Imagination

A **software-first** approach to…

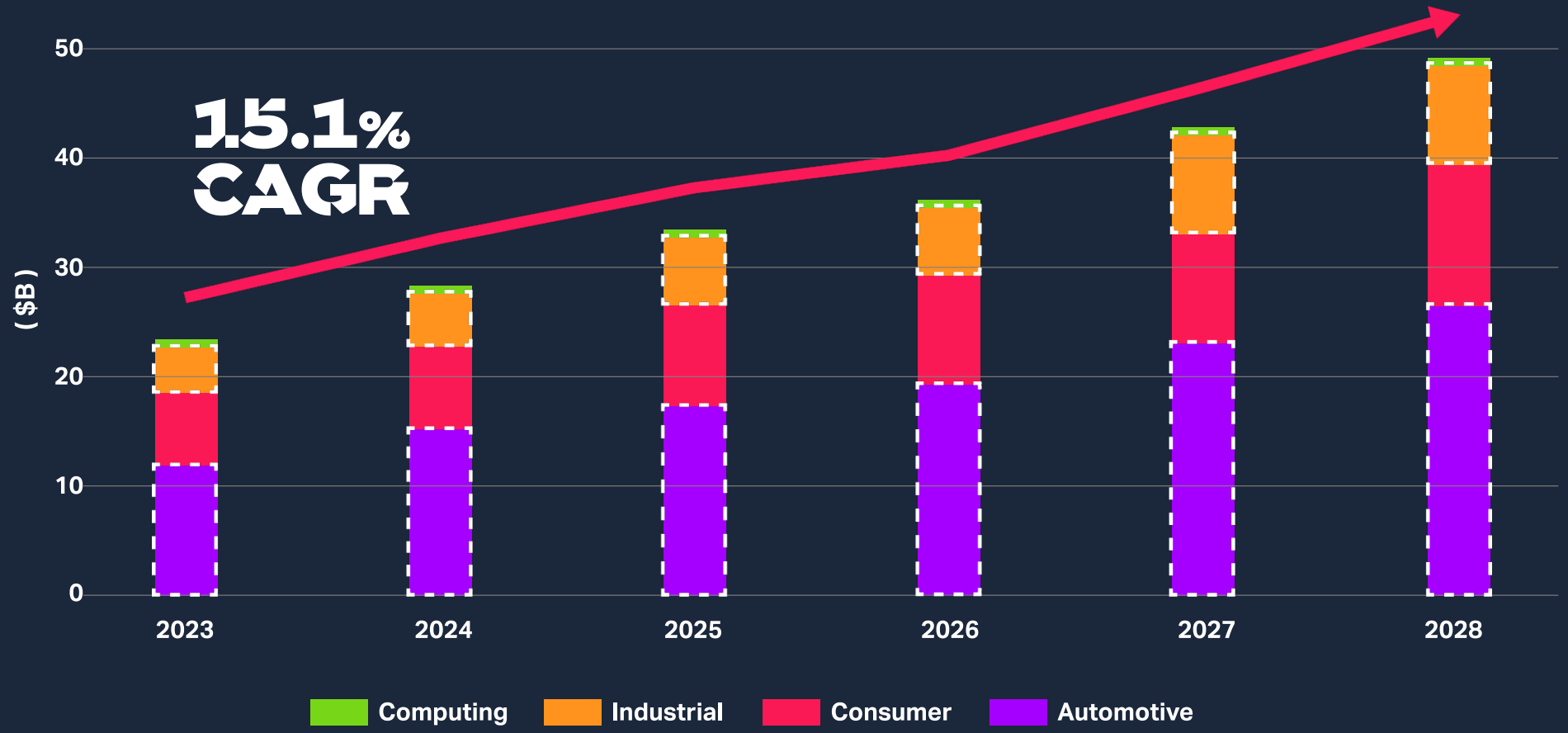
AI AT THE EDGE

Speaker:

Tim Mamtora
Chief of Innovation
& Engineering,
Imagination

Huge societal opportunity for AI to “do good”

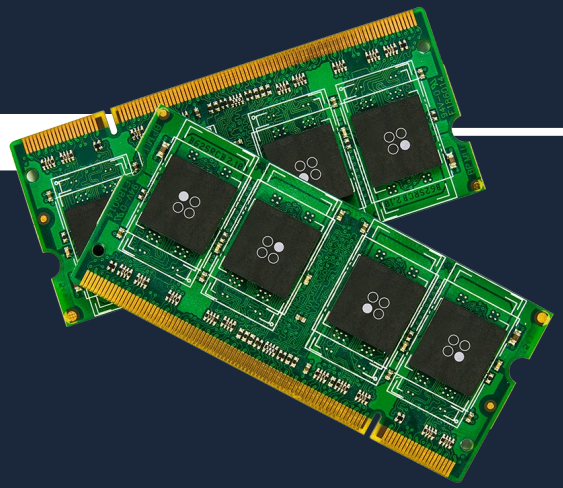
Worldwide Edge Endpoint AI Processor and Accelerator Forecast, 2024–2028



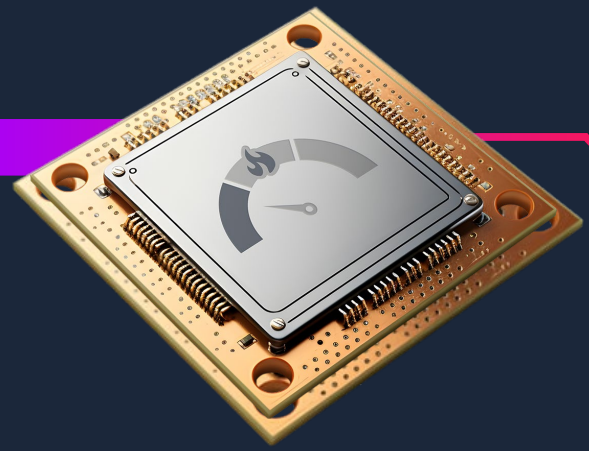
Source: IDC, #US50888824, May 2024

Challenges at the edge are all too familiar

PERFORMANCE EFFICIENCY



HETEROGENEITY



THERMAL MANAGEMENT

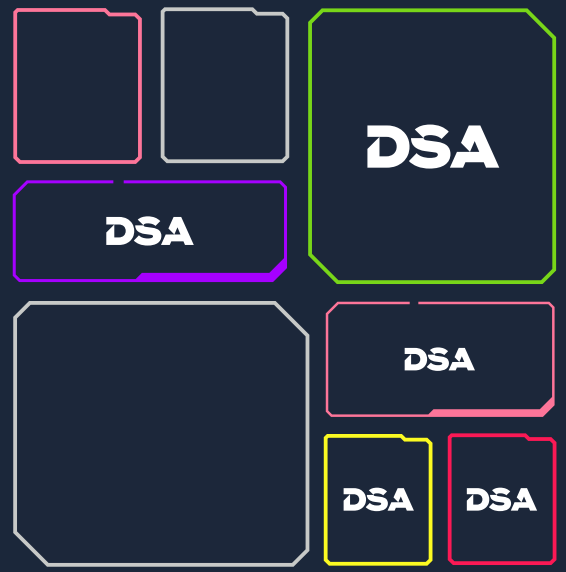
Two key lessons: Scalable and Open

“**The biggest lesson** that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin”

“**Directly competing with open source is a losing proposition…** Open-source models are faster, more customizable, more private and pound-for-pound more capable”

Taking a hardware-first view…

Traditional Edge Device Model
Domain Specific Accelerators (DSAs) and rich IP ecosystem



- Hotspot management challenges
- Silicon integration challenges with increasing transistor count per device
- IP selection and supply challenges beyond 2nm
- Heterogenous programming challenges

Future Edge Device Model
Borrowing from cloud design philosophy

- Simpler thermal management
- More tractable silicon integration whether monolithic, 2D or 3D
- Simpler IP management scenario
- It scales!
- It’s simple to use!

Taking a software-first view…

SUPPORTED
Yes, you can run it

FUNCTIONAL
It works

PERFORMANT
Sufficient Performance

OPTIMAL
As good as it gets

Computational Generality and Ecosystem Compatibility

- Non-proprietary entry points
- Avoidance of premature optimisation

Ease-of-use and Good SW and Libraries

- Avoidance of premature optimisation
- Good quality and extensive libraries and tools

Solution-level Energy Efficiency

- Sufficient degrees of freedom to address workload specific needs

User can exert **fine control** over the hardware

Enter the **Unified Acceleration Foundation (UXL)**

OUR MISSION

- Build a multi-architecture, multi-vendor software ecosystem for all accelerators
- Unify the heterogeneous compute ecosystem around open standards
- Build on and expand open-source projects for accelerated computing

UXL
Unified Acceleration Foundation

Steering Members

arm, FUJITSU, Google Cloud, Imagination, intel, Qualcomm, SAMSUNG, vmware by Broadcom

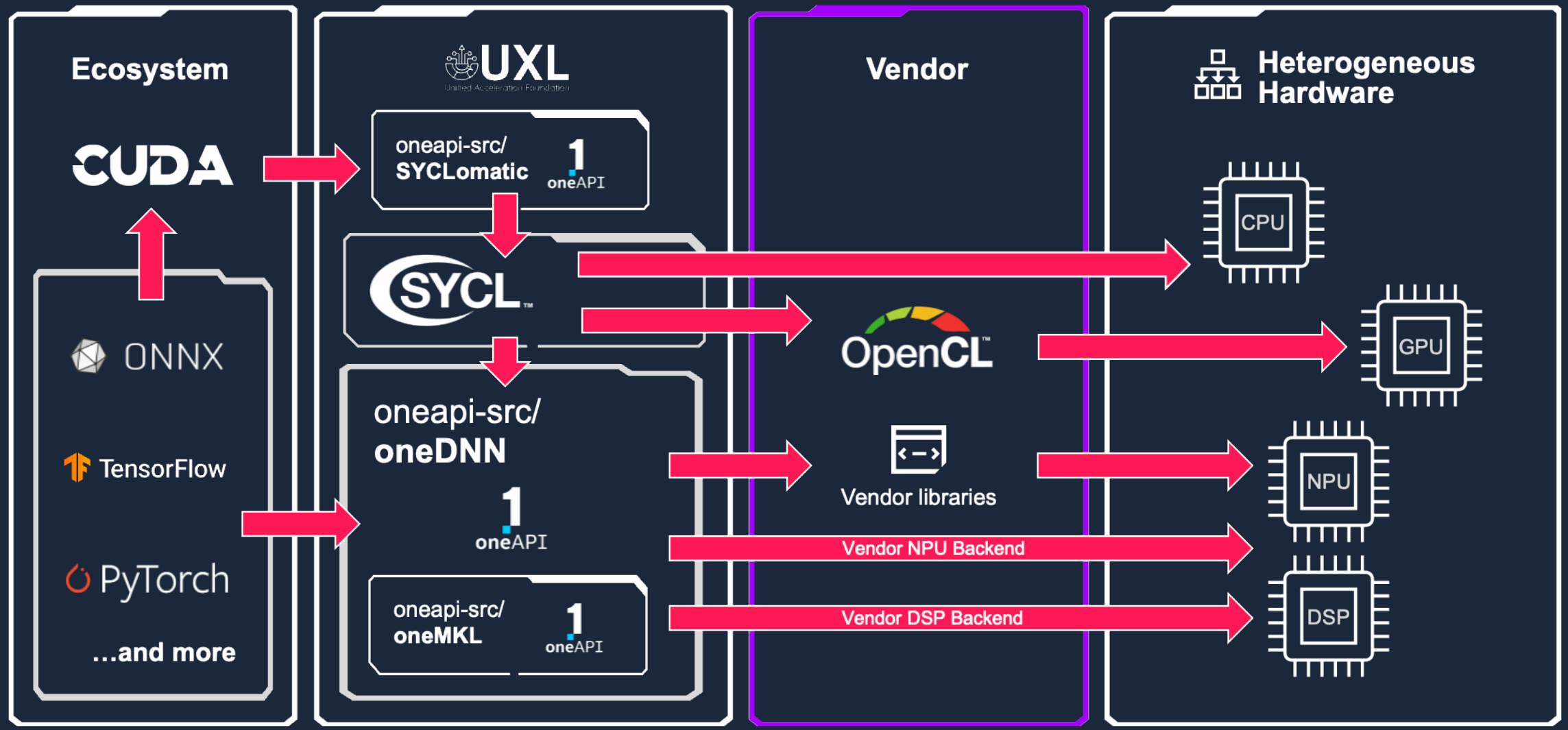
Contributor Members

AKHETONICS, AXELERA, CloudsAI, codasip, CODE RECKONS, DENSO, EMBECOSM, FIXSTARS Speed up your Business, G, InspireSemi, MARKET POTENTIAL, wipro, UNTETHER AI, XIANGDIXIAN

Affiliations

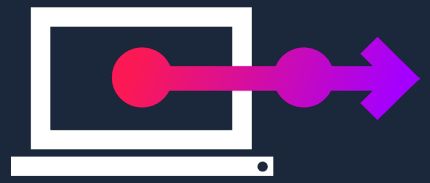
THE AUTOWARE FOUNDATION, KHRONOS GROUP

A fully supported developer journey



Libraries, models, toolkits and profilers *are* the product

Leverage open-source for speed / flexibility



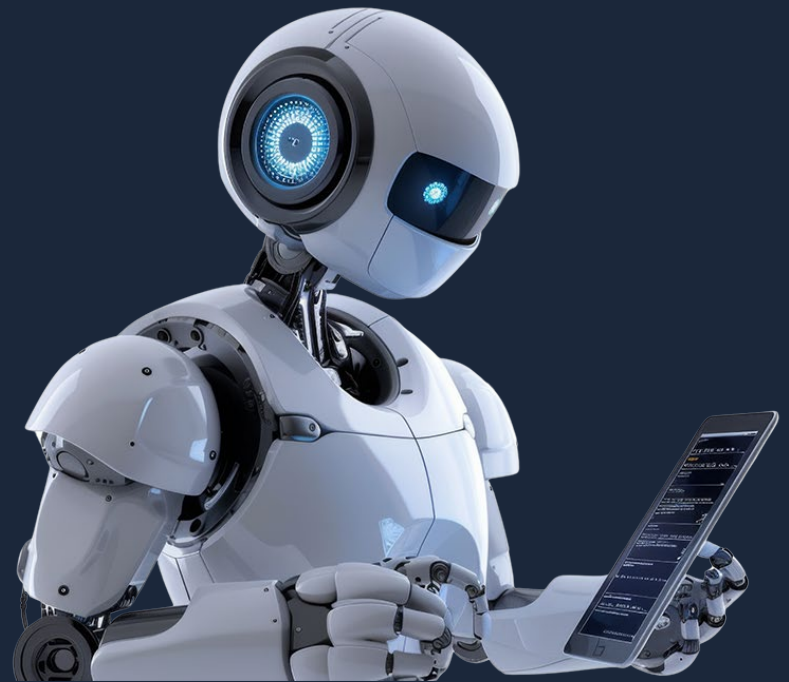
Shift left with performance-indicative models

Guide workload deployment through reference stacks

Flexible architectures for dynamic operation

Edge autonomy moves us on even further…

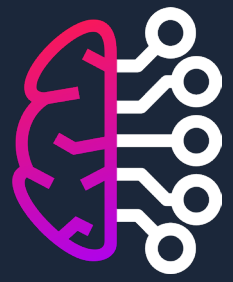
REAL-TIME PROCESSING



SAFETY

… and emphasises the importance of a software-first mindset

Functional safety / SIL



Constitutional AI / XAI

On-device tuning

Open, modifiable software

SUCCESS AT THE EDGE

Conclusions: Take a software-first approach

1

Edge opportunity **“for good”** is enormous

2

Hardware innovation continues abound

3

Defragmentation essential through open software collaboration

Imagination

4

Autonomy increases need for flexibility and transparency

5

Success at the edge requires a software-first mindset