



AXELERA
ARTIFICIAL INTELLIGENCE

Simplifying & Accelerating AI at the Edge

Fabrizio Del Maffeo

CEO & Co-Founder

September 2023

fabrizio.delmaffeo@axelera.ai

Our opportunity is AI for edge devices

AI is expanding from cloud computers to edge devices

Existing technology does not fit market needs

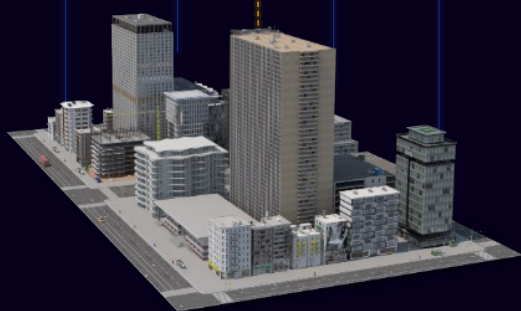
Market is highly fragmented, and incumbents keep focusing on cloud computing, automotive & consumer electronic

40+ billion connected things

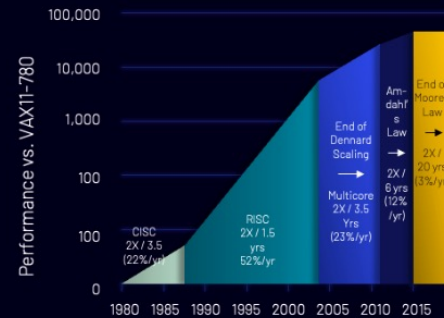


54% of data stored at the edge

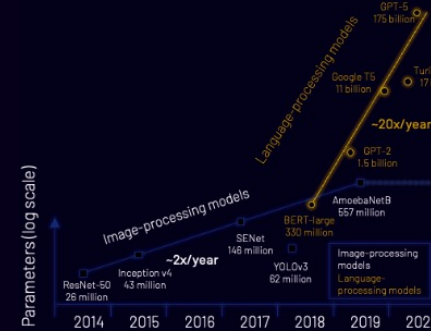
82% of data on internet is images and video



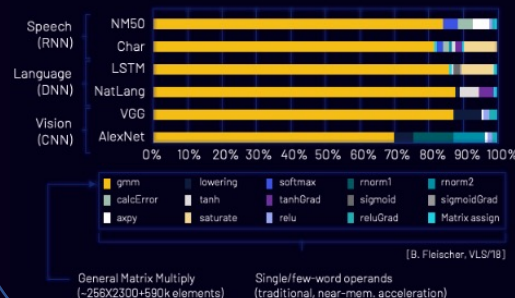
End of exponential growth in computing performance



Neural Network size increasing exponentially



Computer technology too much general purpose, not optimized for AI workloads

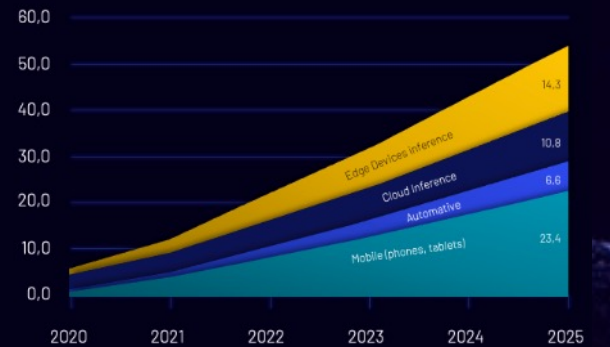


Customers need more than just efficiency or throughput



Excluding Automotive, Consumer and Cloud Computing Inference, all the other market domains are highly fragmented with thousand small/medium customers

AI Chipset for Inference - Revenue Forecast* \$B



- Mobile (phones, tablets)
- Cloud Inference
- Automotive
- Edge Devices Inference



Retail



Security



Robotics &
Drones



Industry 4.0



Healthcare



Automotive

A powerful, efficient, cost-effective and user-friendly **hardware and software AI platform** to accelerate **computer vision and generative AI** in edge devices

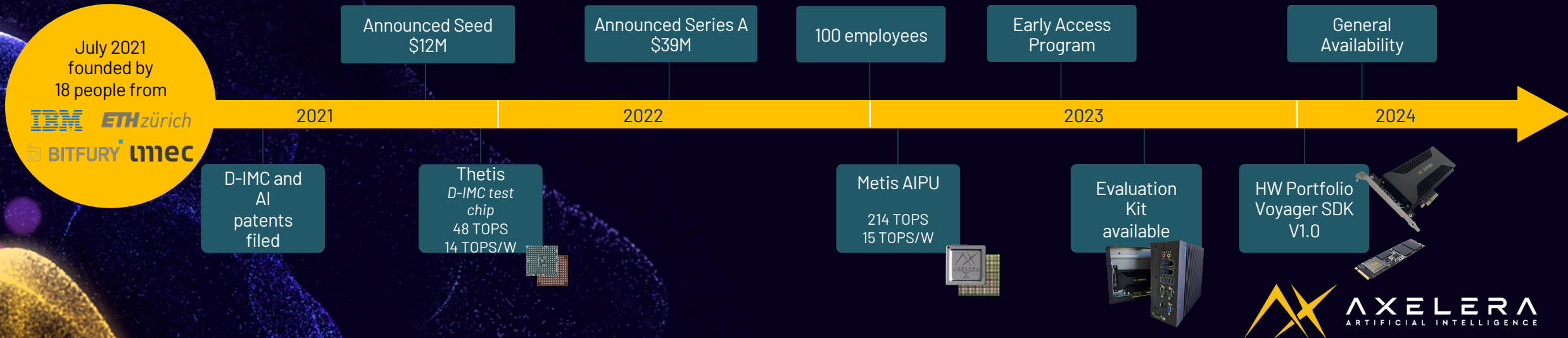
Axelera AI in a nutshell

- Established in July 2021, now 130+ employees, 50+ PhDs in AI, compiler and VLSI design
- HQ in Eindhoven (NL), main offices in Zürich (CH), Leuven (BE), Bristol (UK), Florence (IT)

US operations from August 2023

- A breakthrough in *Digital In-Memory Computing (D-IMC) & RISC-V* for AI processing
- Started shipping Metis, our ultra-efficient cost-effective 214 TOPS Edge AI accelerator

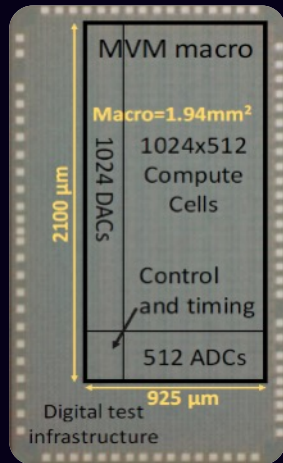
Axelera experts come from



Axelera AI Expertise – Analog In-memory Computing

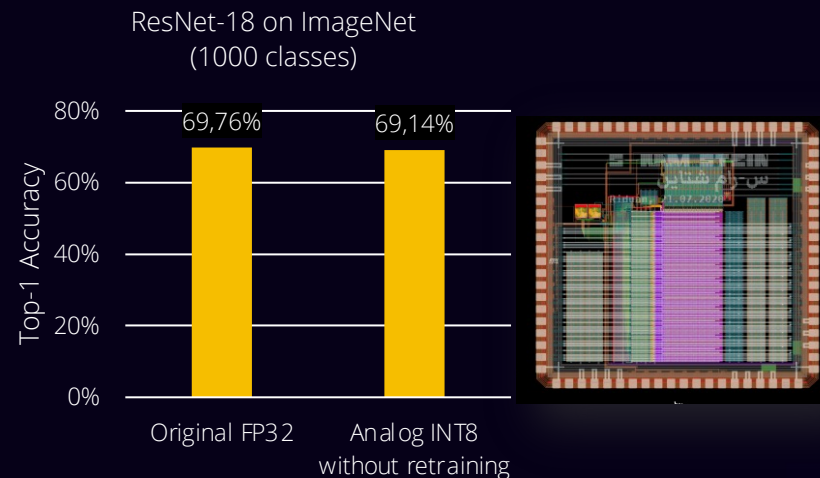
AnIA chip (2019), 22nm –
Maximum **Efficiency**

- 1540 TOPS/W
- 1024 pulse width modulating DACs with 6b + sign input
- 512 SAR ADCs with 6b output
- 1024x512 ternary compute with (-1,0,1) weights



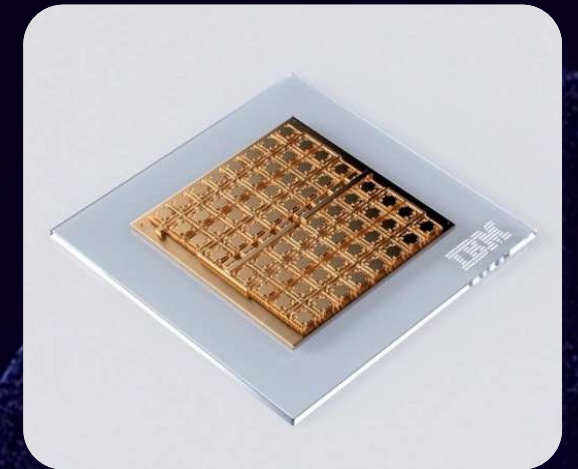
SRAMstein chip (2020), 65nm –
High **Accuracy**

- Accuracy loss with retraining ~ 0.2-0.3%
- Precision scales linearly with time, power & area
- 64 pulse width modulating DACs with 8b + sign input
- 256 SAR ADCs with 8b output
- 64x256 INT8 compute with INT8 weights



*Our team members are co-authors of
IBM HERMES chip (2021), 14nm*

- Throughput: 63.1 TOPS, Energy efficiency: 9.76 TOPS/W
- 64 cores, each core w/ 256 x 256 PCM crossbar array
- 8b input / 4 bit effective weight precision / 8b output



Axelera AI Expertise – RISC-V®

- Strong RISC-V (SoC) expertise in-house:
 - Design of CVA6 (Ariane) and Snitch: Most popular SystemVerilog RISC-V Cores in GitHub
 - Several tape-outs in academia and industry
 - Several high multi-core tape-outs (incl. by Intel)

- RISC-V Vectors (RVV)

Original design of **Ara, Vicuna and Spatz**

- RISC-V cores

Six instances integrated in Metis AIPU SoC:

- **CVA6** is used by:
 - Google, Thales Group, Hensoldt Cyber, Tortuga Logic, CEA/Leti, European Processor Initiative (EPI), GlobalFoundries, Princeton University, Columbia University, IBM, Columbia, Harvard, ongoing DARPA project(s), Uppsala universitet, TU Graz, IntelSelectiv



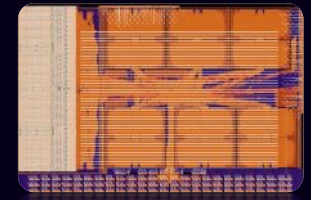
Smallest Linux-booting
RISC-V system



26x RISC-V, 22nm GF



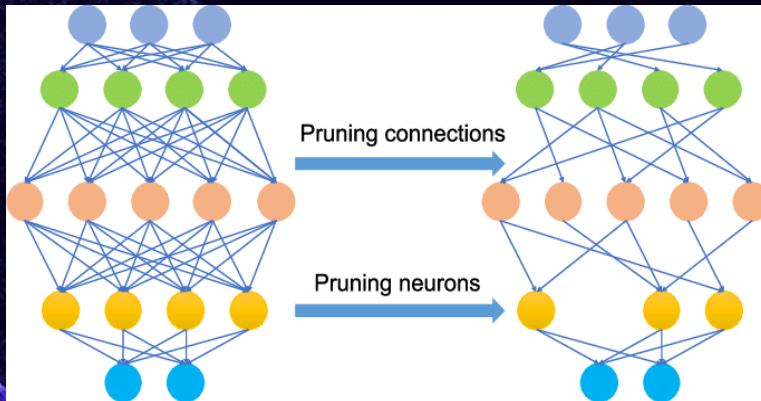
216 RISC-V Cores



Axelera AI Expertise: Machine Learning

DiffPruning: Differentiable Transport Pruning

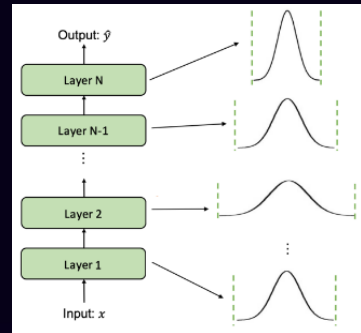
- ~40% reduction in model storage size



Method	Backbone	Base top-1%	Pruned top-1%	Top-1 drop%	Speedup
DiffPrune (Axelera AI)	ResNet-50	76.13	74.26	1.87	3.06x

Grad-PTQ

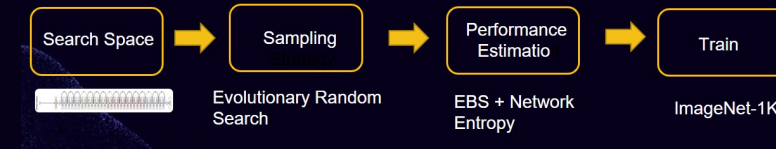
- Computationally inexpensive SGD-based fine tuning of quantization parameters
- Always better than PTQ, at times better than even FP original model



MobilNet v1	Accuracy[%]
Original (FP32)	70.34
PTQ (INT8)	69.03
Grad-PTQ (INT8)	70.45

Network Architecture Search: AxeleraNet V0.1

- similar latency as ResNet-18, but 5% better accuracy
- 36% faster latency, and 1.5% better accuracy, compared to ResNet-34
- Interestingly, AxeleraNet v0.1 has 96 layers compared to 18/34 layers for ResNet-18/34



Network	Accuracy (%)	Latency (ms)
ResNet-18	69.8	0.62
ResNet-34	73.3	1.03
AxeleraNet v0.1	74.8	0.66

A man in a dark blue suit and light blue shirt is smiling and holding a small, black, rectangular AI processing unit with both hands. The unit features a small fan in the center, the Axelera logo (a stylized 'X' with a yellow and orange gradient), and the text 'AXELERA ARTIFICIAL INTELLIGENCE'. The background is a blurred outdoor scene with green grass, a body of water, and modern buildings under a clear blue sky.

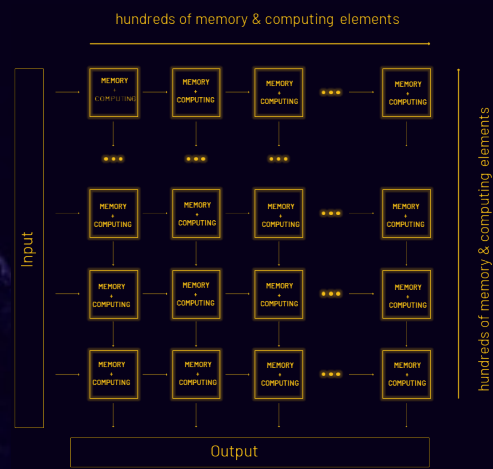
Introducing
METIS AI Processing Unit



Axelera AI – an AI-Native Computing Architecture

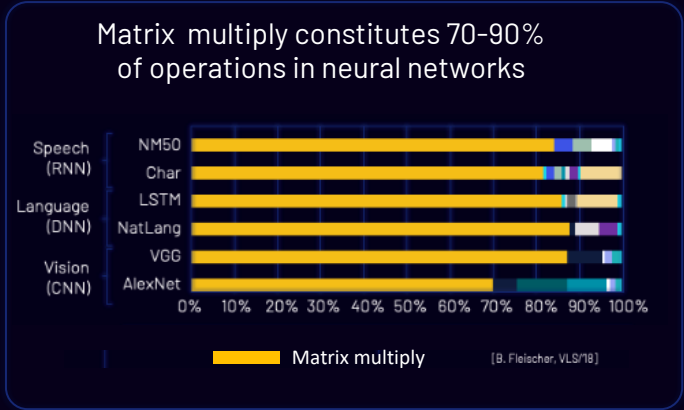
Digital In-Memory Computing (D-IMC)

Densely interleaved memory and compute elements



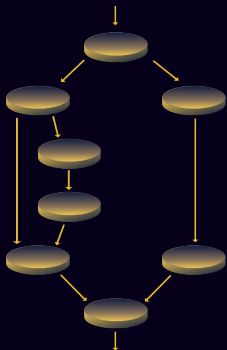
Application-first approach

Balance performance, programming efficiency and silicon cost



Precise scheduling of memory and graph processing ahead of time

General operations and data flow



Massively parallel execution
= **High Throughput**

Weight cells and compute merge
= **Small Die Area**

Local data & low data movement
= **Power-efficient**

FP32 and INT8-32 data paths
= **FP32 Level Accuracy**



End-to-end optimization
= **Real-world performance**

Open-source AI toolchain
= **Reduce risk**

Any activation function supported
= **Future-proof**

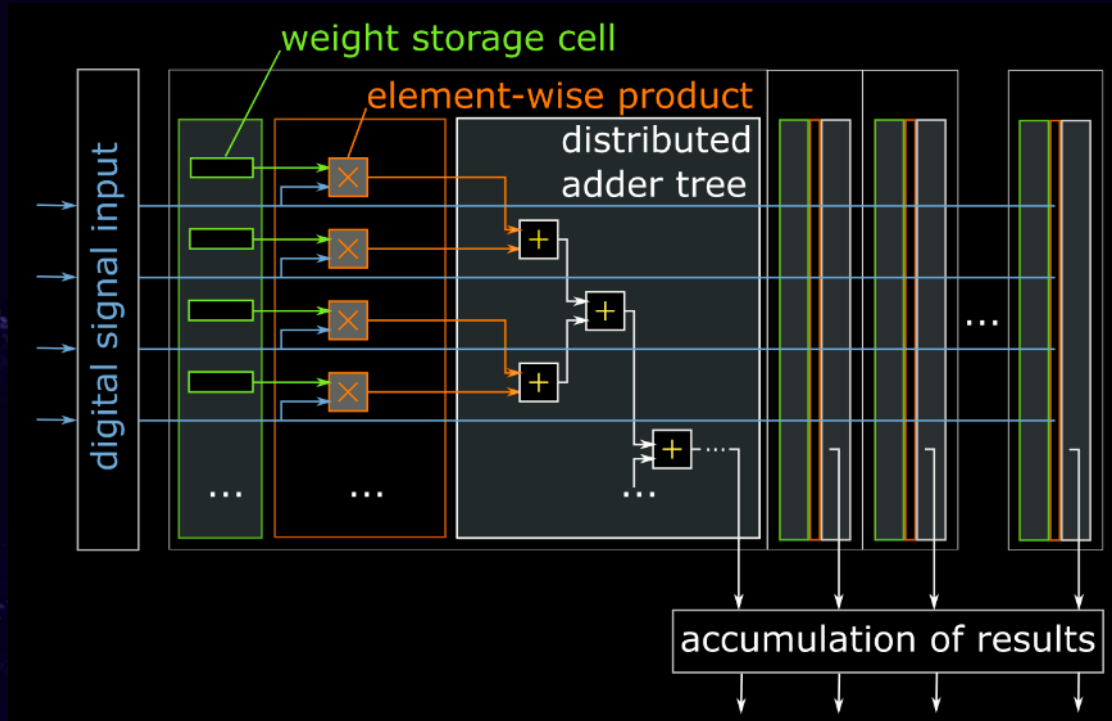
Data prefetch keeps AI cores busy
= **High Performance**

Highly programmable
= **Usability**

Scalable architecture
= **Protect investment**

Digital In-Memory Computing: how it works

Digital In-Memory Computing



- **SRAM-based Digital In-Memory Computing engine**
- Immune to noise and memory non-idealities that affect precision of analog IMC
- Interleaved weight-storage and compute units in an extremely dense fashion
- Multiple weight-sets allowing simultaneous prefetching and processing
- INT8 activations and weights, with INT32 accumulation to maintain full precision
- Technology commensurate with CMOS scaling to any low lithography node

The Digital IMC engine is a handcrafted, full-custom design

Introducing Metis AI Platform – a big leap forward for Edge AI capabilities



Breakthrough performance

- 3,000+ fps ResNet50@ 99.9% FP32 accuracy
- 5,000+ fps SSD-MobileNet V1@ 99.7% FP32 accuracy
- Ultra-low latency for real-time responsiveness
- Simultaneous multi-model & multi-stream



Excellent usability

- Fully programmable, AI acceleration platform
- Easy build and optimization of complete pipelines
- No costly retraining necessary



Exceptional power efficiency

- 4-7W typical power consumption
- > 10X power efficiency vs. leading GPUs
- Up to 15 TOPS/W core energy efficiency



Superior value

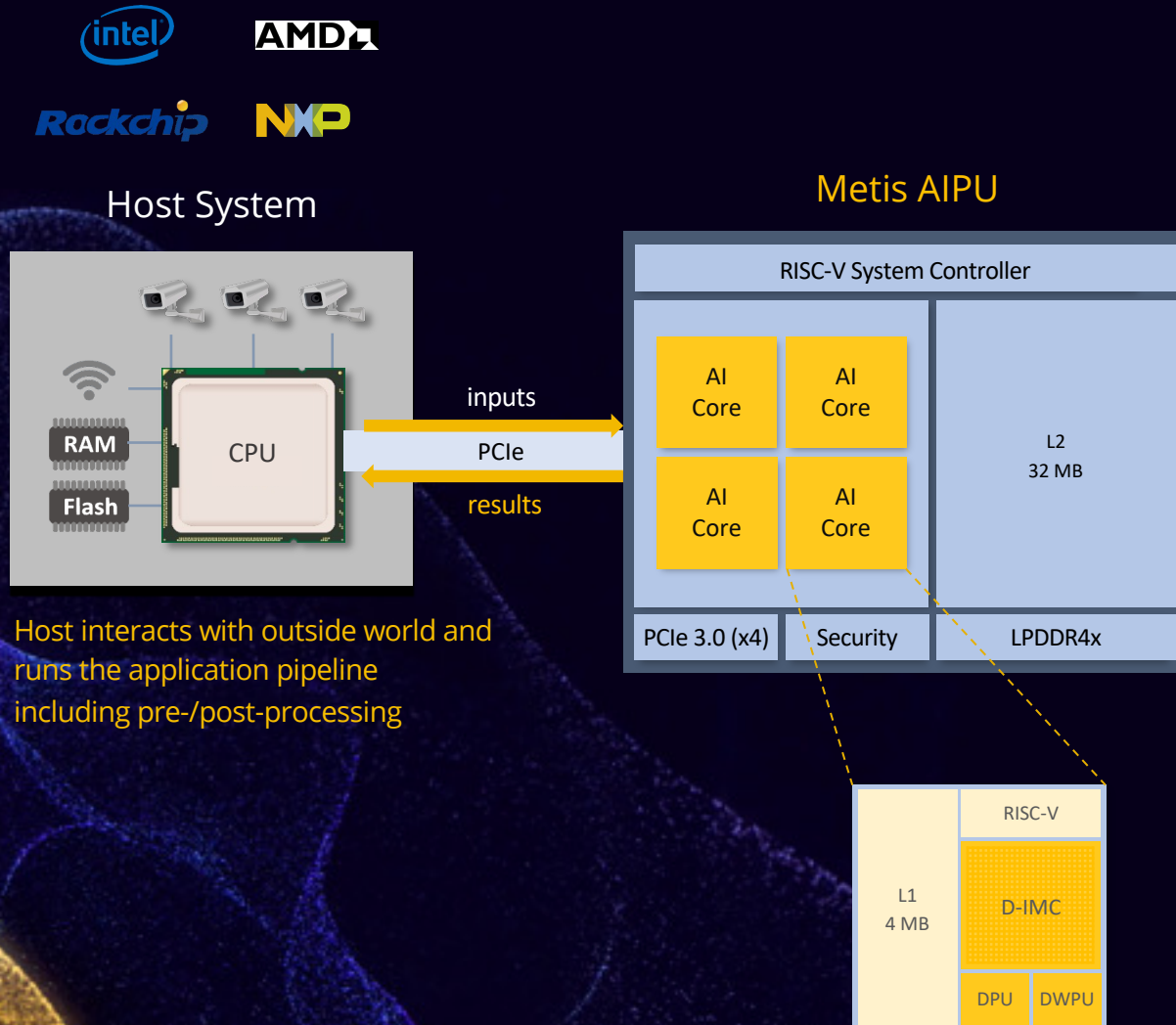
- Industry's lowest cost per AI stream & inference
- Single-chip M.2 to high-performance PCIe
- Small silicon die = competitive pricing

VOYAGER SX
Tools Suite



Metis
AI Processing Unit

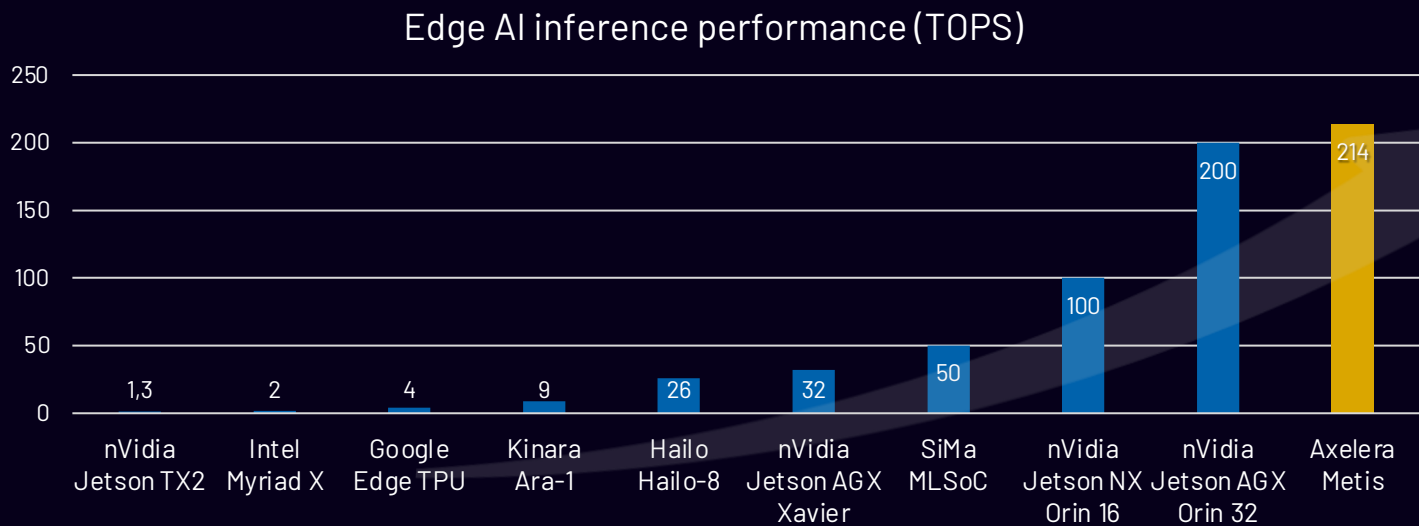
Metis AIPU – system block diagram



- **AI Engine with four AI Cores**
 - D-IMC : matrix-vector multiply, densely interleaved memory and compute with >1M elements
 - DPU : element-wise vector operations and activations
 - DWPU : sparse ops; pooling, depth-wise convolution, up-sampling etc.
 - AI Cores run independently or in concert
- **Hierarchical compiler-managed memory**
 - Execute-overlapping prefetch and data reuse allocation
 - D-IMC : 4 x 1 MB
 - L1 AI Cores : 4 x 4MB
 - L2 Shared : 32 MB
 - DDR : Up to 4 GB to fit large models, minimize latency and context switching overhead
- **Security unit**
 - Secure key storage and management, encryption and decryption
- **Performance and physical data**
 - 214 tera-operations-per-second (TOPS)
 - 4-7W typical power consumption
 - 12 nm CMOS technology
 - -40°C to +85°C extended temperature range

Metis AIPU – performance and usability at a fraction of the price and power consumption

Given power and cost constraints, AI's appetite for compute at the edge is... big



START
1-5 TOPS

Simple, few-layer CNNs

Low accuracy

YESTERDAY
5-50 TOPS

More complex CNNs
and pipelines

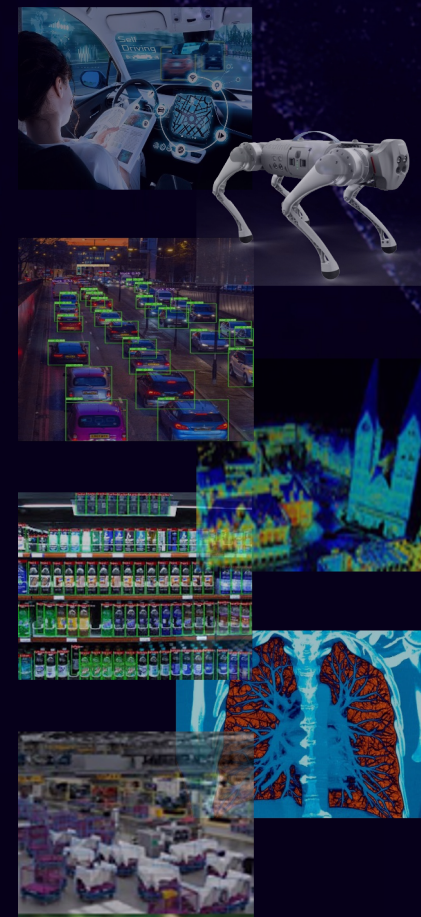
Medium accuracy

FOR NEW DESIGNS TODAY
>100 TOPS

**Transformers, LLMs, generative,
increasing resolutions,
ultra-low latency**

High accuracy

* TOPS numbers as reported in official datasheets. nVidia numbers reported as 'Sparse TOPS' (2x 'Dense TOPS')

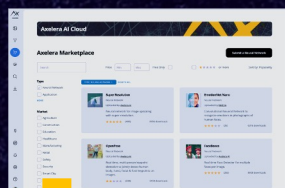


Voyager SDK: streamlined development, yet open and flexible for experts

Application Development

Model Zoo

Ready-to-use models and application templates



Low-code Pipeline Development Kit

Declarative definition (YAML) of pipeline with pre-/post-processing and multiple CNN steps

```
pipeline:
  preprocess:
    - resize:
        width: 3024
        height: 3024
    - convert-channels:
        mean: 104, 117, 123
        std: 1, 1, 1
  postprocess:
    - faceboxes:
        dimension: "1024:1024"
        thr: 0.5
        conf: 0.7
  - inception:
    preprocess:
      - convert-channels:
          dimension: "128:128"
          totensor:
            postprocess:
              - landmarks:
                  mode: "faceboxes"
                  dimension: "160:160"
                  totensor:
                    postprocess:
                      - dynamicnormalize:
                          mode: "faceboxes"
                          file: "/home/ubuntu/embeddings.json"
                          thr: 0.8
```

No-code Cloud Application Composer

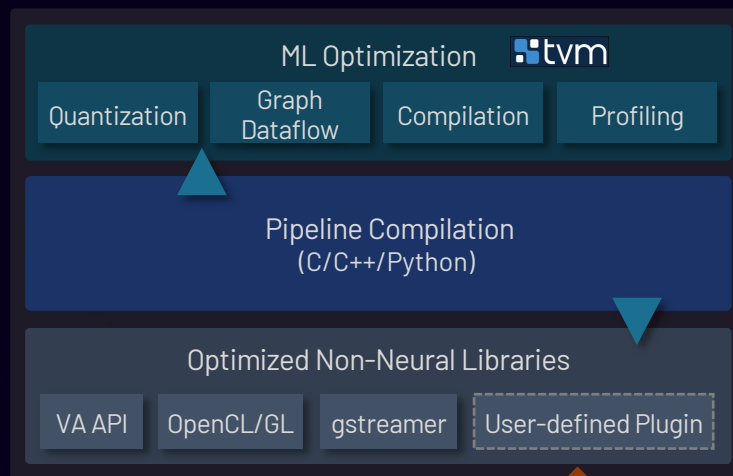
Drag and drop application pipeline building

Custom Pre-trained Models



Optimization

Comprehensive framework for ML optimization and AI application optimization



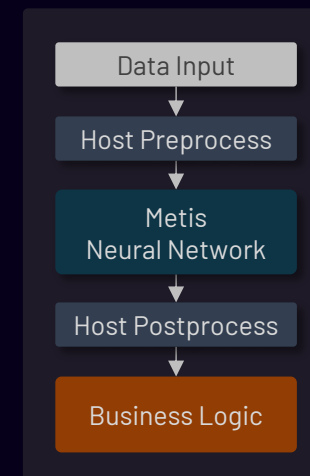
User Legacy Code



Deployment

Application Pipeline

Inference seamlessly integrated with the business logic application



Package & Deploy



Metis AIPU evaluation Kit for Early Access Program (EAP) partners - Shipping Now!

VOYAGER.SDK



Shipping now
to qualified early access
program partners

Metis Evaluation Kit	
Host system	Advantech MIC-770V3 Compact Fanless Industrial PC Intel Core i5-12500TE Size 77 x 192 x 230 mm + PCIe expansion box
AI acceleration	Axelera 1x Metis 214 TOPS w/4x GB LPPDR4x PCIe card
Video source	HD IP Camera, AXIS ME2035-LE
OS	Linux Ubuntu 20.04 / 22.04
SDK	Voyager SDK
ML framework	PyTorch
NN models included in Eval Kit	Image Classification: SqueezeNet, ResNet-18, ResNet-34, ResNet-50 Object Detection: SSD-MobileNet-v1, YOLOv5s/m/l Segmentation: Mask-R-CNN, U-net
Support	Direct engineering support to assist performance evaluation, pipeline porting and custom ML requirements

Metis AIPU - A complete product portfolio of high-performance cost-effective accelerators

Partner System Hardware

AI Vision Gateway



SECO PHOENIX

Compact embedded
160 TOPS AI

AI Vision Edge Server



Advantech MIC-770V3

Intel i7 Industrial PC
856 TOPS AI

AI Vision Appliance



Dell PowerEdge XR4000

2U Intel Xeon server
1.7 PetaOPS AI

Axelera AI Modules and Cards



Metis M.2 2280M

Price: From 159 USD



Metis PCIe x4

Price: From 209 USD



High Performance 4x Metis PCIe x16

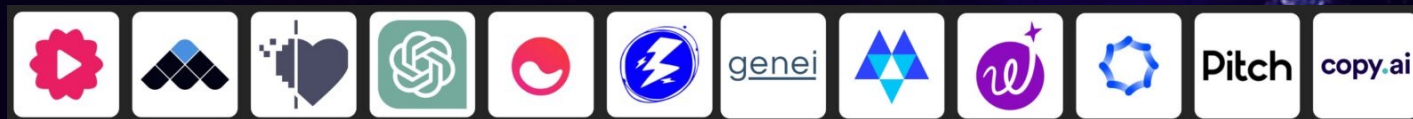
Price: From 549 USD

Axelera AI's advantage: Performance/\$ vs Current Technology



Expected Customer **Savings** for 1K units:
USD 2.5 – 3.0 Million CAPEX

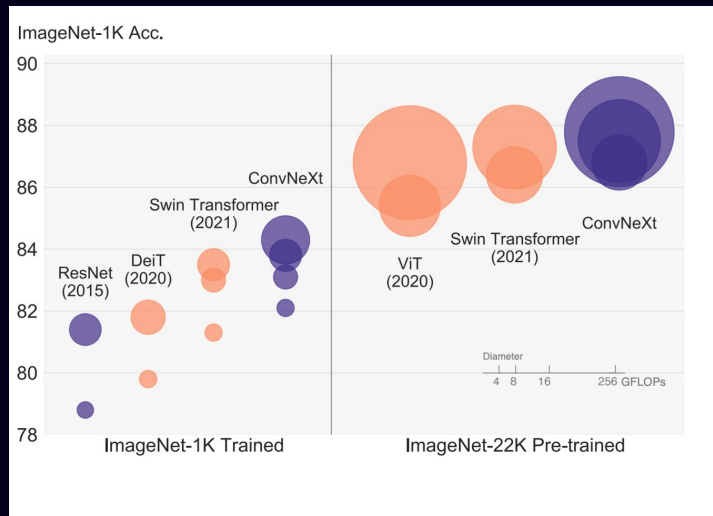
What's next ?



Observation 1: neural networks trends

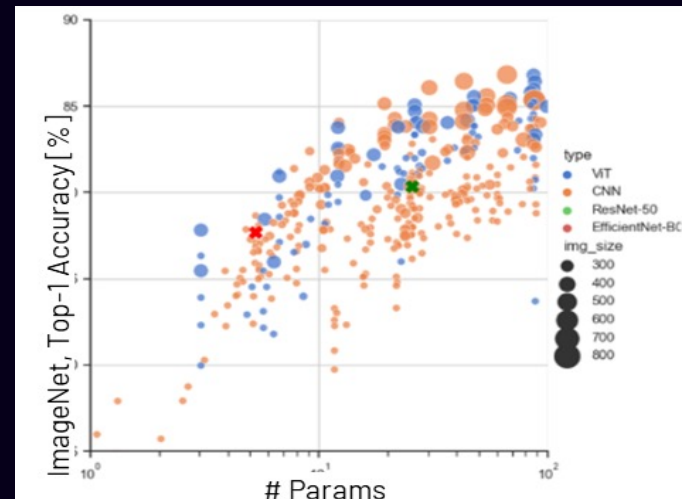
Larger CNNs on larger frames

- Evolution of CNNs seems to have stalled, ConvNeXt being a peak performer
- Further gains likely to be had from topology changes, and operating on larger frames



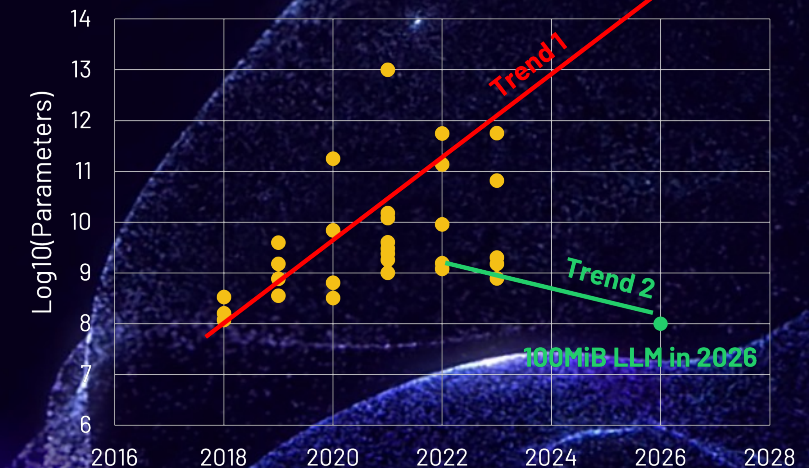
Vision Transformers

- Important trend, likely to continue and become more dominant through 2026
- No clear breakthrough/benefits in terms of accuracy and parameters compared to CNNs



Generative AI and LLM

- Trend 1 (ChatGPT/Dall-E etc.) : towards larger models, cannot be covered by Edge device
- Trend 2: smaller mobile models can be intercepted by Edge devices



Observation 2: the importance of memory bandwidth for accelerating LLMs

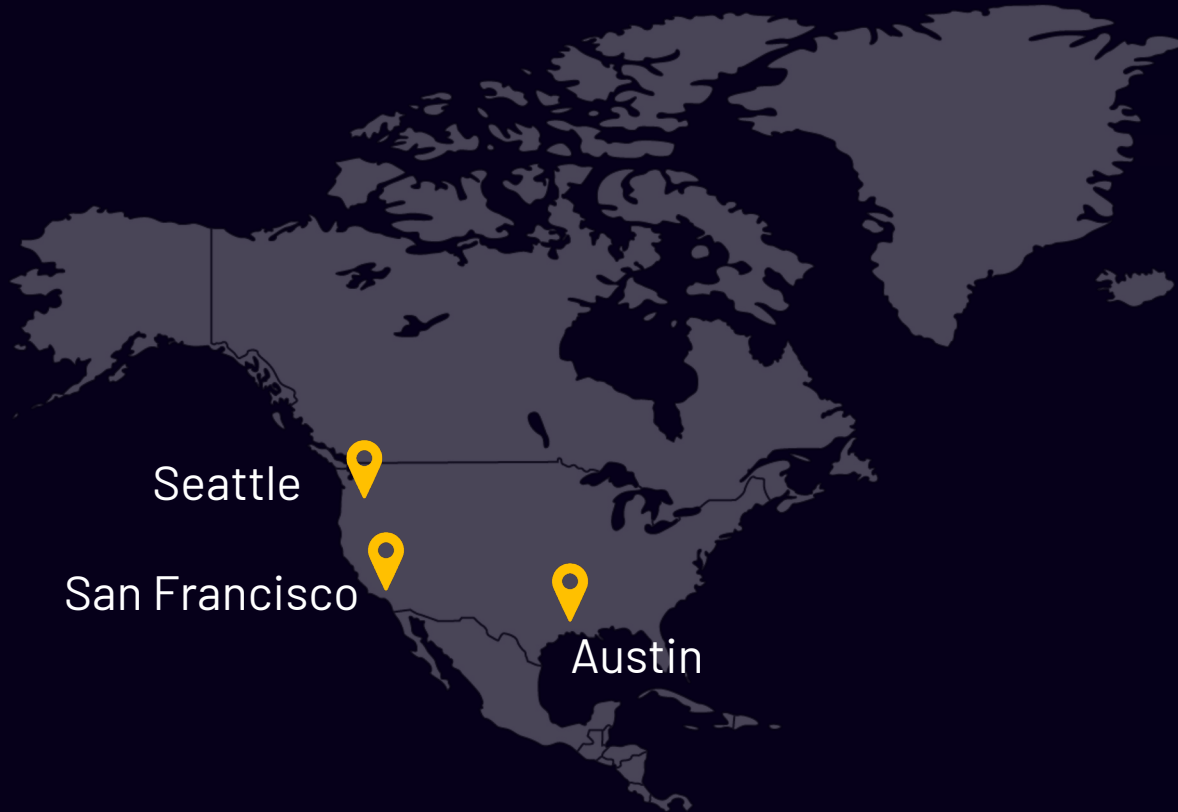
Network	S	E	L	Model [GiB]	Model + Cache [GiB]	GOP / token	100 eTOPS 4xLPDDR4x	100 eTOPS 8xLPDDR5
LLM-7B	64	4096	32	6.44	6.5	26	5.2	15.4
	512	4096	32	6.44	6.7	26.5	5.1	14.9
	2048	4096	32	6.44	7.5	28.5	4.5	13.3
Large Proposal	64	2048	32	1.6	1.62	6.5	21.0	61.7
	512	2048	32	1.6	1.74	6.8	19.5	57.5
	2048	2048	32	1.6	2.15	7.8	15.8	46.5
Medium Proposal	64	1024	32	0.4	0.411	1.6	82.7	243.3
	512	1024	32	0.4	0.47	1.8	72.3	212.8
	2048	1024	32	0.4	0.67	2.3	50.7	149.3
Tiny Proposal	64	512	16	0.05	0.052	0.2	653.8	1923.1
	512	512	16	0.05	0.067	0.25	507.5	1492.5
	2048	512	16	0.05	0.117	0.37	290.6	854.7

S = Sequence Length
E = Embedded Dimension
L = number of layers in the encoder and decoder

Table 1 – Estimating throughput in terms of tokens/s for a variety of models and for two different platforms: an edge machine with 100 effective TOPS (terra-operations per second) of compute capabilities and either 4 channels of LPDDR4x (+/- 34GiB/s) or 8 channels of LPDDR5 (+/- 100GiB/s). Inferring networks like LLM-7B for large sequence lengths is infeasible on such devices, network designers should focus on developing better, smaller models in order to increase throughput on such Edge devices. S is the sequence length, E is the embedding dimension (E2=4xE1) and L is the number of layers in the encoder and decoder.

BLOG: <https://www.axelera.ai/decoding-transformers-on-edge-devices/>

We are expanding our operations in the US!



Hiring Actively across roles in 2023-24

- General Manager and President Americas
- Regional Sales Director
- Alliance and Ecosystem Director
- Business Development Executive
- Key Account Manager
- Field Application Engineer



Disclaimer

Presentation Disclaimer

This presentation contains information and data that is subject to change without notice. While every effort has been made to ensure the accuracy and completeness of the information included, the presenter and their associated parties make no guarantees or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability with respect to the presentation or the information, products, services, or related graphics contained herein for any purpose. Any reliance you place on such information is therefore strictly at your own risk.

Forward-Looking Statements

Some information in this presentation may contain forward-looking statements that involve risks and uncertainties. Actual results may differ materially from those expressed or implied.

Third-Party Trademarks and Logos

All product names, logos, and brands are property of their respective owners. All company, product and service names used in this presentation are for identification purposes only. Use of these names, logos, and brands does not imply endorsement. Specifically, the logos of IBM, Intel, AMD, NXP, Meta, OpenAI, Google, TensorFlow, PyTorch, ChatGPT, Llama and Rockchip are trademarks of their respective owners, and their inclusion in this presentation does not imply any affiliation with or endorsement by them.

Intellectual Property

All original content in this presentation, including but not limited to text, graphics, and code, are the intellectual property of the presenter and may not be used, reproduced, or distributed without explicit permission.

Limitation of Liability

In no event will the presenter or their associated parties be liable for any loss or damage including, without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use or interpretation of the information in this presentation.



Thank You!