A Re-Imagination of Embedded Vision System Design

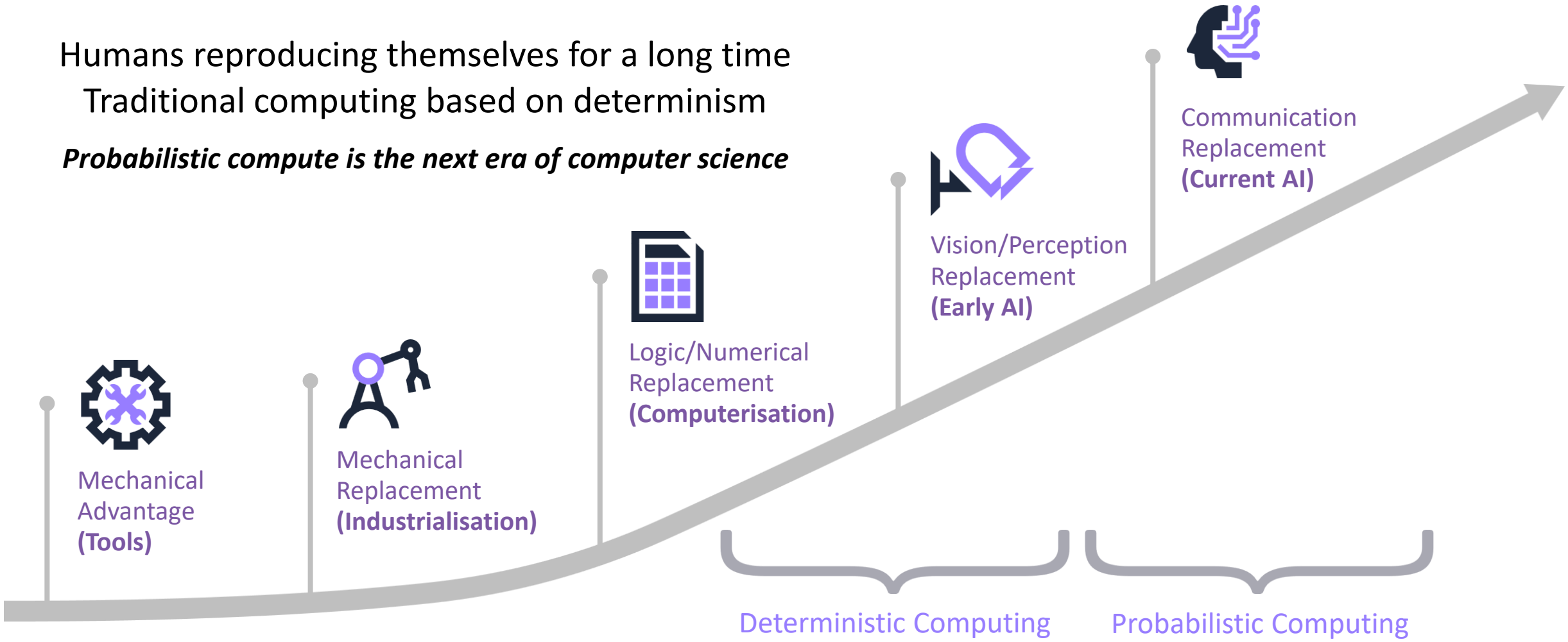
Dennis Laudick
VP of Product Management - Imagination Technologies

Imagination

AI Not a New 'Workload', It's a New Standard for Software

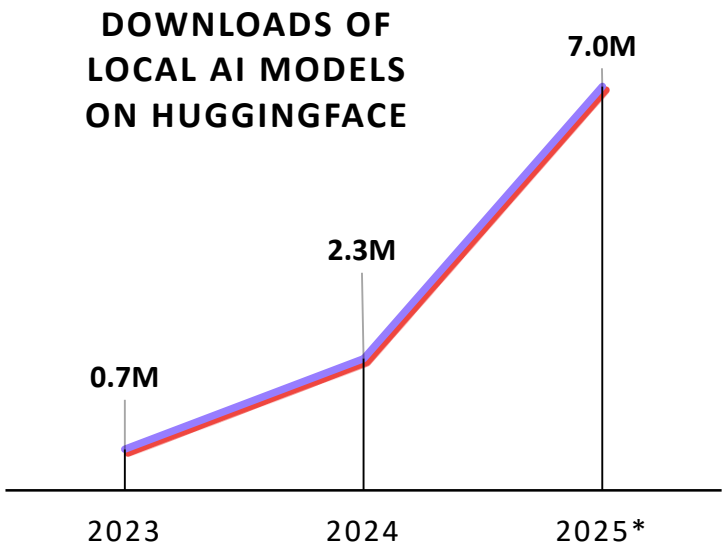
Humans reproducing themselves for a long time
Traditional computing based on determinism

Probabilistic compute is the next era of computer science



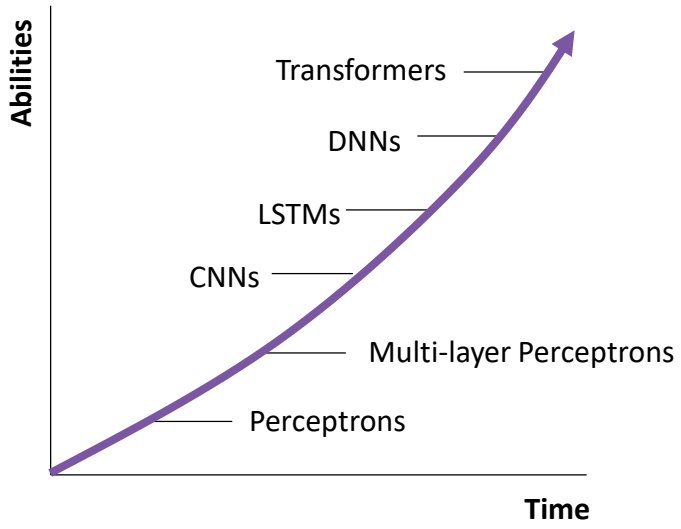
AI Software Growing, But Will Change, Forever

More Models are Available, with More Capabilities

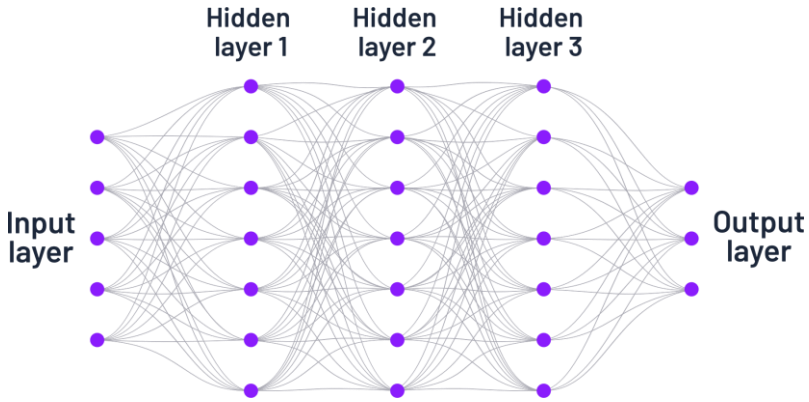


Source: *Huggingface Stats, Github*
** = Forecast*

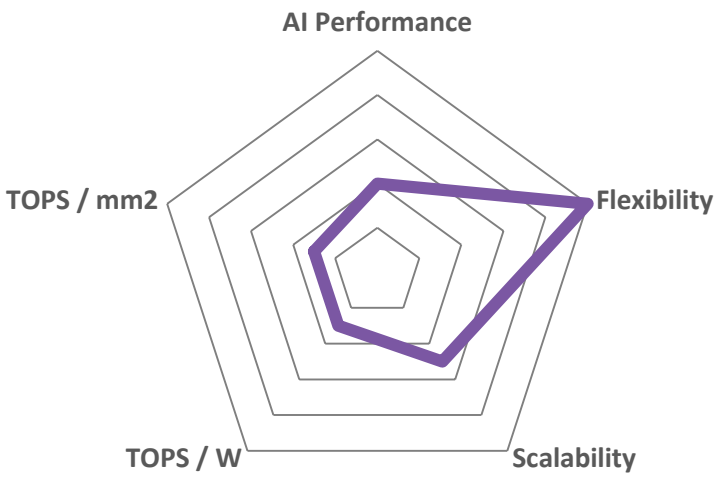
AI Algorithms Still Evolving



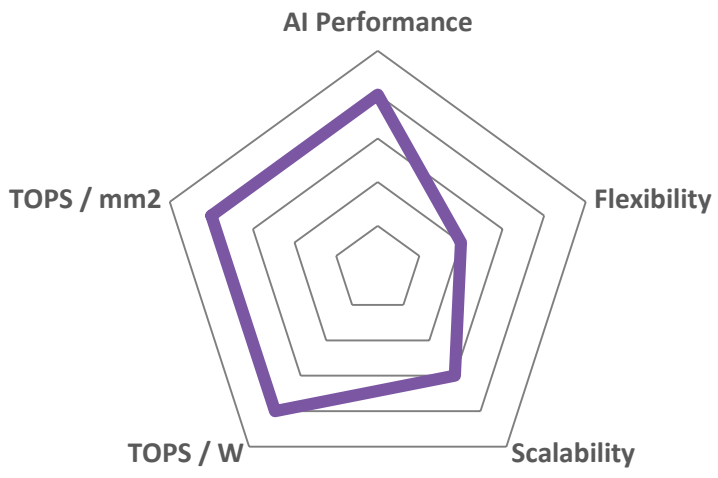
AI Remains a Parallelisation Problem



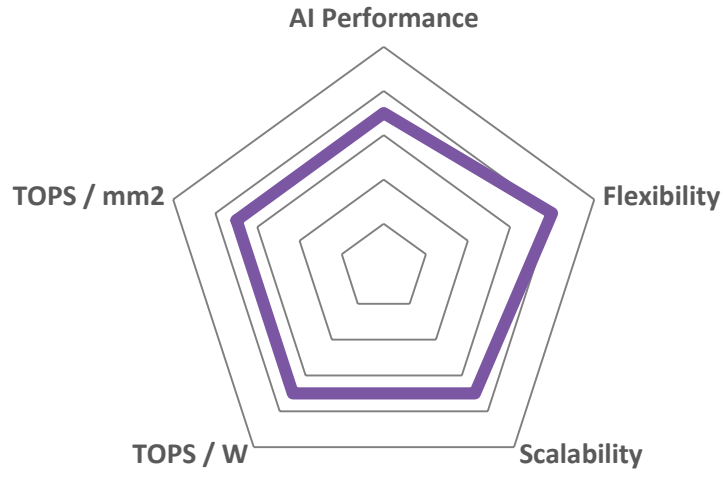
But Hardware Needs to Adapt



CPU



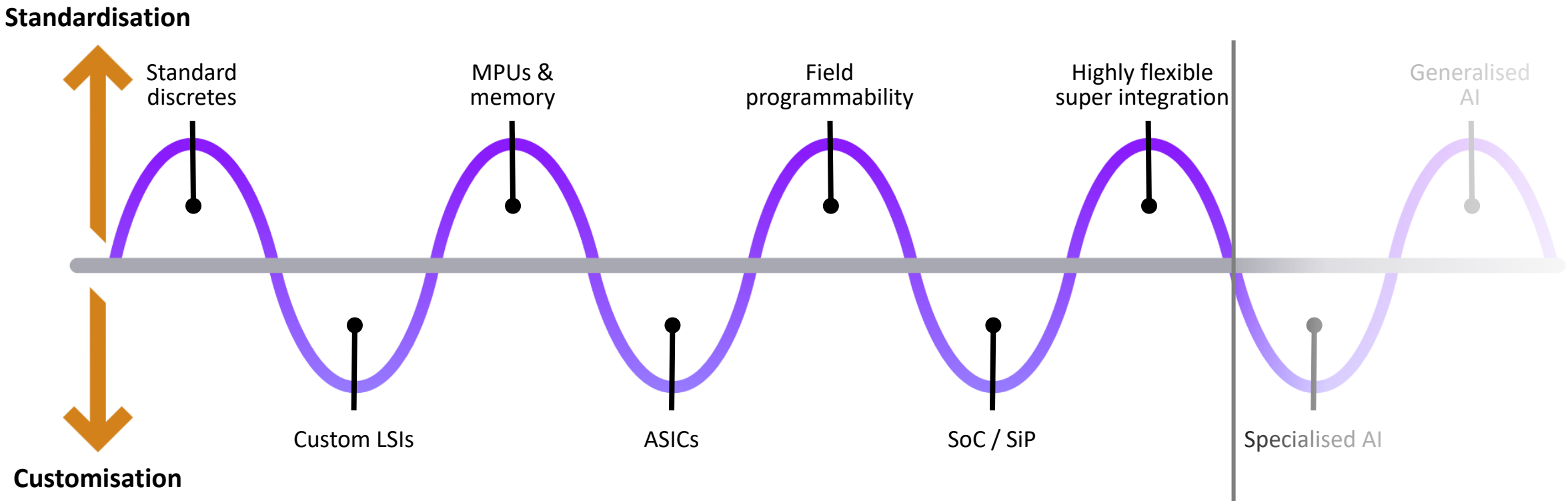
NPU



GPU

Makimoto's Wave and AI Hardware

There will always be a use for specialised hardware, but the macro trend remains

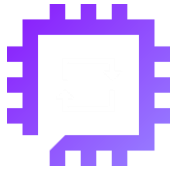


Bringing Parallel Processing Innovation to AI

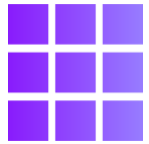
Tradition of Firsts

- Tile Based Deferred Rendering
- Mobile GPUs
- OpenCL on Mobile
- Priority Based Rendering
- Full HW Virtualisation
- FuSa Certified GPU
- Advanced Ray Tracing

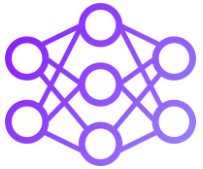
Foundations For High Efficiency AI



Optimised
data management



Tile-Based rendering
→ **Tile-Based compute**

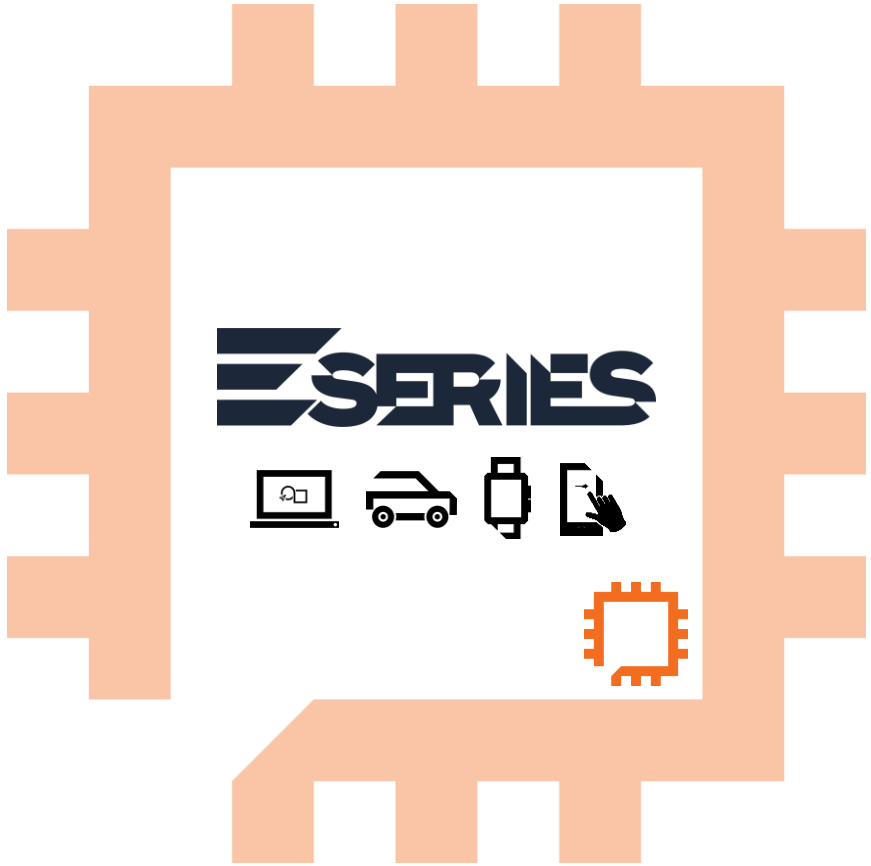


Deferred rendering
→ **Deferred compute**

High utilisation graphics
→ **High utilisation AI**

The New Imagination E-Series GPU IP:

Transforming EDGE system design with programmable AI acceleration



Ultimate EDGE efficiency

New **Burst Processors** pipeline structure boosts average power efficiency by **35%** for graphics and AI workloads.

AI for every device

Scaling from 2 - 200 TOPS, the new **Neural Cores** deliver unprecedented compute density.

Developer & system flexibility

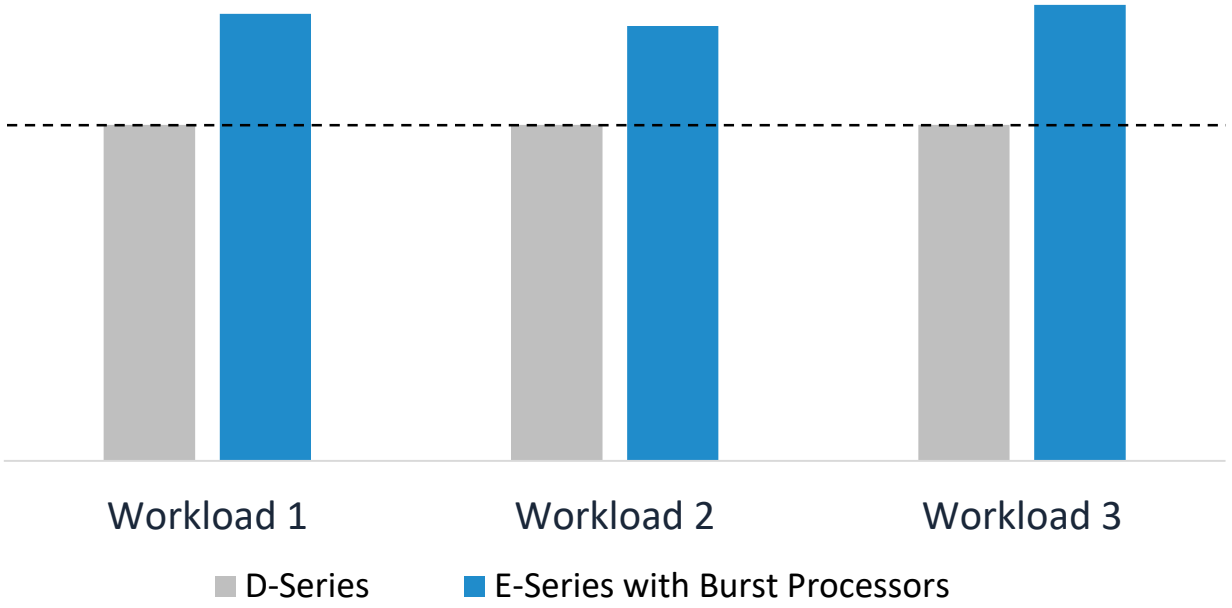
A highly versatile edge processor for graphics or AI acceleration - or both simultaneously.
It is easy to get code running on E-Series, and from there find optimal performance

E-Series Redefines Efficiency at the Edge

Power efficiency matters:

- Battery life
- Performance sustainability
- Operating costs
- Trade for more performance

+35% Average Power Efficiency (FPS/mW) Gains



Burst Processors Reduce Data Movement

Legacy ALU Architecture

Dependant on GPU Register Store for all data read & writes

Burst Processors

Localized data processing to reduce data movements

Key advantages of Burst Processors:

- Scheduled bursts of instructions minimise controller overhead
- Redesigned ALU Pipeline with reduced pipeline depth lowers power consumption and improves occupancy
- Reusing data from local storage, not the big power-hungry GPU register store reduces power consumption
- Works within the new Neural Cores to deliver power-efficient on-GPU AI processing
- Also improves efficiency of graphics workloads

E-Series' Neural Cores Deliver Up to 200 TOPS INT8 AI Performance

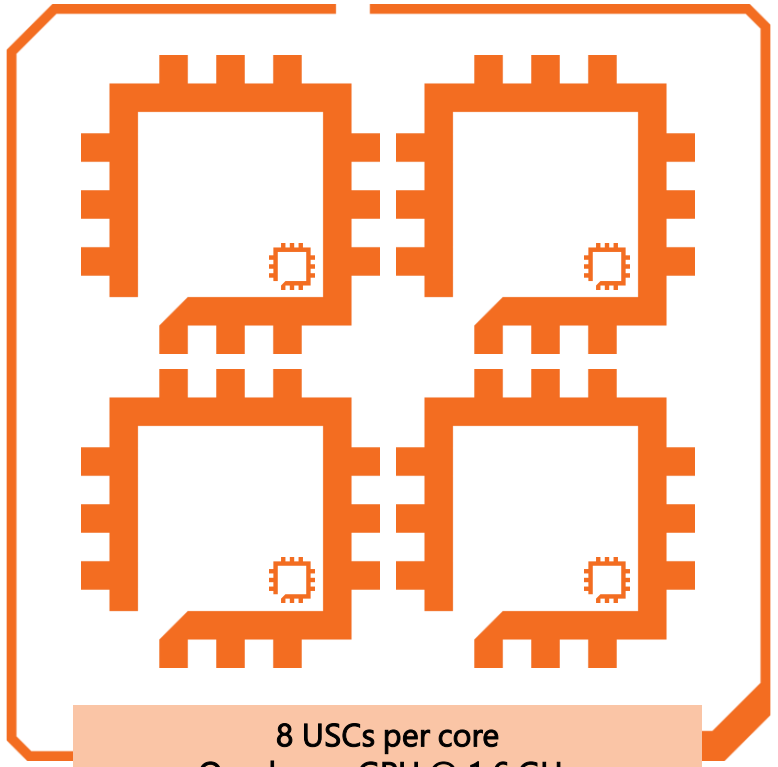
EMBEDDED VISION SUMMIT

Power-efficient AI operations are up to **16x faster**

400
Gpixel/s

13
TFLOPS FP32

>200
TOPS INT8



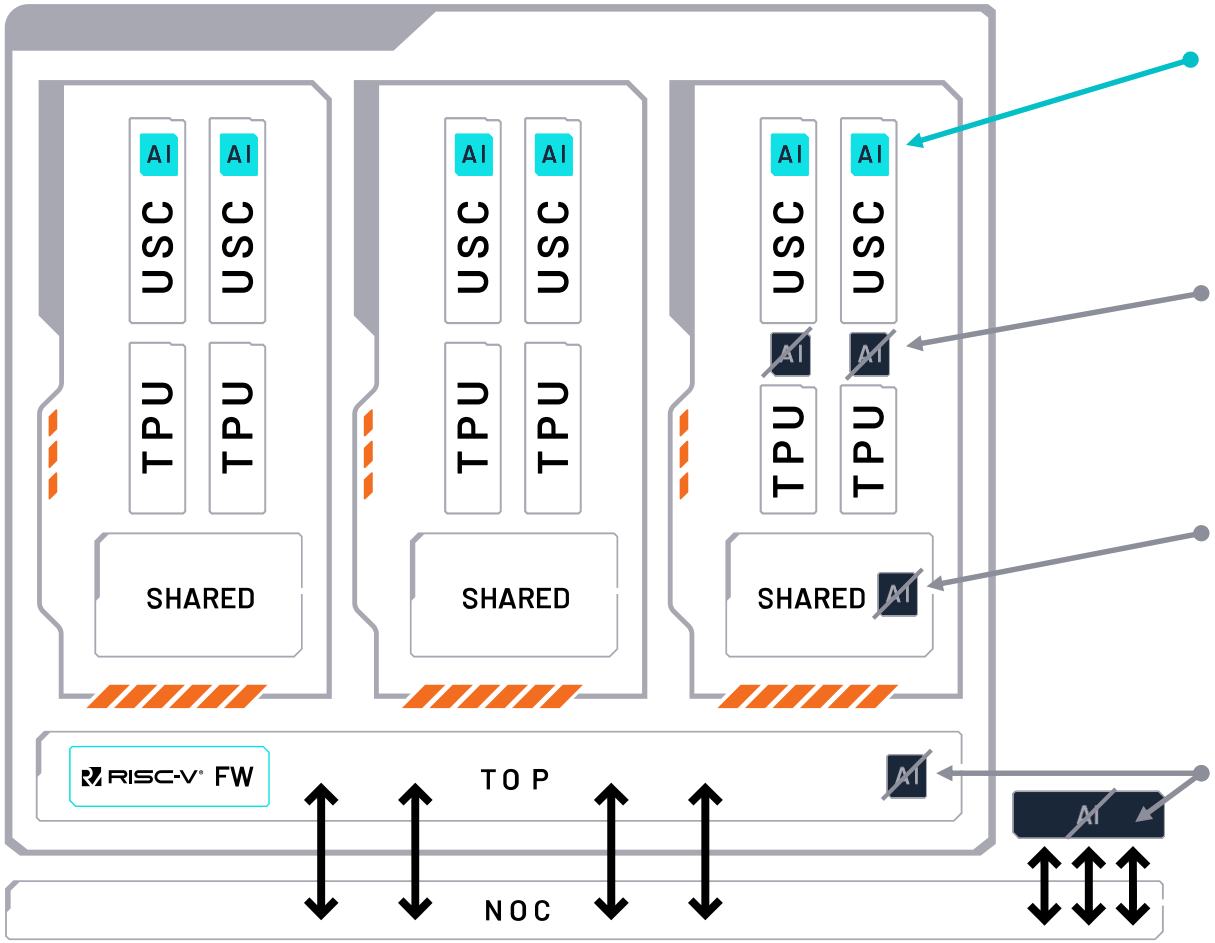
100
TFLOPS BF16

3.6x
TOPS/mm2

Focus on compute density for **area-efficient AI**

FP32, FP16, BF16, INT8, FP8, MXFP8, FP4, MXFP4
SUPPORTED

There is Only One Approach for GPU-Based AI Acceleration



Deeply Embedded Integration

- Shares Registers / SRAM with classic GPU ALU USC
- Minimal data travel distance for re-use – near memory compute
- Matches modern OpenCL and Vulkan AI / Compute Extensions

GPU Accelerator Level Integration

- Non-shared registers, extra SRAM cost for local storage
- Data copies needed to co-op with classic ALU pipes
- Mismatches modern extensions, extra power costs

Shared Logic GPU Level Integration

- Even further away means ever more dedicated SRAM
- Large distance data movement for co-op between ALU and AI unit
- Mismatches modern API extensions

Top Level Or NOC Level Integration

- Loosely coupled, massive dedicated SRAMs, very large data travel distance and latency, poor power efficiency

Imagination's AI-Friendly Memory Hierarchy Sustains Performance and Lowers Power Consumption

32KB Local Memory per USC

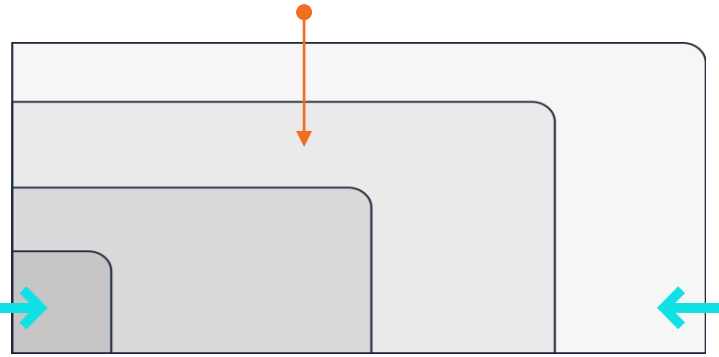
512KB Register Store per USC

Dedicated Texture Cache per TPU

Dedicated Data Cache per SPU

Up to 2MB GPU Specific L2 Cache

Configurable System Level SRAM



- Focus on near memory compute
- Keep data close and on-chip for best power efficiency.
- Give processing pipelines the storage resources to operate properly.
- Reuse shared data efficiently.
- Reduce memory traffic to external caches.
- Lower pressure on (expensive!) registers.
- Speed up data access.
- Improve data broadcast across USC pipelines.

3rd Party NPU

Flexible Hardware based MailBox signalling between IPs

Programming Model and Compatibility Advantages of E-Series

FEATURE	E-Series GPU	CPU	DSP+MMA	NPU
Parallel Processing	High	Low	High	High
Programming Model	Standard	Standard	Proprietary	Proprietary
Programmability Ease	High	High	Low	Medium
Developer Experience	Good	Good	Poor	Poor
AI Data Formats	Broad	Limited	Specific	Specific
Workload Flexibility	High	High	Low	Low

Key advantages of E-Series GPU:

Faster AI compute, with specialized acceleration for quantized models

Industry-leading power efficiency with Burst Processors technology

Standard APIs & comprehensive developer tools matching industry developments & innovation in AI

Flexible architecture that supports simultaneous graphics & AI compute workloads

Multitasking Mechanisms Within E-Series

Asynchronous **Compute**

2D Graphics

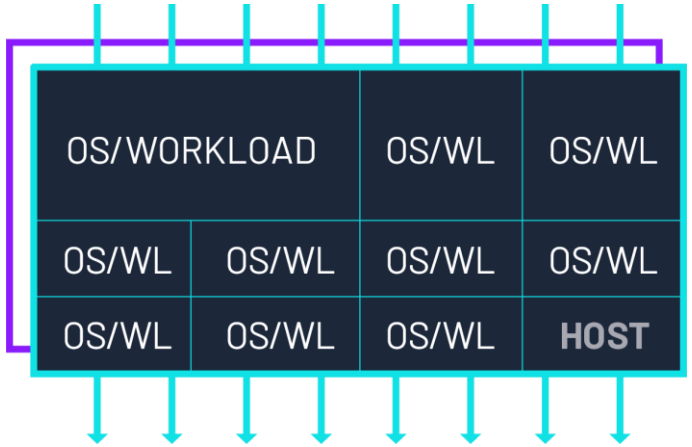
3D Graphics

Compute

Housekeeping

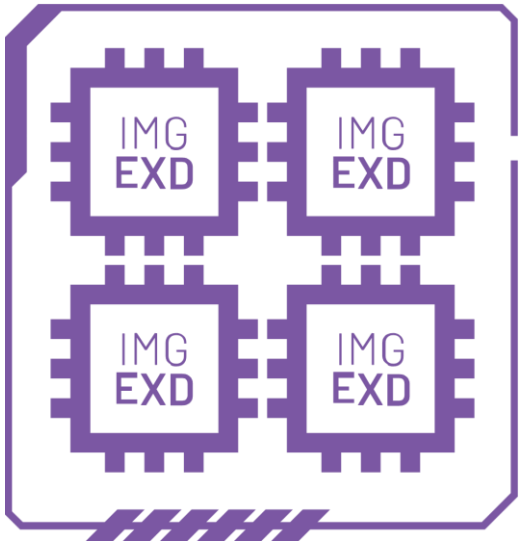
Concurrent processing of different task types

High Performance **Virtualisation**



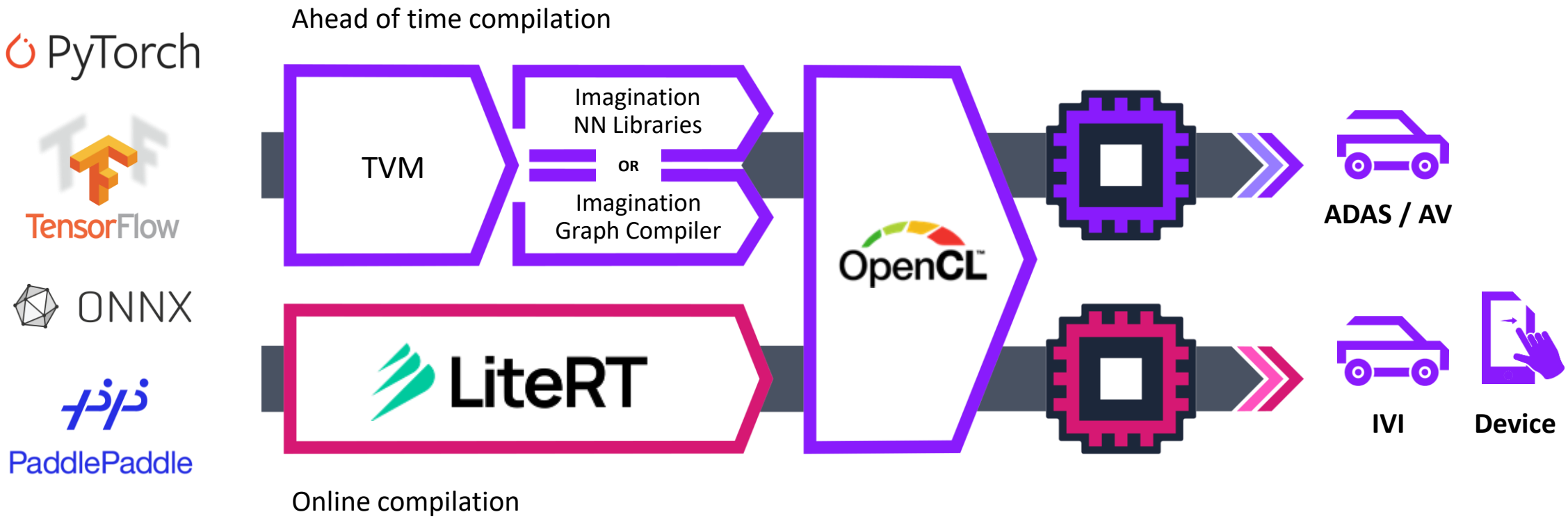
Up to 16 virtual machines supported in hardware with QoS – double D-Series capabilities.

Multi-core **Scaling**



Scaling from 1 to 4 cores for extra flexibility and performance

Clear Paths for Porting Models



E-Series is Designed for Developers

BROAD OS & API COVERAGE

Linux, Android & Windows
Vulkan, OpenGL ES, OpenCL, DirectX

IMAGINATION DEVELOPER HUB

Documentation & Tools
Software Development Kit
Demos & Sample Code

ADVANCED TOOLING

PVRStudio – IDE & Debugging Environments
PVRCarbon – Frame Capture & Analysis
PVRTune – Real-time performance analysis

OPEN SOURCE DRIVER PROGRAMME

Designed from the ground up
for open source

WORKING WITH OPEN PROJECTS:

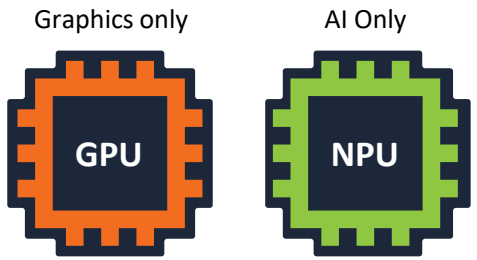
e.g. The UXL Foundation,
ONNX, TVM, LiteRT.

GAME OPTIMIZATIONS

Game engine optimizations, manual & automated QA, ecosystem development.

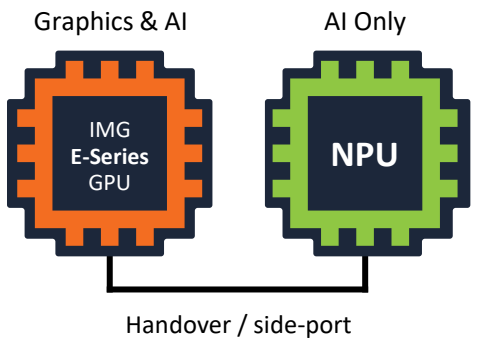
E-Series – Unlocking Flexible, Cost Optimised AI in SoCs

Traditional Approach



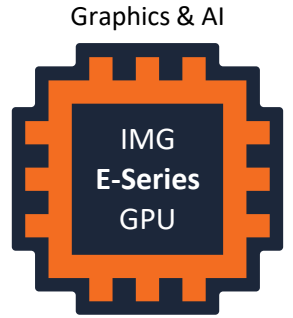
Simple:
Very simple design but expensive and high risk of not supporting future AI networks

GPU Provides Future Proof AI



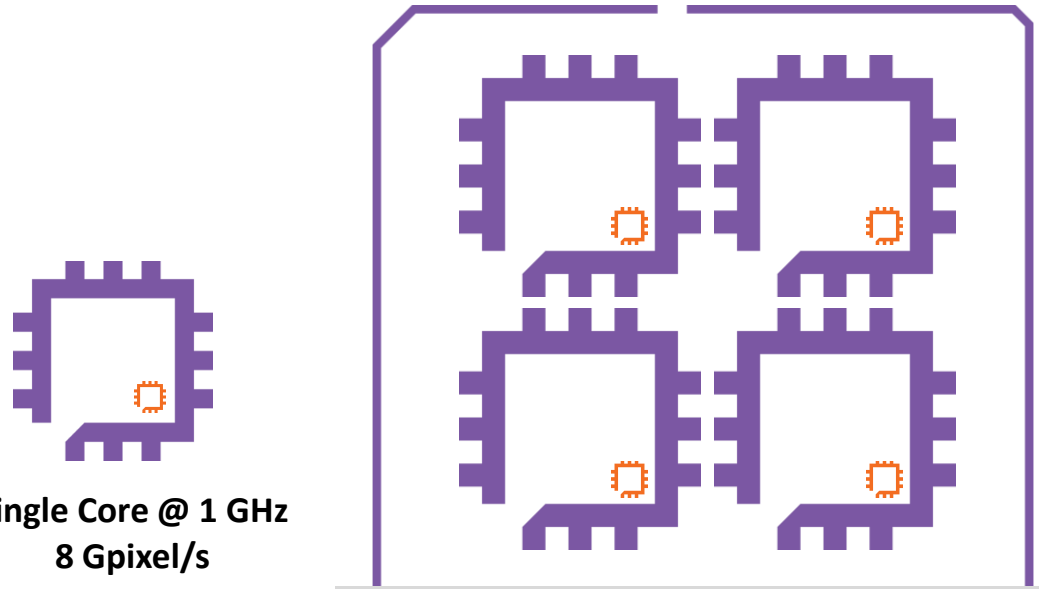
Balanced:
High performance AI runs on NPU while GPU provides flexible support for future (unknown) AI networks

Larger GPU Provides Flexible AI



Cost Optimised:
Additional GPU performance provides good, flexible support for future (unknown) AI networks

Imagination E-Series GPU IP Range Covers Every Market and Device



Single Core @ 1 GHz
8 Gpixel/s
0.25 TFLOPS FP32
2 TOPS Neural Core

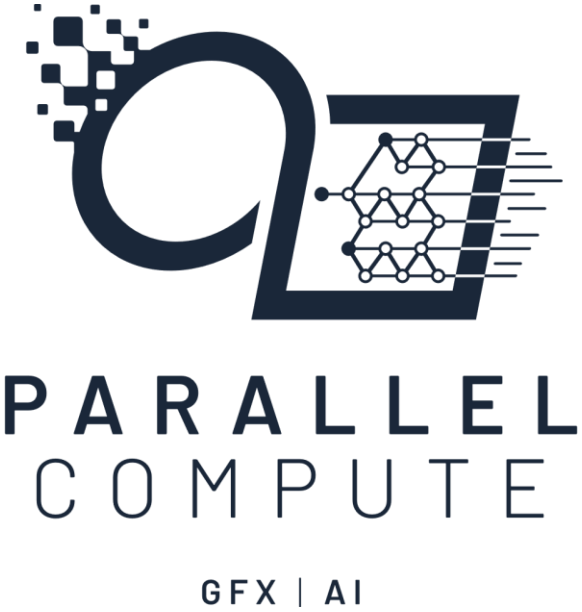
Quad core @1.6 GHz | 400 Gpixel/s
12.8 TFLOPS FP32 | 200 TOPS Neural Core

Scales from 2 to 200 TOPS.
First configurations arriving Autumn 2025

- AI PC & Data centre
- Mobile & Consumer
- Automotive
- Industrial

Summary

- **AI a fundamental change in software**
  - Hardware needs to change to match it
- **Need flexible, generalised parallel compute hardware**
  - Along with understood programming model, libs and tools
- **Imagination E-series delivers efficient graphics and AI**
- **Visit us:**
  - Booth #908
  - **Imaginationtech.com** for more information



THANK YOU