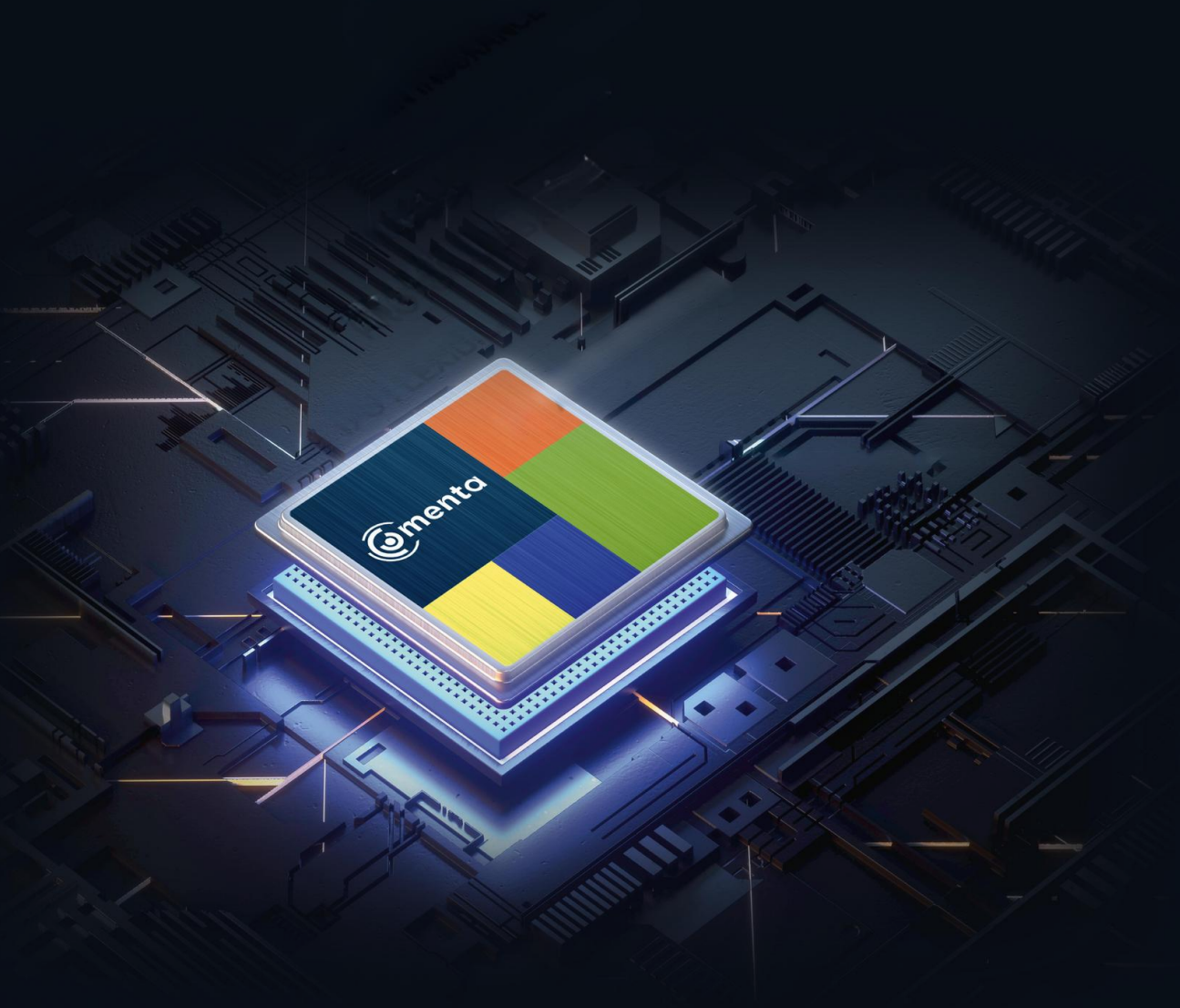


menta
Unlimited Power
of Technology

Adaptable Compute Architectures for Physical AI Systems

Because the only constant is change

Jayson Bethurem
VP, Business Development

Intelligence without adaptability quickly becomes obsolete

Physical AI requires hardware that can evolve as fast as the world around it

AI Algorithms Evolve Faster than ASIC Lifecycles

Physical AI systems are deployed for years or decades – Vehicles, Satellites, Industrial Equipment, Defense Systems

AI models evolve constantly! New neural network architectures, tensor formats, quantization models, inference techniques

Sensor Fusion Continuously Changes

Physical AI combines multiple sensor types including radar, vision, ultrasonic, thermal, IMUs, & more

Sensor fusion pipelines continuously evolve, and critical systems such as autonomous driving, robotics, aerospace & defense depend on the ability to adapt and optimize these processing architectures over time

Deterministic Low-Latency Processing

Cloud AI tolerates latency. Physical AI cannot. Real-time systems like collision avoidance, beamforming, motor control, drone stabilization, industrial safety, and financial trading require deterministic low-latency processing where delays directly impact performance, safety, and outcomes.

Workloads are Still Undefined

Physical AI is still in its early stages, and no one fully knows what the optimal AI architectures, sensor pipelines, interconnects, standards, or future safety requirements will ultimately look like. This uncertainty makes adaptable hardware critical, enabling intelligent systems to evolve and optimize long after deployment.

FPGA Value

Adaptable HW allows post-deployment HW updates, AI accelerator evolution, algorithm refinement & differentiation

Adaptable HW enables new filtering pipelines, updated synchronization logic, changing protocols, improved preprocessing and optimized dataflow architectures. Increases Market SAM

Adaptable HW provides deterministic HW execution, parallel processing, ultra-low latency and real-time response

Adaptable hardware reduces that risk by allowing systems to evolve after deployment.

The only constant is change!

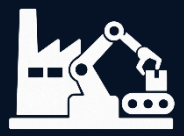
menta
Unlimited Power of Technology

Where Physical AI interacts with the Real World

Sensing, deciding and acting locally with deterministic latency, safety & reliability

Autonomous Vehicles & Transportation

Cars, trucks, robotaxis, rail, drones, shipping, logistics robots, autonomous mining and agriculture vehicles



Industrial Automation & Robotics

Factory robotics, cobots, PLC/PAC systems, machine vision, predictive maintenance, warehouse automation & smart manufacturing

Aerospace & Defense

Autonomous drones, EW systems, radar, sensor fusion, secure communications, missile guidance, battlefield AI & edge autonomy

Agriculture & Precision Farming

Autonomous tractors, crop monitoring drones, smart irrigation, livestock monitoring and robotic harvesting

Edge AI Consumer Devices

AR/VR, smart appliances, drones, AI cameras, personal assistants and adaptive consumer electronics

Humanoid & Service Robotics

Human-assist robots, healthcare robots, elder care, hospitality, delivery robots and domestic automation

Smart Infrastructure & Smart Cities

Traffic systems, surveillance, energy optimization, transportation coordination, public safety and environmental monitoring

Healthcare & Medical Devices

Surgical robotics, AI imaging systems, wearable monitoring, prosthetics, rehabilitation systems and autonomous diagnostics

Energy & Utilities

Smart grid control, autonomous inspection, power optimization, oil & gas monitoring and renewable energy balancing

Financial & Mission-Critical Edge Systems

High-frequency trading acceleration, real-time risk analysis, telecom infrastructure and ultra-low latency networking

In the cloud, AI generates tokens… In the real world, AI generates insurance claims!

Why Physical AI on Adaptable Hardware?

Maximize efficient and adaptability for constantly evolving algorithms

Hardware Tailored to the Algorithm, not the other way around

Fixed AI accelerators leave performance on the table. Menta eFPGA fabric delivers efficient implementations for operations like convolution and pooling, with deeply pipelined logic, DSP blocks, and shift registers arranged to maximize throughput. This flexibility enables highly optimized neural-network execution—especially valuable as algorithms evolve.

Tightly Coupled Memory Maximizes Performance

For AI designs, dataflow is critical. Tightly-coupled foundry memory minimizes slower & non-deterministic off-chip memory accesses. Use as streaming buffers to avoid stalls and accelerate computation

Match Precision to your Accuracy Requirements

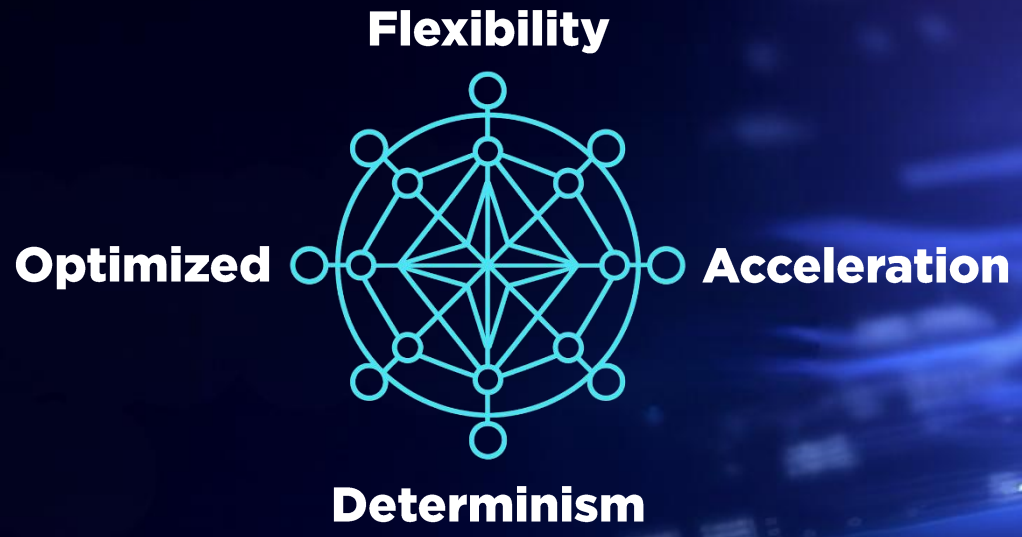
FPGAs excel at low-precision math. Instead of expensive floating-point, quantize data to 8-bit down-to binary formats with mixed precision, per-layer scaling and saturation-aware rounding to preserve accuracy where it matters most

Deterministic Processing

Real-time applications require deterministic paths, tight loop bounds and precise memory scheduling – down to the clock cycle

Every Gate Counts

ASIC area is critical – Menta's adaptable hardware enables you to optimize your design for speed, power, accuracy and area!



Architectural Tradeoffs

Hardware	Cost	Flexibility	Performance	Power
CPU	Low	High	Low	High
GPU	High	Medium	High	Very High
ASIC	Very Low	Very Low	Very High	Very Low
FPGA	High	High	High	Medium
ASIC w/ eFPGA	Low	High	High	Low

Efficient FPGA AI QNN AI Implementation

Highly Optimized and Deterministic AI/ML Inference for the Edge

Efficient AI algorithm execution requires

Processing architectures that exactly match the algorithm

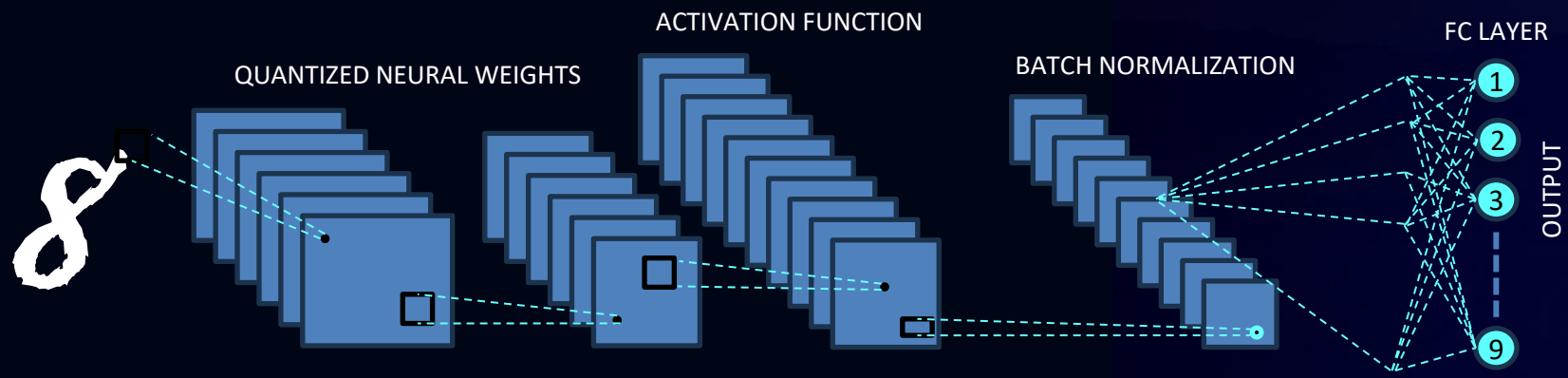
Effective software tool implementation into said architecture

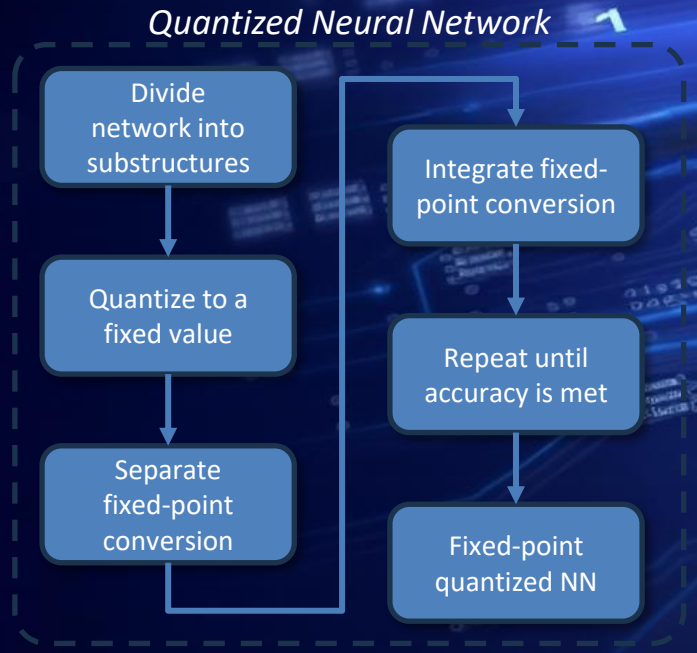
Architectural flexibility architecture to adapt to novel algorithms

FPGA's flexible architectures can adapt to any network

Quantized Neural Networks map efficiently to FPGAs

- Ultra low-latency with high performance & very low memory footprint
- New dataflow compilers are available logic gate networks





Where FPGAs Outperform GPUs in AI

FPGAs can outperform GPUs on latency, determinism, power efficiency, or streaming performance

Small Quantized CNNs

MobileNet
SqueezeNet
Tiny-YOLO
Custom defect detection CNNs
Industrial vision classifiers

Binary Neural Networks

XNOR-Net
BinaryNet
FINN-generated NNs

Spiked Neural Networks

Brain-inspired inference
Event-driven vision
Robotics

Streaming Sensor AI Networks

Radar
LiDAR
RF Signals

Tiny Transformers

Keyword spotting
Sensor fusion
Predictive maintenance
Industrial monitoring

Why eFPGA Wins

INT8 / INT4 / Binary implementations
Deep pipelining
No DRAM bottleneck
Deterministic latency

GPUs process frames in batches.
FPGAs process pixels continuously

Multiply / Accumulate operations replaced with XNOR & POPCOUNT

FPGAs are exceptionally efficient here. Researchers have shown 10x-100x better TOPS/Watt versus GPU for BNN workloads

Event-driven
Sparse activity
Custom routing

GPUs are optimized for dense matrix multiplication
SNNs are not

FPGAs can process every sample as it arrives

GPU often requires:
Collect
Store
Transfer
Process

Tiny transformers with:
- small sequence lengths
- quantized weights
- edge inference

can run efficiently on FPGA

Microsoft Project Brainwave: When FPGAs Outperformed GPUs by >10×

Real-time RNN inference at batch size = 1 achieved more than an order-of-magnitude improvement in latency and throughput versus state-of-the-art GPUs.

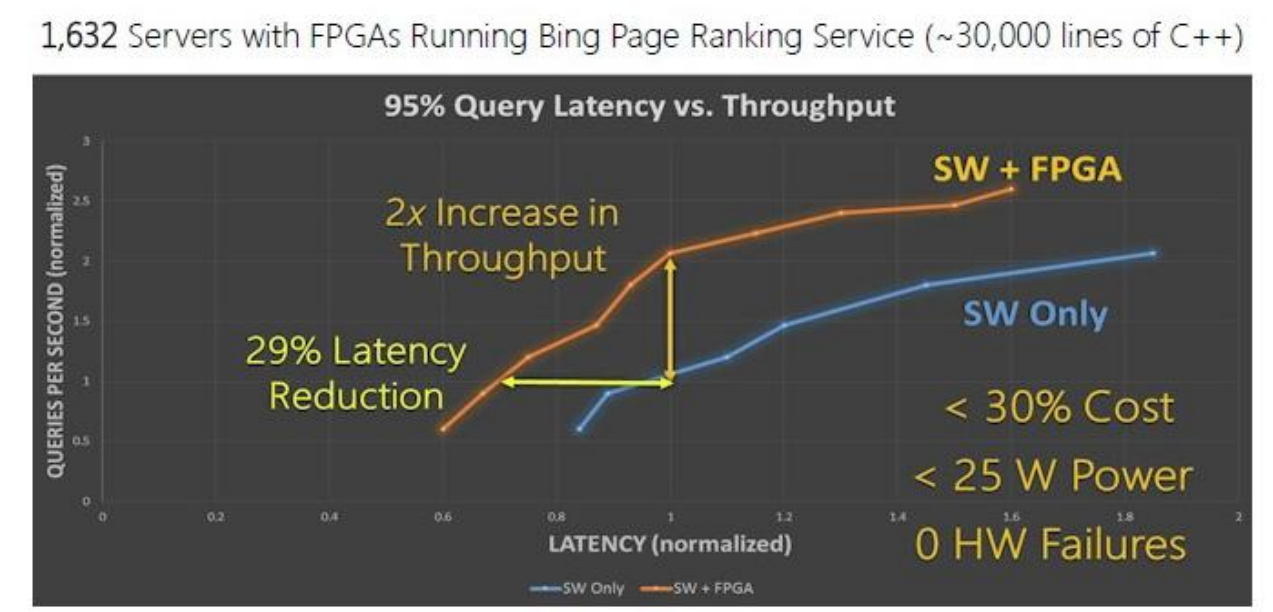
Bing Deployment

Microsoft explicitly states that Project Brainwave was used for:

- Bing intelligent search features
- Azure machine learning services
- Real-time AI inferencing infrastructure

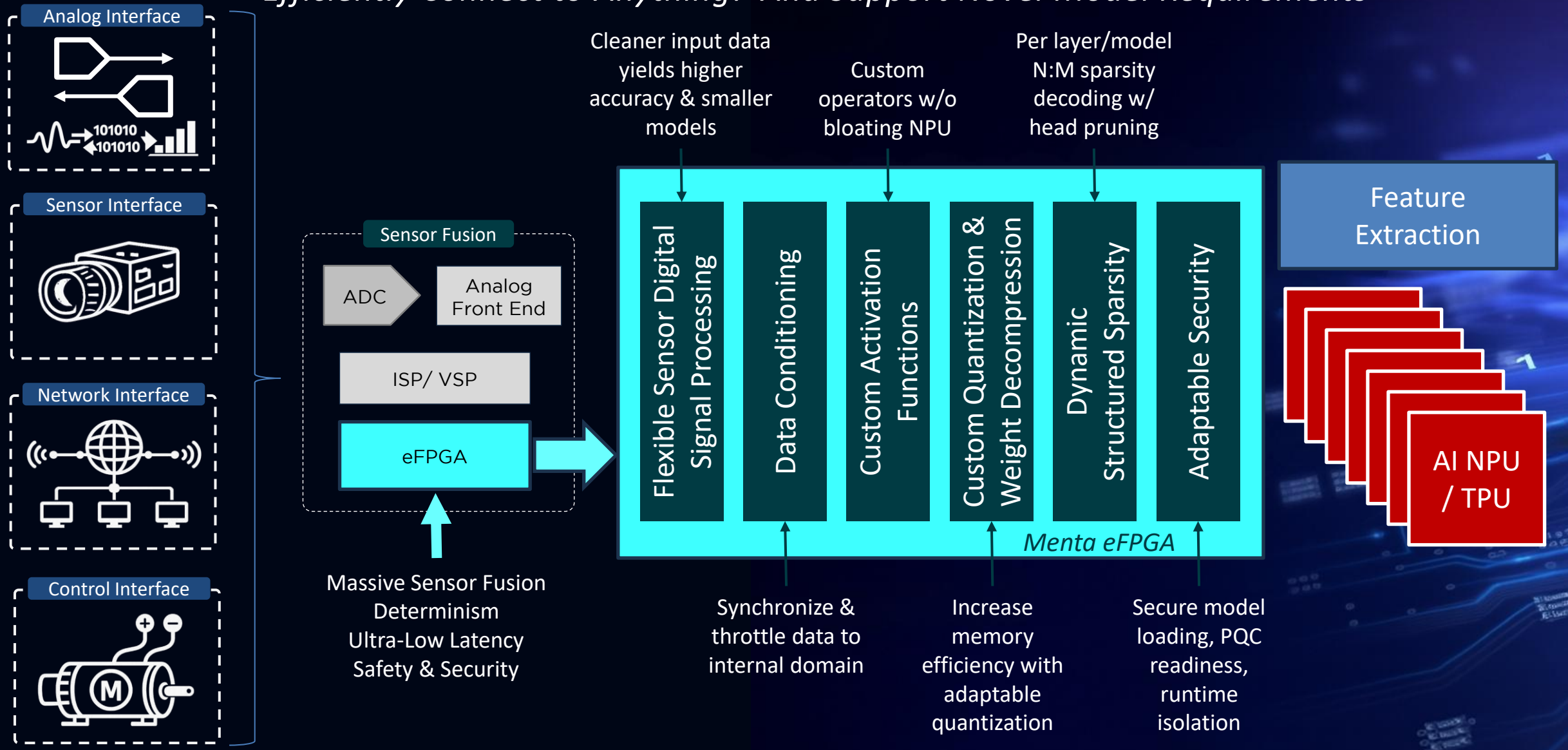
in production

Metric	GPU Approach	FPGA-based Brainwave
Batch Size	Large	1
Latency	Higher	Ultra-Low
Determinism	Variable	Deterministic
Real-time Inference	Good	Excellent
Model Updates	Software	Reconfigurable Hardware



Valuable NPU/TPU Companion

Efficiently Connect to Anything! And Support Novel Model Requirements



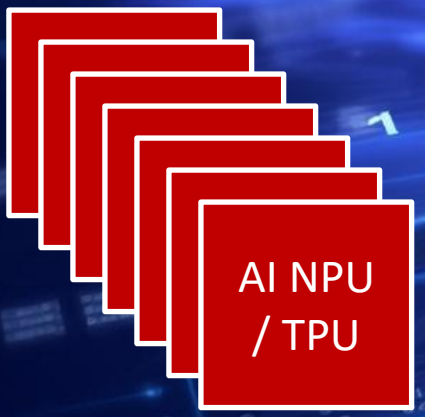
Adaptable Feature Extraction with eFPGA

Feature extraction & preprocessing before data ever reaches the AI accelerator

Examples of Workloads Well-Suited for eFPGA

Function	Benefit
Edge Detection	Reduce Vision Workload, ROI Support *Sobel/Canny Filter, Prewitt/Scharr Operator, DoG*
Transforms	Radar/LiDAR Preprocessing *FFTs, CFAR Algorithms, Clutter Filtering*
Beamforming	Adaptive Radar Systems *Steering, Multi-beam/MVDR Generation, Direction Finding*
Optical Flow	Motion Tracking *Flow support, Point-tracking, Segmentation, Estimation*
Sensor Fusion	Deterministic, Low-Latency Physical AI *Debayering, Scaling, HOG (Histogram Generation)*
Compression	Reduce Memory Bandwidth *ROI Compression, Encoding, Adaptable Quantization*
Quantization	NPU Optimization *Data Conversion, Channelization, Posit Arithmetic*
Sparse Encoding	Reduce Compute Load *Lower Bandwidth, memory, power – Faster Inference*
Cryptographic AI Security	Secure Model Execution *Support novel security algorithms, defend against novel threats*

Feature Extraction



Don't let your $100M AI ASIC become obsolete because the preprocessing pipeline needs to change

Example Physical AI Design on Menta eFPGA

AI Models Change. Sensor Processing Changes. Adaptable Hardware Makes Both Possible

… and Efficient

Pixel Stream

3x3 Line Buffer

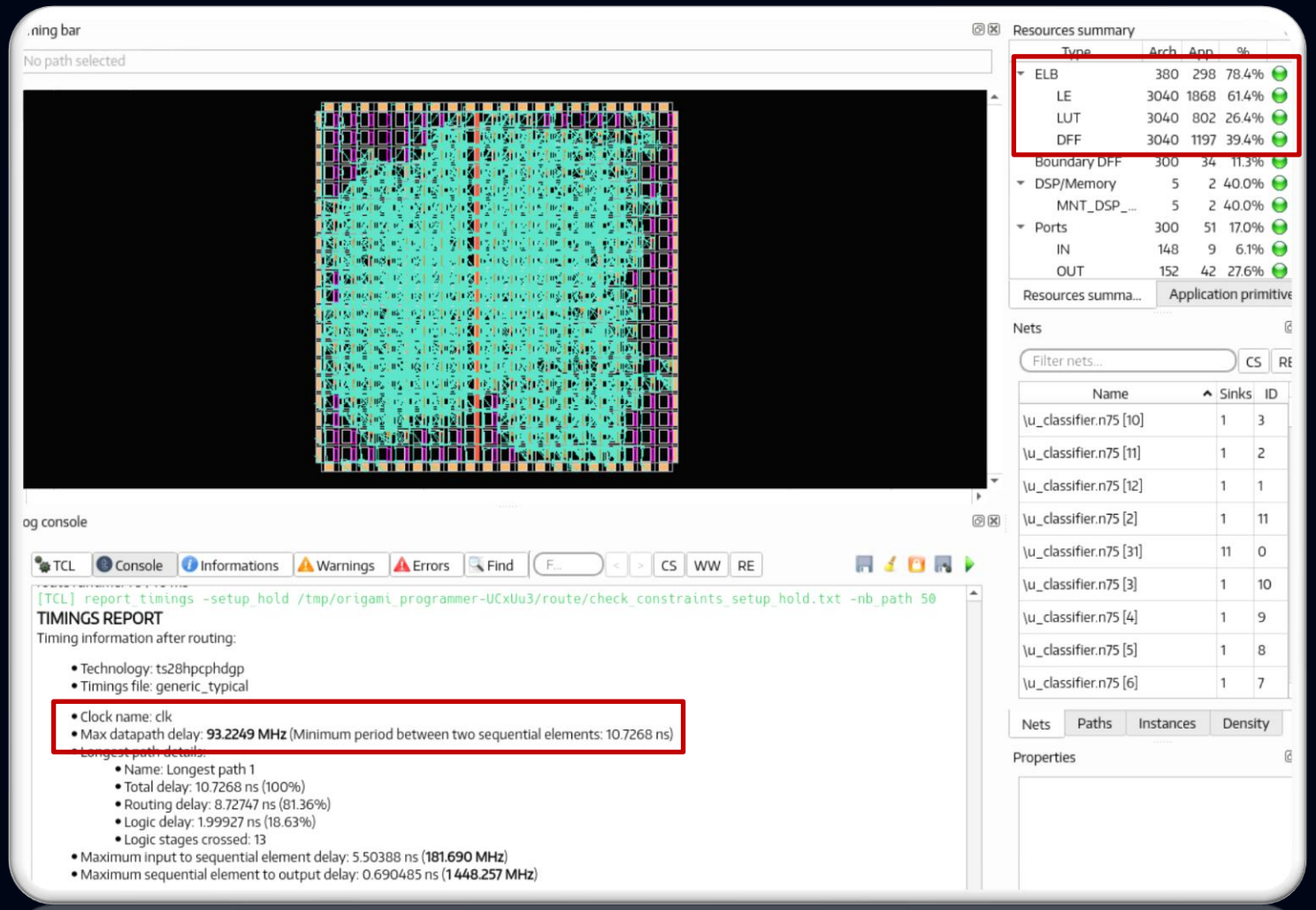
Sobel Edge Detection

Sobel
Optical Flow
Harris Corner
Compression
Sparse Encoding
Quantization

ROI Feature Accumulator

Quantized Defect Classifier

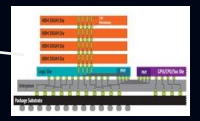
Defect / No Defect



Optimize AI Memory Efficiency with eFPGA

Prevent AI Hardware Obsolescence with eFPGA Adaptability & Memory-Saving Techniques

Memory drives AI solution cost



DRAM
HBM

Large LLM & GenAI models are dominated by weights → memory

Save memory with
- Lossless LLM weight compression
- Structured sparsity
- New numerical formats

Compressed Weights

Weight Decompression

2x

Structured Sparsity Weight Population

2x

INT2/4 → INT8 Weight Remapping

4x

INT8/16 Weight Interface

TPU
AI Engine

Activation Functions & New Operators

Adapt to changing AI Algorithms
Memory shim maximizes memory efficiency between models with different weights

- Different models & layers can employ different decompression & remapping techniques
- Deterministic latency with no loss of throughput

Menta also supports Output Activation & new operators in **existing HW**

Menta eFPGA IP

Applications of eFPGA in SoCs & ASICs

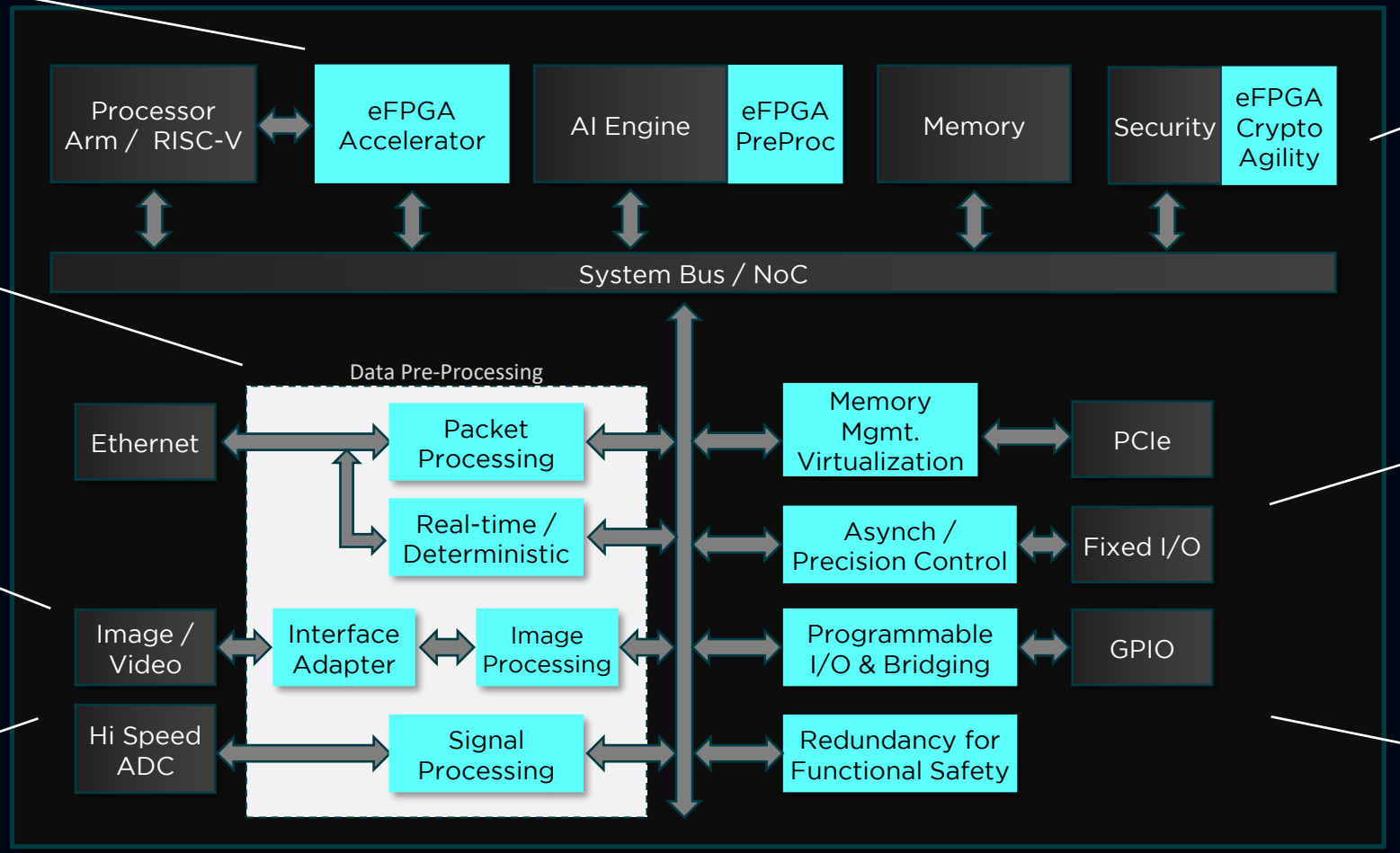
Security. Flexibility. Lifecycle Extension

Accelerate complex algorithms with custom ISA Instructions

Programmable Data Plane Processing Deterministic Industrial Networking

Adaptable Image Signal processing adapts to more sensors, increasing SAM

DSP Filtering Transforms



Protect against evolving threats & quantum computing

Obfuscate secret IP

Enable security autonomy

Expand Physical and Virtual functions Bridge & DMA data

Interface Flexibility – Adaptable to different markets & applications

Eliminate unused ports – minimizing surface attack exposure

Menta eFPGA IP

5th Generation DESIGN ADAPTIVE eFPGA IP

Embedded Logic Blocks
Menta 6-input LUT

DSP Blocks
Customer Defined

Configuration
SPI | AXI | JTAG
AHB | Custom

DFT
Standard ASIC test scan chain interface

Embedded Memory Blocks
Any 3rd-party RAM

I/O Block
Define # of pins

Embedded Custom Block
Customer or 3rd-party IP

Clock / Reset
Control Input Block

Silicon-proven IP integrating logic, memory & custom blocks

Highly Differentiated eFPGA IPs

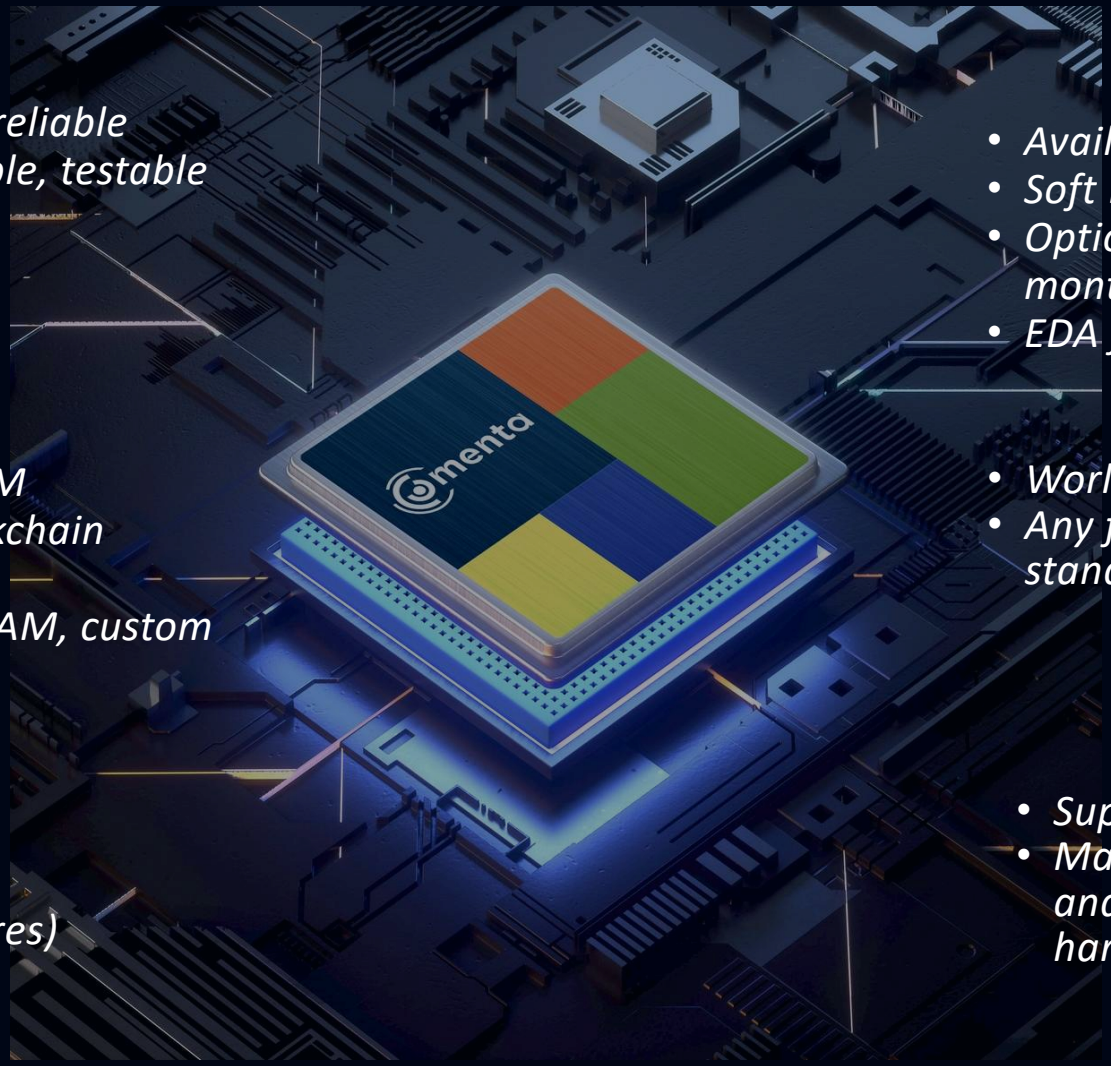
No Custom Cells. 100% Standard-Cell. Fully Adaptable

Core Technology

- *100% Standard Cell — portable, reliable*
- *D Flip Flop Config – robust, reliable, testable*
- *75% less config memory*

- *Advanced Blocks - LUT6, DSP, RAM*
- *Unique IP ID - watermark & blockchain traceability*
- *Flexible Integration - 3rd-party RAM, custom blocks, optimized routing*

- *Configurable DSP (width & features)*
- *Best Test & Fault coverage*



Deliverables

- *Available as Soft IP or Hard IP*
- *Soft IP delivery < 2 weeks*
- *Optional: physical design in 1-6 months*
- *EDA flow agnostic*

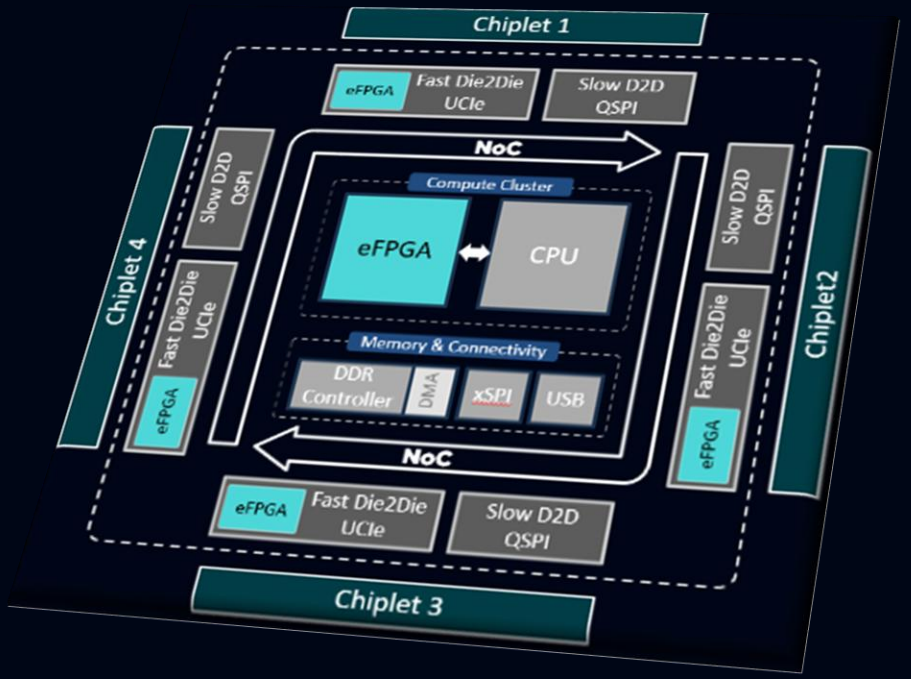
- *Worldwide support throughout*
- *Any foundry, process, metal stack, standard cells & process options*

- *Superior reliability, aging & yield*
- *Make use of any specialized process and standard cells - Auto, Space (rad-hard)*

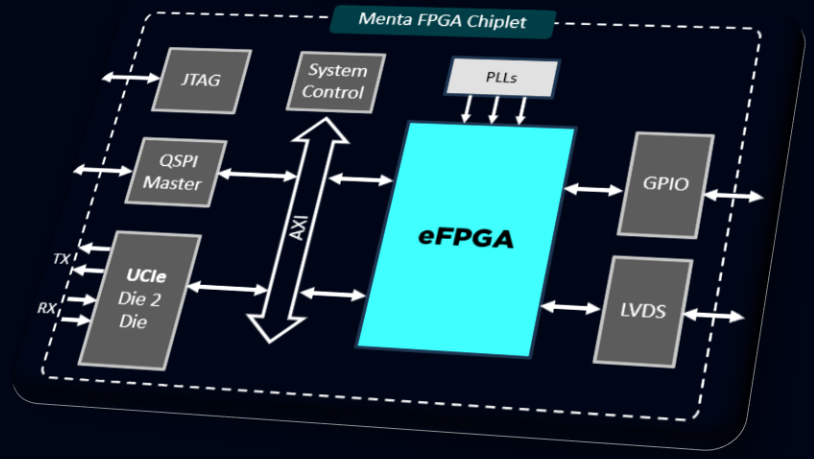
menta
Unlimited Power of Technology

Introducing MOSAIC SiP Development Platform

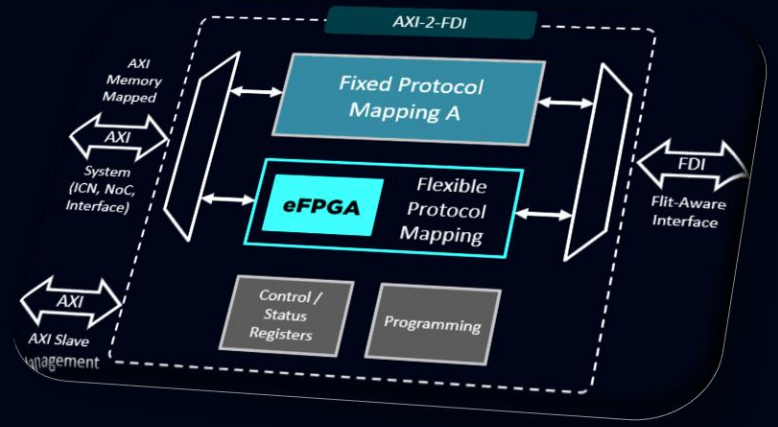
MOSAIC CHASSIS



Menta FPGA Chiplet

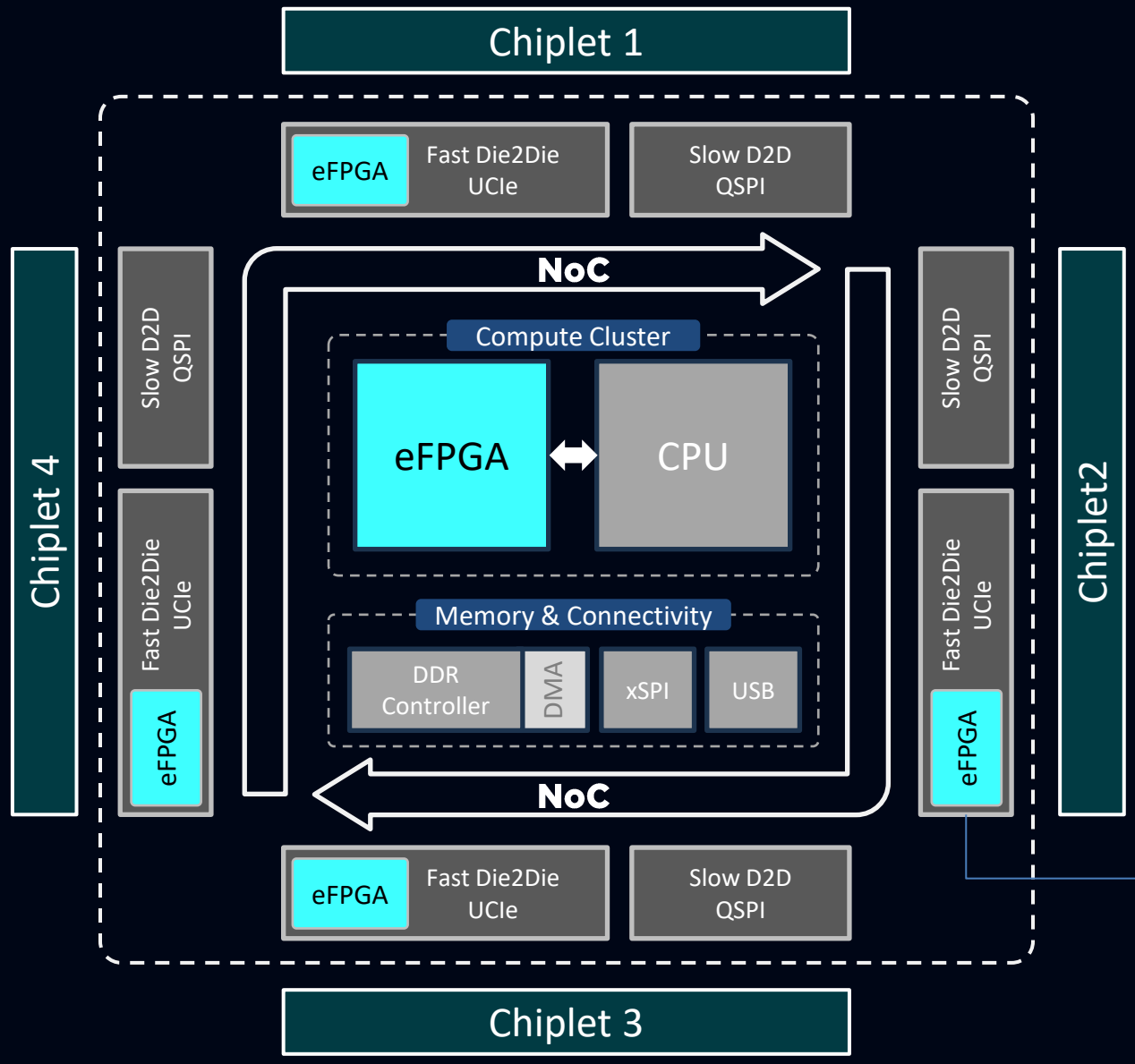


Menta AXI-2-FDI Adapter



cea ARTERIS IP SYNOPSYS® SIEMENS

Introducing MOSAICs Chiplet Chassis



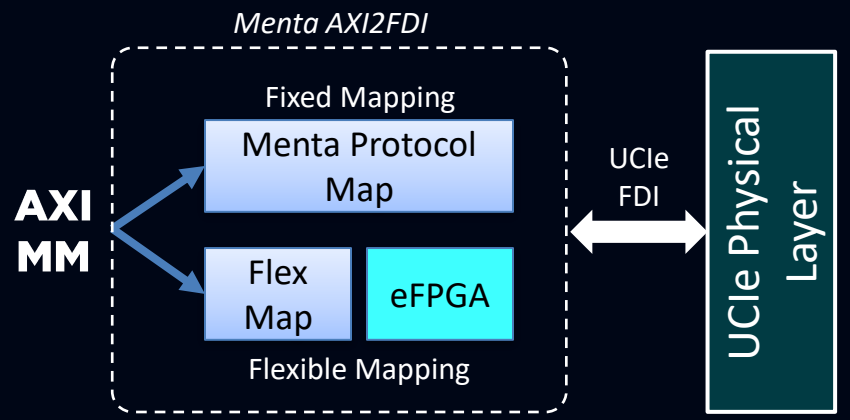
Seamlessly design and develop heterogeneous SiP solutions with unmatched flexibility

SoC Infrastructure

- RISC-V CPU
- Menta eFPGA
- Arteris NoC
- Flexible DMA
- LPDDR4/DDR4
- xSPI Flash

Chiplet Connectivity

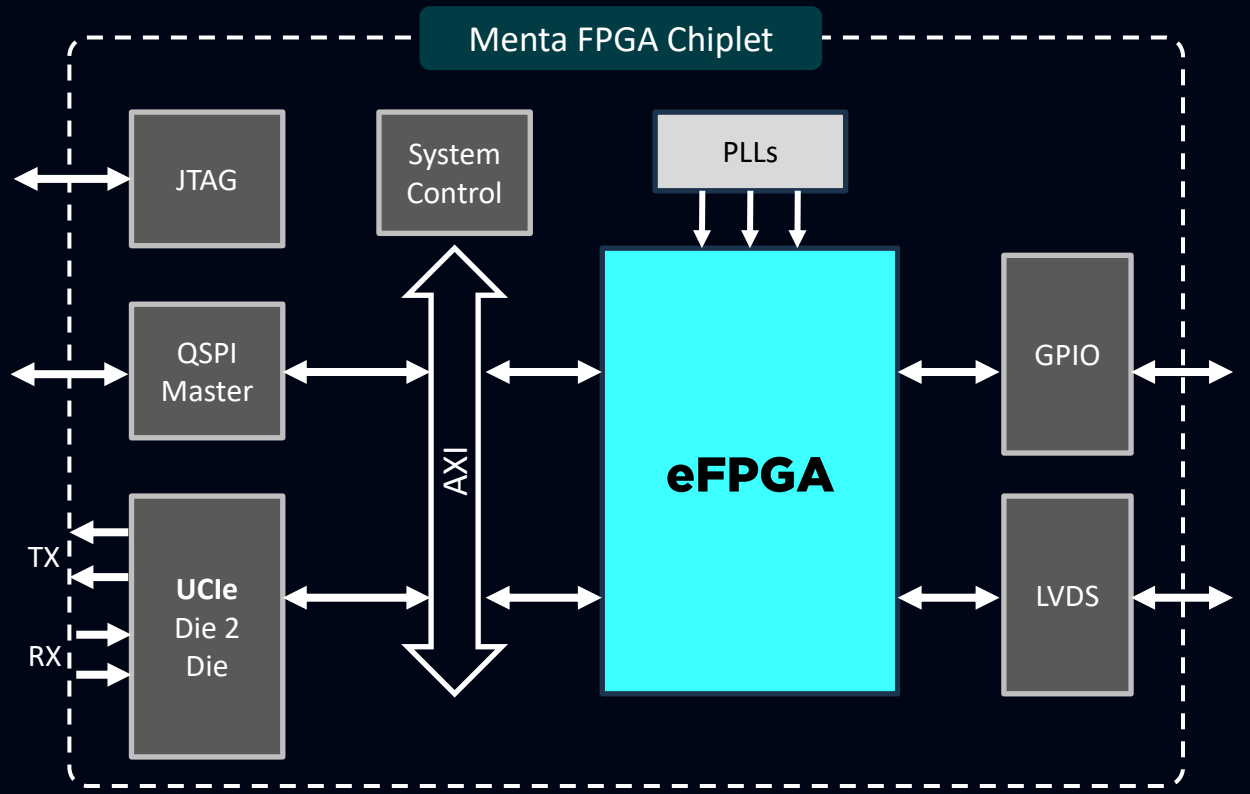
- 4x Fast UCIe Flexible eFPGA Transport Interfaces
- 4x Slow QSPI Interfaces
- USB3 Off-chip Connectivity



Menta FPGA Chiplet

Endless Applications & Benefits

- High-Speed Protocol Adaptation & Bridging
- Real-Time Signal Processing
- Security, Crypto-Agility & Key Management
- Low-Latency Networking & Data Movement
- AI/ML Pre- and Post-Processing
- Storage & Memory Subsystem Adaptability
- Aerospace & Defense Mission Logic
- Industrial & Robotics Control
- Automotive & ADAS Applications



Adaptability to evolving protocols and standards

Deterministic low-latency performance

Hardware updates protect the solution investment

Why Leading Innovators Choose Menta eFPGA

Cost & Time Efficiency

- **Lowest Overall Cost & Highest and predictable yield**
  Small die area, competitive licensing
- **Fastest Time to Market**
  IP delivery < 2 weeks
- **Foundry sign-off flow**
  Reliable, 99.8% test coverage

Control & Flexibility

- **Full Architectural Control**
  IP delivery to your requirements
- **EDA Ownership**
  No reliance on 3rd-party tools, redistributable
- **Future-Proof Design**
  Easily reconfigurable

Partnership & Support

- **Dedicated IP Partner**
  Focused on IP & EDA only
- **Global Support Network**
  Responsive & expert technical support
- **Planned R&D Expansion**
  Long-term innovation roadmap

Menta is the Clear Choice for eFPGA

Physical AI Demands Adaptability

Adaptable Hardware is great for…

1. … evolving quantized AI algorithms
2. … evolving sensor fusion + determinism + low-latency
3. … bespoke AI feature extraction demands

Unlimited Power of Technology

Thank you!

info@menta-efpga.com

us_sales@menta-efpga.com

apac_sales@menta-efpga.com

https://www.menta-efpga.com/