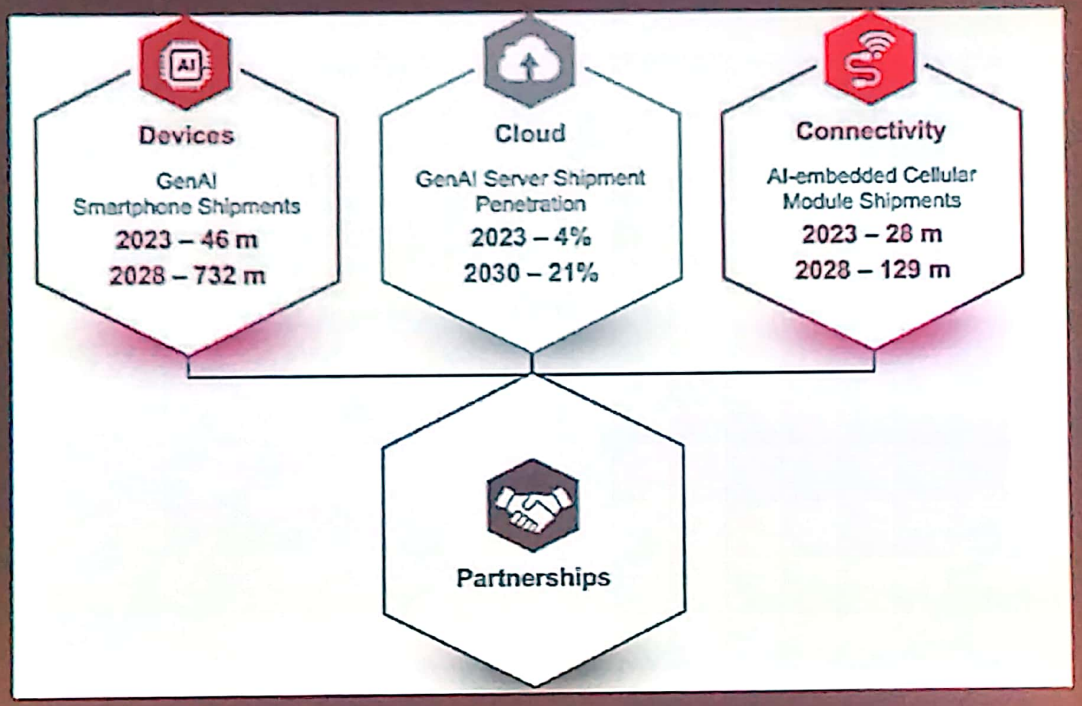
Privacy

- Cloud vs. Device
- Operator vs. Hyperscaler

Service capacity and coverage

- Data rate, latency/QoS and handover interuption
- Token/s @cell edge

Essential components of a *Ubiquitous AI* Solution Provider



1. AI Device (Phone, IoT): Capable of running an LLM on device, smooth user experience, multi-modal content I/O
2. Cloud: Increased AI processing -> More data centers -> Merchant/Custom AI accelerators ->Growing ASIC Data Center vendors who possess advanced Fabless skills (ex: Compute /Interconnect IP, specialized memory, supply chain eco-system)
3. Advanced Connectivity Solutions – Cellular/WiFi/IoT/BT IP's
4. Eco-system Partnerships – Device/Cloud/Connectivity player collaborations

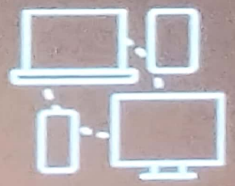
*Counterpoint Research

Wireless Technology Evolution – Priorities from Device Perspective

User Experiences

→

- Latency / Data Rate
- Coverage
- Power



Use Cases

→

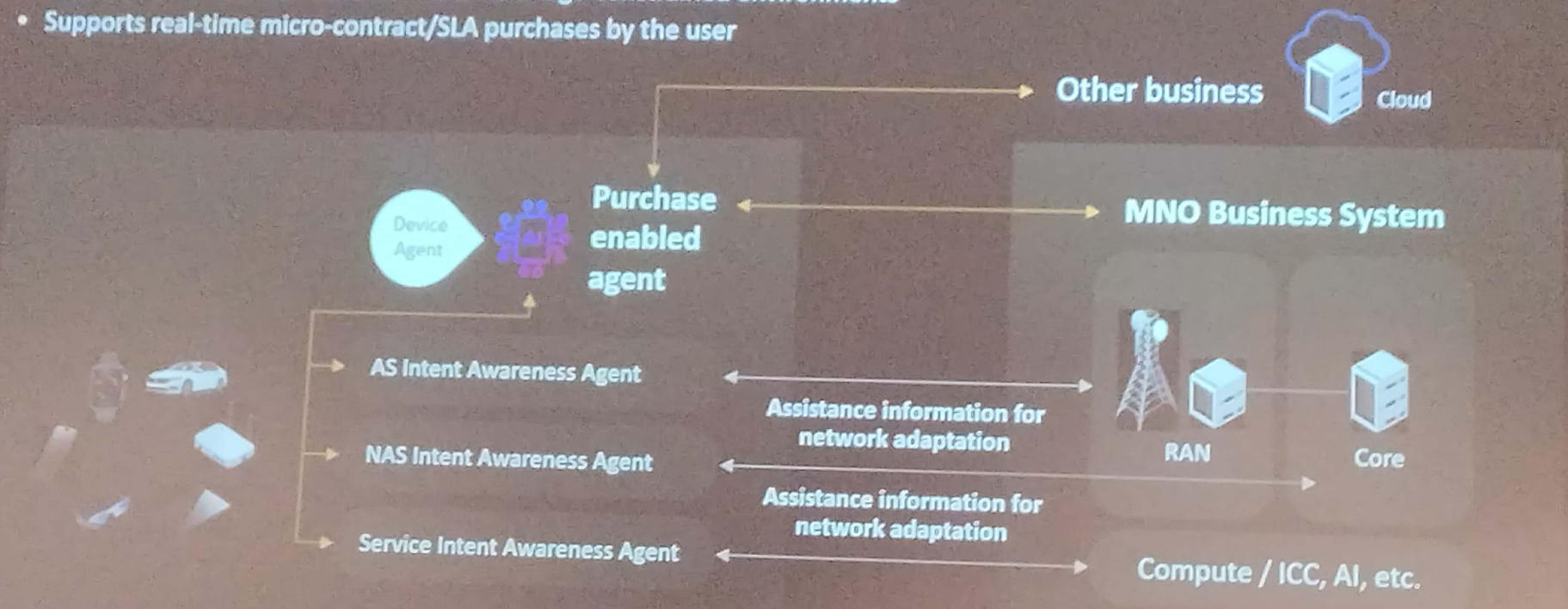
- Applications
- Device Types

On device Agentic-AI : Support multiple applications and Provide best QoS!

Intelligent Systems : Service Awareness using a Device AI-Agent

A lightweight device agent that connects context, application, and intent awareness with network resource

- Enables on-demand, fast adaptation for an exceptional user experience
- Optimized for resource-constrained or coverage-constrained environments
- Supports real-time micro-contract/SLA purchases by the user



6G-enabled seamless AI, with devices serving as user-intent-driven gateways to intelligence

Intelligent Systems : MediaTek in-Modem AI (MMAI) Is Available Today

Using AI/ML to distinguish different types of 5G data transmission to deliver the best performance

- **Predictive traffic power savings**
- **Faster network search and reconnection**
- **Connection robustness with seamless handover**
- **Customizable NN per market, operator, device designer requirements**

Data Traffic Pattern
4G/5G Signal Pattern

AI traffic type classifier

4G / 5G switch forecaster

Traffic type

- VOIP
- Video call
- Audio Streaming
- Realtime Gaming
- Live streaming
- Video streaming
- Large file transfer
- Standby

Signal type

- Seamless 4G/5G handover to avoid inter-RAT back/forth

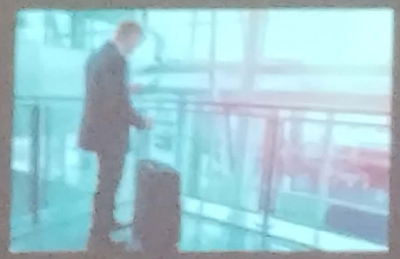
Up to 10% Power Saving

>40% Lower Latency

>20% Lower Latency

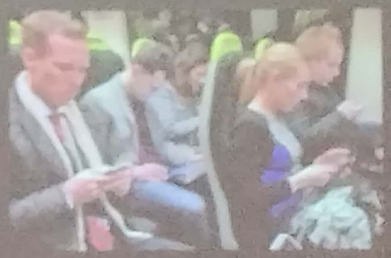
Examples for UE-Device AI: Intelligent Connectivity Optimization

Airport Mode



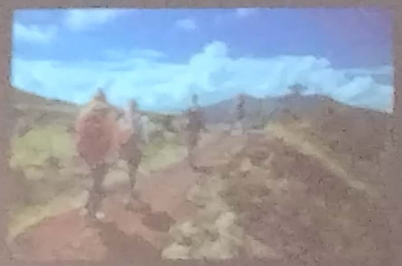
+50% faster roaming power-on

Transportation Mode



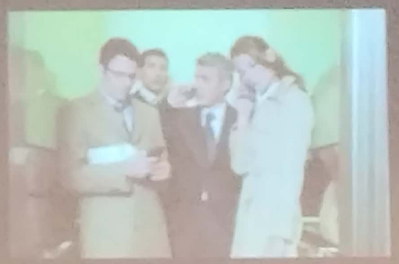
+20% smoother gaming / live video

Hiking Mode



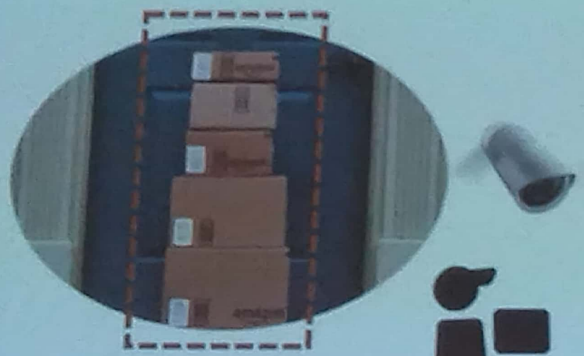
+10% power saving in country area

Elevator Mode



+45% faster 4 / 5G recovery

Gen-AI Gateway – Enabler for Agentic-AI



Camera detected motion at the front door

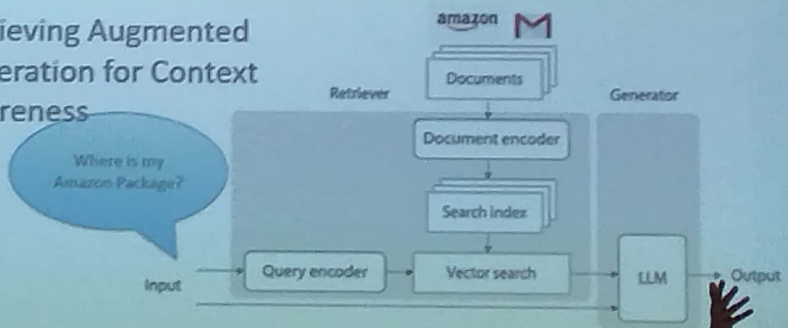
Image/video sent to Gen-AI gateway for object recognition & event analysis

Gen AI Gateway
Llama 3

Query RAG database learned from user's emails

Notify user of parcel arrival

Retrieving Augmented Generation for Context Awareness



MWC25

MEDIATEK

What is "AI native 6G"? Combination of "Wireless-for-AI" and "AI-for-Wireless"

- **Wireless-for-AI:** 6G designed to **optimize AI based application** in terms of **service capacity and QoS boosting**
  - **Integrated communication, computing and sensing** — computing collaboration across cloud, network and edge with joint latency optimization
  - **Token communication** — protocol and air interface designed for fully utilizing token/semantics characteristics with environment adaptations

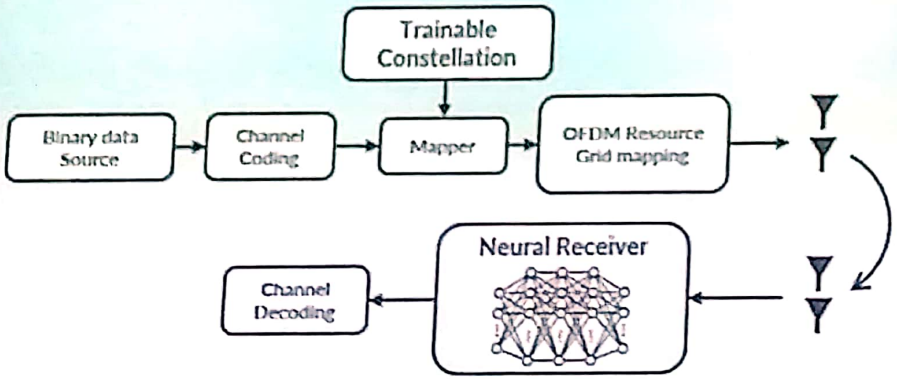
- **AI-for-Wireless:** 6G designed with **AI assisted capabilities** in terms of **personalized and site-specific optimizations**
  - Overhead compression/reduction (two sided CSI, UL JSCC, DMRX-less TRX etc.)
  - Energy efficiency (scenario/service/DoU/location awareness)
  - Security
  - QoS on demand (Modem agentic comm.)
  - LCM/data collection framework

AI-For-Wireless in 6G Era: DMRS-less transmission with AI

Potential for UL Spectral Efficiency Boost

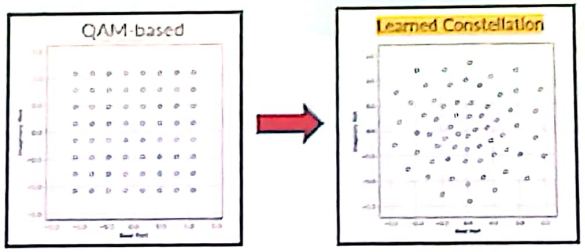
Pilotless transmission and learnable constellation with site-specific adaptation for improved spectral efficiency

Autoencoder-like framework to jointly optimize overall perf.



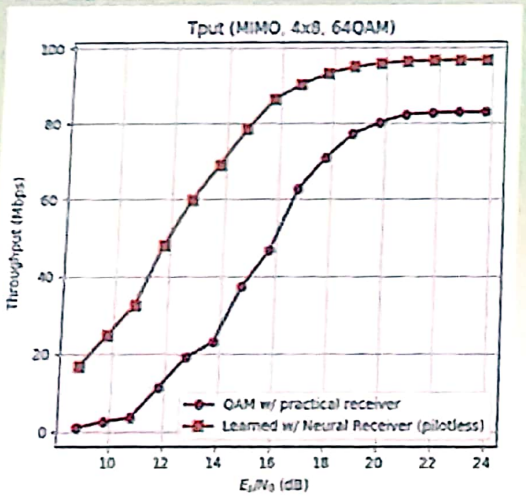
Trainable Constellation to provide additional shaping gain

Allows constellation to be optimized by learning from data to adapt specific channel conditions or RF impairment



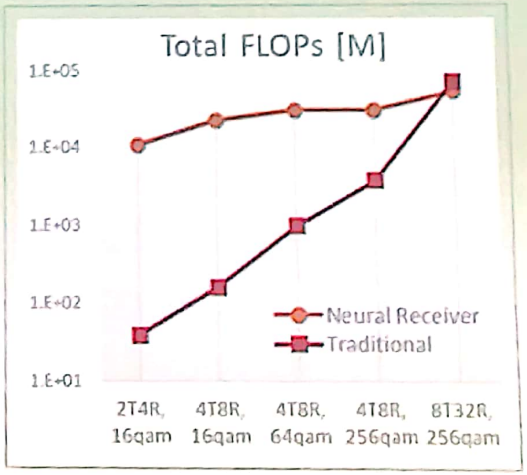
Performance boost

15-45% Tput enh.
depending on MCS
Thank to pilotless transmission and better data recovery capability



Mild Complexity scaling

makes it more feasible for **massive MIMO config.** w. comparable FLOPs w. conventional receivers

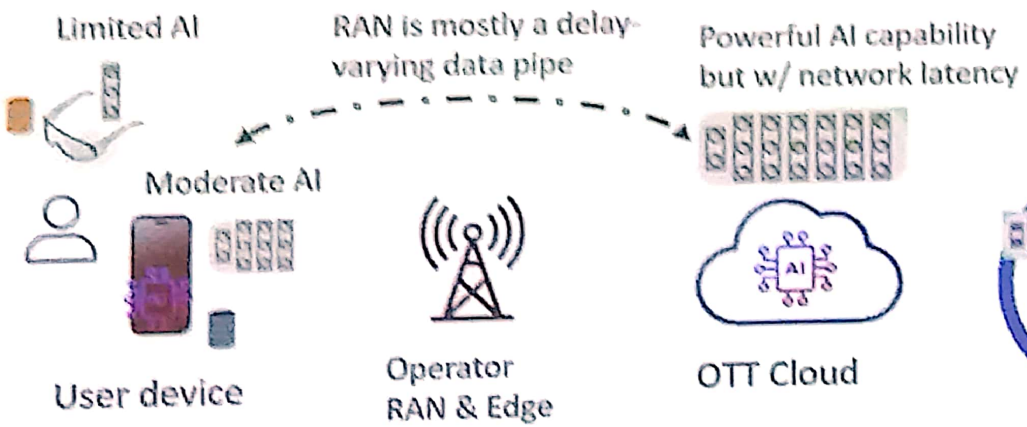


Wireless-for-AI in 6G Era: Compute Architecture Evolution

AI Compute Architecture Evolution in Wireless Network

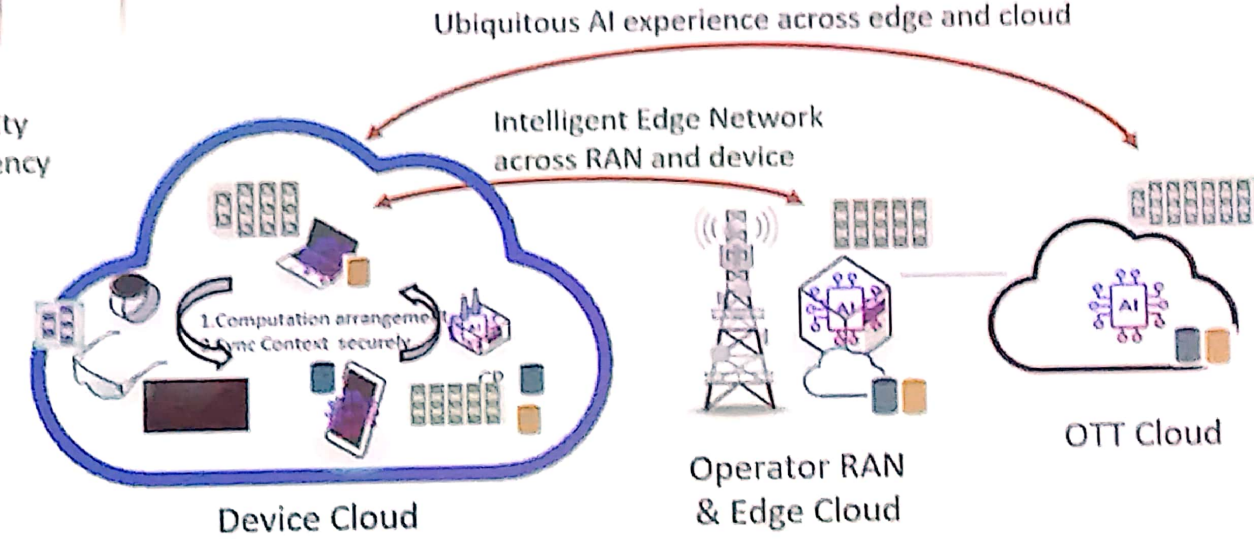
AI on 5G Today – Varying Experiences

Independent Smart Assistance at Edge and Cloud with **limited cross-system collaboration**



AI on 6G Vision – Ubiquitous intelligence with Hybrid-

Agentic AI and **inter-agent collaboration across Device, network Edge, and Cloud**



AI unit (compute + storage) User service context

Ubiquitous AI Evolution towards Intelligent Edge Device Cloud

Enable Premium AI Experience on Wide Range of Device Form Factors

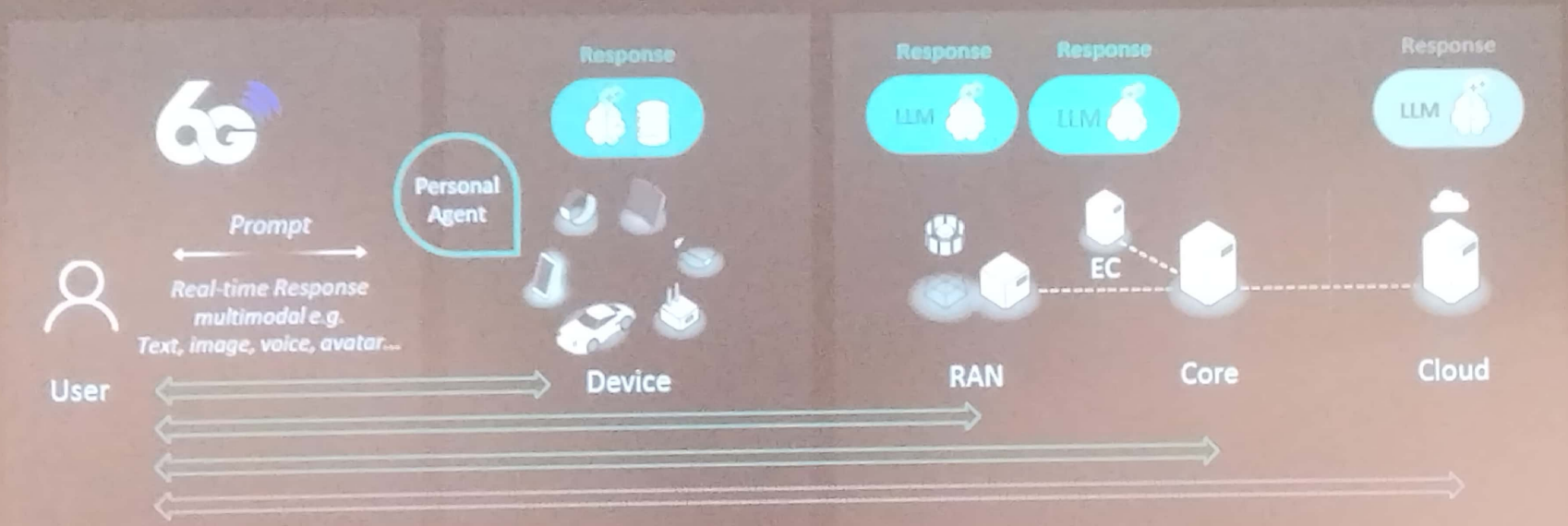
Edge Device Cloud:
Orchestrated computation across edge devices - smart glasses, watches, drones, CPEs, edge servers

Device Collaboration:
Data sharing and task distribution across a network in real-time, improving both efficiency and speed

Multi-Agent System architecture
for distributed AI compute sharing. Help improve scalability, flexibility, and robustness.

- Multiple autonomous agents collaborate – Independent agents communicate and coordinate to achieve shared objectives

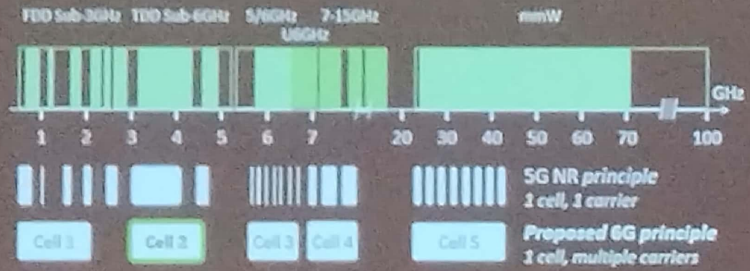
Multi-Agent System examples:
Autonomous Vehicles, Public Safety/Disaster Rescue Robots, Healthcare Coordination, Smart Cities, Transportation Systems, Smart Power Grids



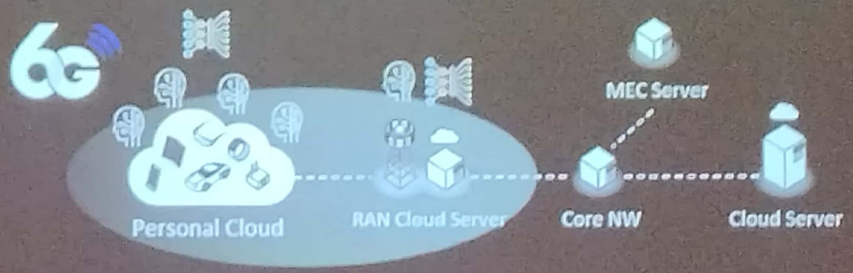
MEDIATEK

Advanced Wireless Technology Enabling Ubiquitous AI

High-Fidelity Air Interface on New Spectrum
(Capacity / Reliability / Scalability)



Ambient Intelligence (Computing / Storage)



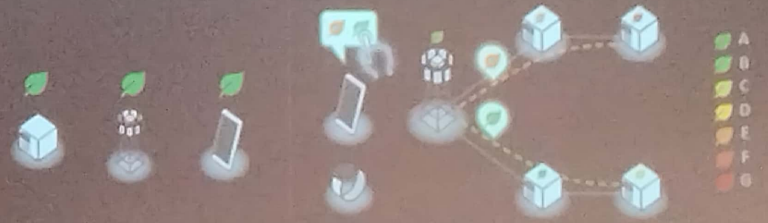
New Edge Cloud = Personal Cloud + RAN Cloud

Non-Terrestrial Network, NTN (Coverage)

TN / NTN Spectrum re-use

Massive MIMO beamforming and adaptation in satellite and device

Sustainability (Power consumption)



Energy / bit → Energy and Carbon / bit

Example Eco-System Enabler for AI Leadership

- Internally as MediaTek – Strong in-house device expertise (Cellular/Connectivity/Automotive/Smart home solutions)
- AI Partnerships - Nvidia, Microsoft, Google, Alibaba (Access to CPU's/GPU's HW/SW, LLM's others)
- Supply Chain Eco-system enables access to latest tech (2nm/3nm), advanced packaging (2.5D/3D) for custom AI chips
- Enabling a large developer community to build attractive AI apps and hence widen the user base

MediaTek Dimensity Chipsets Enable Edge-AI and Agentic AI

NPU and AI

- New MediaTek NPU990 with 2nd Gen-AI Engine
- Newly Added Super Efficient NPU – Industry's first CIM Based
- NPU to support Agentic AI Applications
- 56% lower peak power use
- 2X faster token generation speed
- 4K resolution for high-quality image generation

- Optimized to run various LLM's on-device efficiently – support Agentic-AI applications
- AI engine supports features like BitNet 1.58 bit-model processing and integrated in-compute memory architecture to speed up
- LLM output and reduce power consumption

Takeaways

Edge AI Emerging

- Recent technological advancements have significantly lowered the cost of AI inference, making edge AI more practical and enabling enhanced privacy and personalized experiences.
- Effective data collection remains complex due to issues such as user's consent, security, fragmented data sources, and ownership, etc.

Value Creation Beyond AI Computing Race

- MediaTek is focusing not just on arms race for AI computing power, but on delivering real benefits for users and carriers—such as improved user experience, network capacity, and lower TCO.
  - –AI-for-Wireless → MMAI; 6G designed with AI capabilities (increased Tput, network optimization, Security, QoS)
  - –Wireless-for-AI: 6G ICC & Device Agent

New Revenue Opportunities for Telcos and OEMs

- AI-enabled chipsets are unlocking new consumer market by new device type and new revenue streams by powering innovative services, driving the next wave of value creation in the telecom and OEM industries.
- Building strong ecosystem partnerships is essential to maximize value creation and ensure seamless integration of AI solutions.