JULY 9–13, 2023

MOSCONE WEST CENTER

SAN FRANCISCO, CA, USA

MENTIUM
TECHNOLOGIES

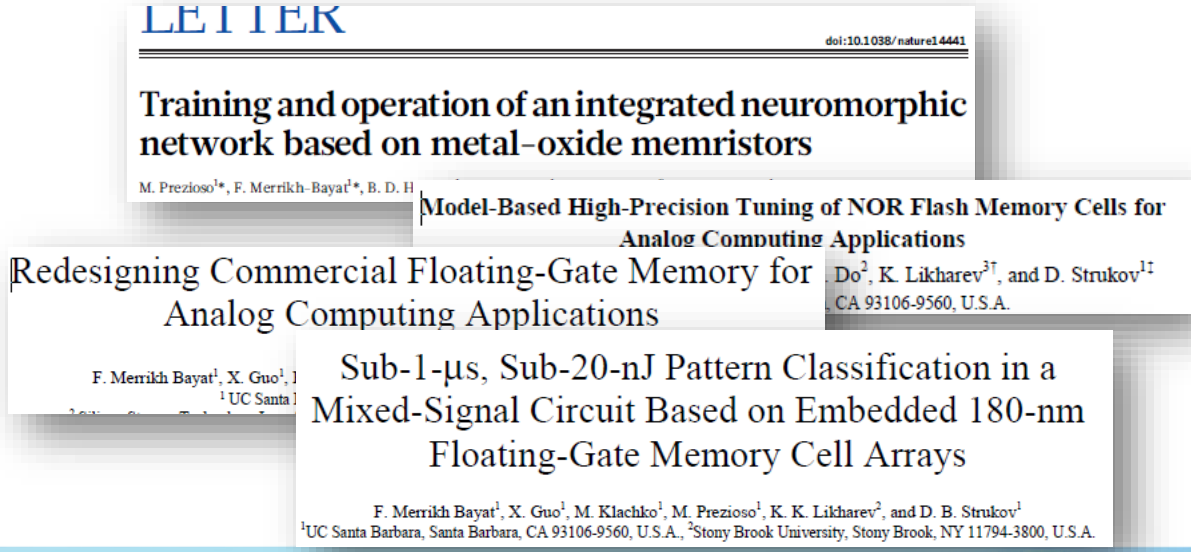
Evolving Edge

AI For mission-critical applications

Introduction

It all started at UC Santa Barbara…

- I and Farnood (CTO) joined Prof. Strukov's lab in 2013:
- By 2016 the team showed great results in in-memory computing for Neural Networks
- We incorporated Mentium Technologies in 2017



MENTIUM TECHNOLOGIES

The Managing Team

Mirko Prezioso, PhD
Co-founder, CEO

First demonstration of neural computation on integrated Memristors

UCSB

National Research Council of Italy

Craig Ensley,
Executive Board Member

CIRRUS LOGIC

pSemi A Murata Company

ATOMICA

Farnood M. Bayat, PhD^2
Co-founder, CTO

First demonstration of a in-memory AI processor with eFlash Devices

UCSB

Mark Ross
VP of Engineering

CYPRESS EMBEDDED IN TOMORROW

NeXT

CISCO

Sun microsystems

Paul Pickering,
Marketing Strategy

TELEDYNE FLIR Everywhereyoulook™

ATOMICA

Microsemi

Prof. John Bowers,
Board Member, advisor

UCSB

CALIENT OPTICAL SWITCHES

NOKIA

George Jones, advisor

NXP

500

Sand Hill Angels

WCP

Pete Rodriguez, Board Observer, advisor

Si SILICON CATALYST

NXP

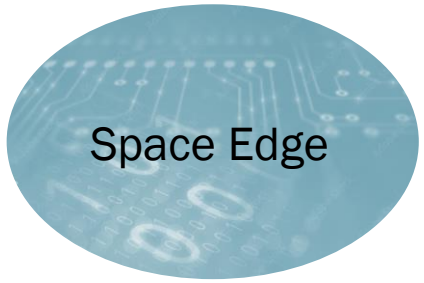
EXAR

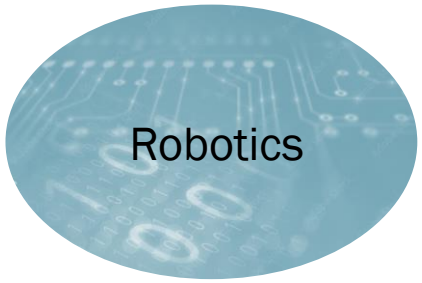
60

MENTIUM TECHNOLOGIES

AI co-processors for complex applications at the EDGE

- **Top Quality AI** for **complex problems** at **the Edge Power**
- **Expanding** existing **systems** by **easy integration:** we do **not compete** with **SoC manufacturers**
- **Adding Orders-of-magnitude better** performance at **ultra-low power**
- **Extremely scalable, Broad** horizontal **market** applications
- **Offered also as Rad-hard** solution



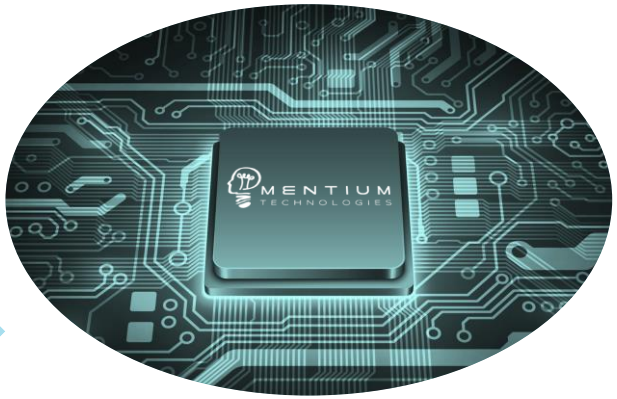


Key differentiators

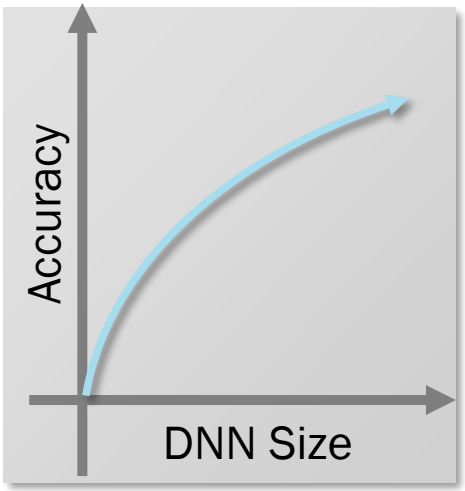
- Pure co-processing approach
- Unique hybrid digital-analog in-memory architecture
- Rugged design for harsh environments

Whenever the customer needs **dependable AI** inference

Mentium is **#1 solution**

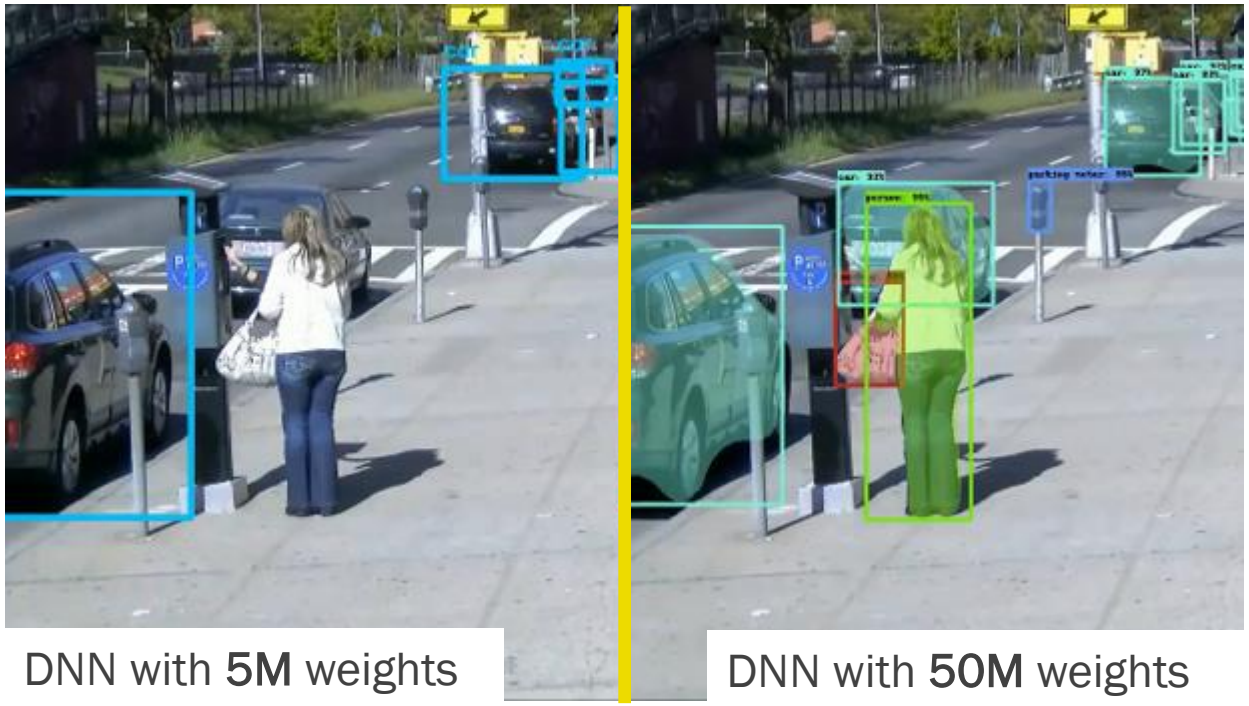


AI Analytics are only as good as your Neural Network



Current Edge Performance

Cloud Performance



DNN with **5M** weights

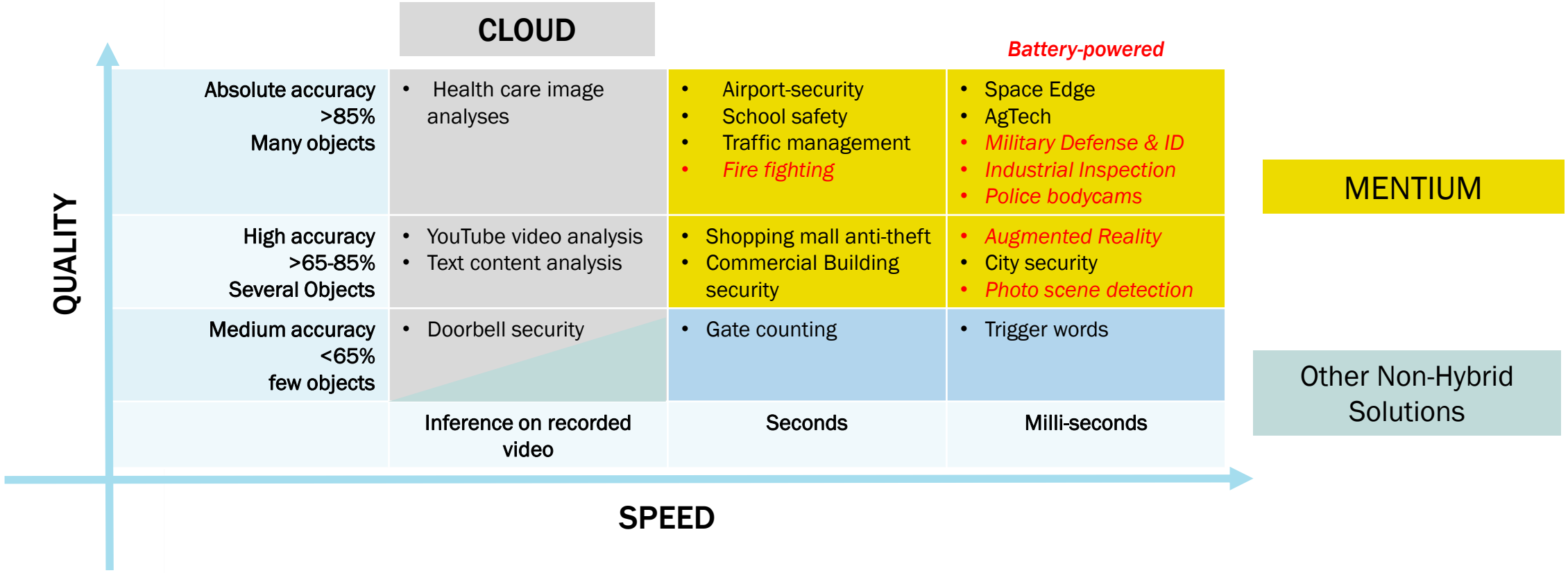
DNN with **50M** weights

Cloud Issues
Latency
Privacy
Reliability
Costs
Scalability

<u>Mentium's equipped IoT devices can run the most complex models That are currently ran on Cloud</u>

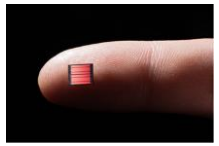
- Higher quality in real-time
- Operating Costs reduction **100s times** cheaper!
- New applications

Mentium Enables Broad Classes of Applications that SoCs alone or Cloud AI cannot support



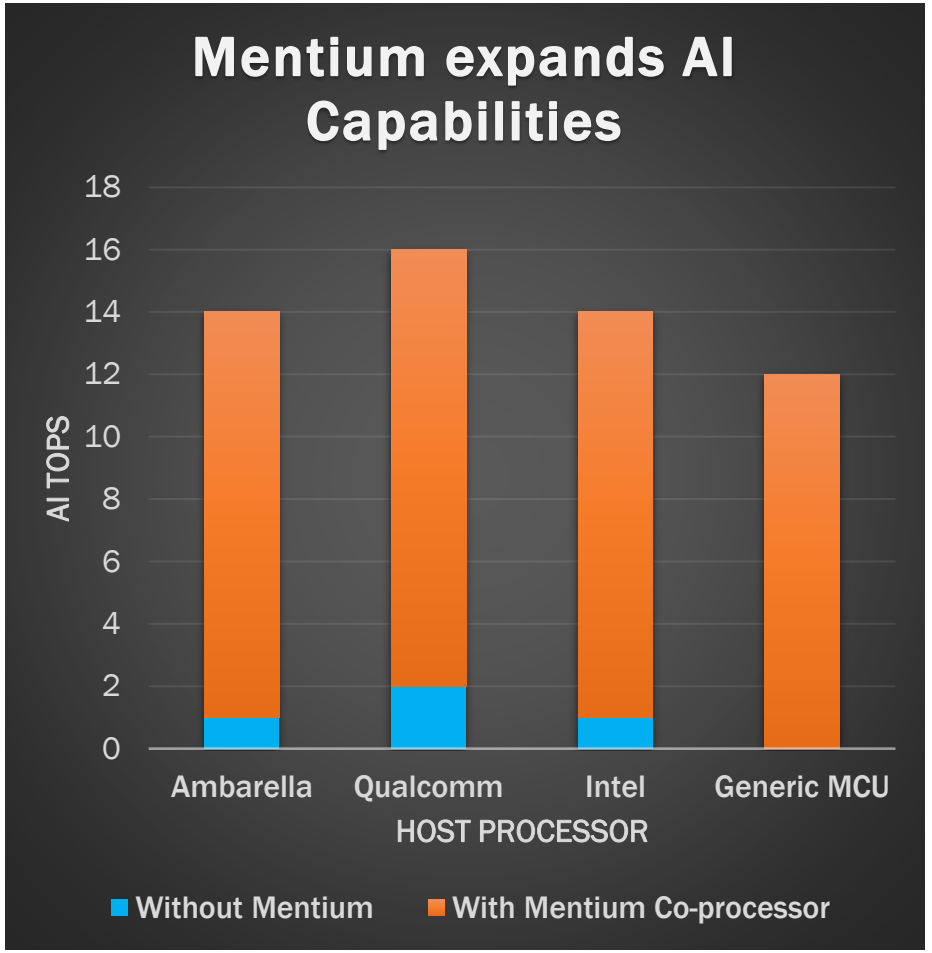
Edge Co-Processor For Cloud-Accuracy AI Analytics

+



Simple **Bolt-on solution.**
Completes and **expands** AI capabilities of **existing systems** with their SoCs:

- ✓ **Cloud-quality** inference at **0.1 to 0.5 watts**
- ✓ **Adding 12 AI TOPS with less** than **1W TDP**
- ✓ **No need for external memory**
- ✓ **USB/PCIe** interfaces
- ✓ **Easy** hardware and software **integration** across **multiple platforms** and BUs
- ✓ **Re-use** of customer's DNN deployment **toolchain**
- ✓ **Rugged Design** for harsh environments

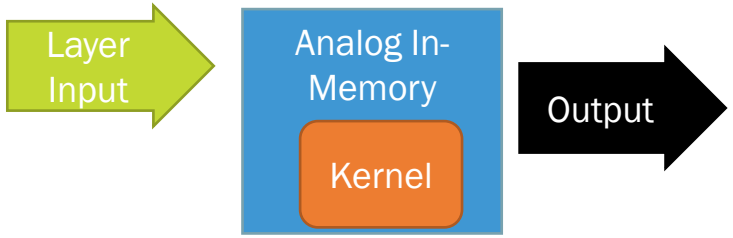


MENTIUM TECHNOLOGIES

How does Mentium reach this efficiency and throughput?

DNNs operation (VMM): Layer Input X Kernel = Output

PRO: No ADC and DAC overhead
CONS: A lot of energy and time spent on Input and Kernel transfer



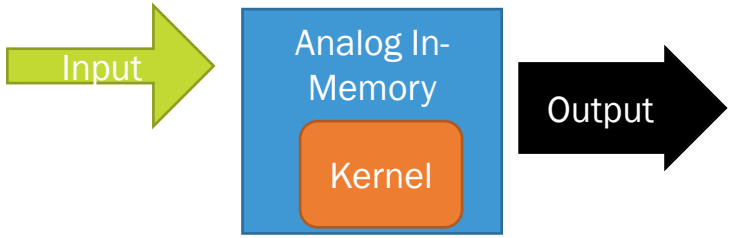
PRO: No time or energy spent on Kernel transfer
CONS: A lot of energy and time spent on Input DAC and Output ADC.

How does Mentium reach this efficiency and throughput?

DNNs operation (VMM): Layer Input X Kernel = Output

PRO: No ADC and DAC overhead
CONS: A lot of energy and time spent on Input and Kernel transfer

BEST: Small Kernels, lot of repetitions



PRO: No time or energy spent on Kernel transfer
CONS: A lot of energy and time spent on Input DAC and Output ADC.

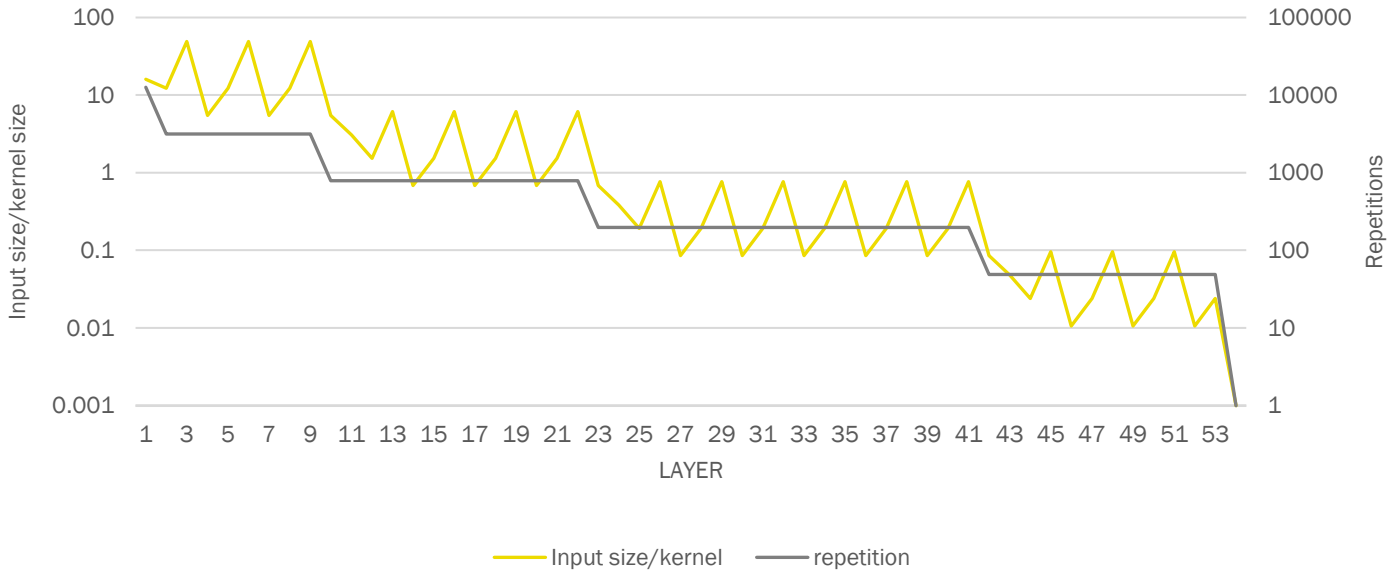
BEST: Large Kernels x Small inputs, few repetitions

Example: Let's look at Resnet 50

$$\frac{Input\ Size}{Kernel\ Size}$$ and *repetitions*

vs

Layer



Digital Processing	I/K ratio & Repetitions / Kernel Size	Analog In- Memory Processing
• Small Kernels • many repetitions		• Large Kernels • few repetitions

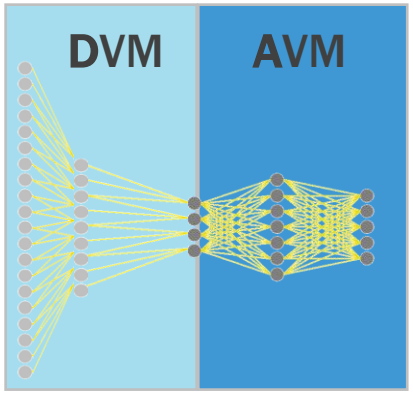
Hence: Mentium Hybrid Architecture

Fast – Accurate – Efficient

Based on two equally important parts

Digital Computing (DVM Core)

- Co-designed with AVM
- Proprietary architecture
- On par with AVM efficiency and speed
- 10x better for **sparsely** connected layers



Analog in-memory computing (AVM Core)

- No more memory bottleneck
- 40x storage density advantage
- Massively parallel
- Unmatched on **densely** connected layers

DVM & AVM co-designed synergistic system
Delivering best-in-class efficiency at high throughput

MENTIUM TECHNOLOGIES

The hybrid architecture delivers

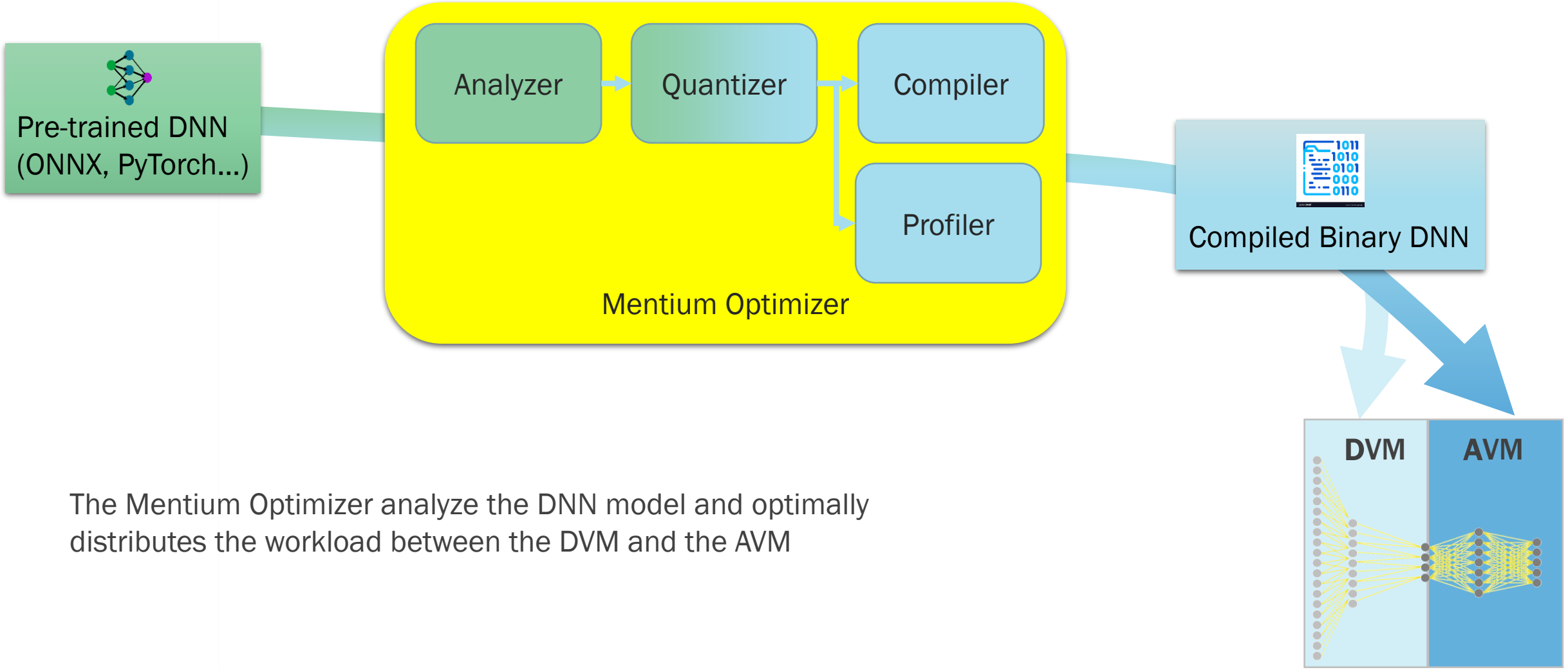
Security Camera Sequence Analysis	Intel Myriad X	Qualcomm Q8250	Mentium AIM	
Image Recognition Benchmark (Resnet-50 224x224 input size)			Full speed	Real-time On battery power
Inference Speed IPS	20	25*	650	30
Power (watts)	1.3	3	**0.4**	**0.1**
Production Object Detection (SSD+MobileNetv2 300x300 input size)				
Inference Speed IPS	14.7	18*	210	30
Power (watts)	1.6	3	**0.5**	**0.1**
Large DNNs (Reference models)	NO	NO	**YES**	

- **SSD-Mobilenet_v2 input size 300x300, 8-bit precision**

***Derived from benchmarks on similar networks**

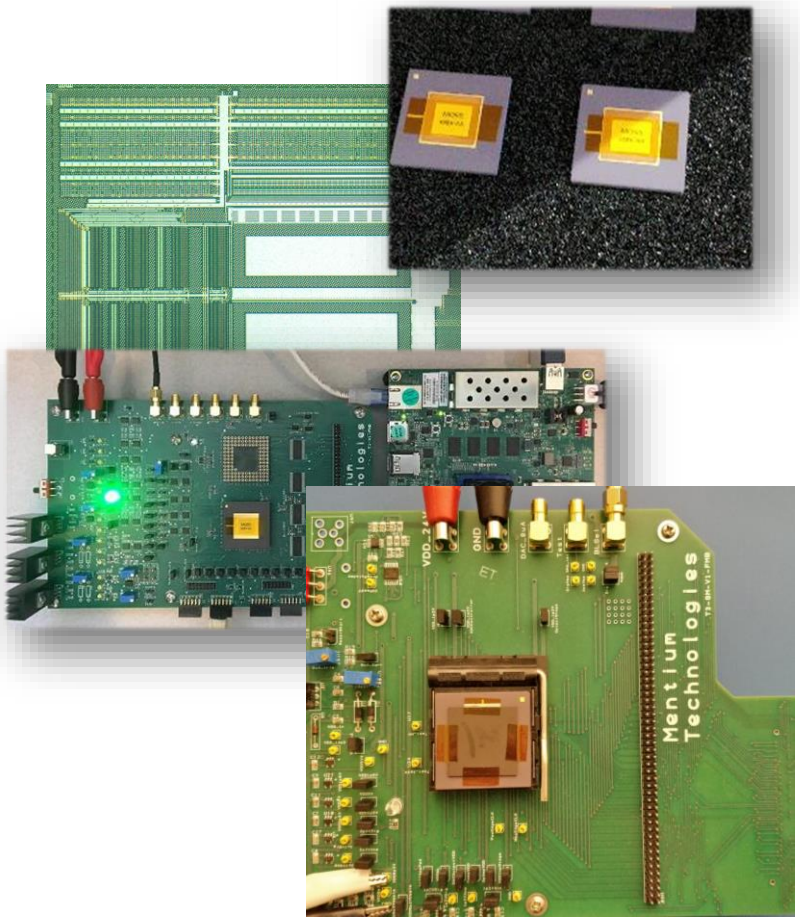
MENTIUM TECHNOLOGIES

Software pipeline



The Mentium Optimizer analyze the DNN model and optimally distributes the workload between the DVM and the AVM

Road map



- AVM tested on Silicon
- DVM validated in post-layout Synthesis

- 2023Q4 DVM Tape-out
- 2024Q1 Dev Kit Sampling
- 2024Q4 Production tape-out

Wrap-up

UNIQUE HYBRID APPROACH

SIMPLE INTEGRATION, COMPLEMENTARY TO ALL SOCs & MCUs

#1 IN MISSION-CRITICAL AI AT LOWEST EDGE POWER

Happy to talk more! mprezioso@mentium.tech

LinkedIn:

MENTIUM TECHNOLOGIES