AI HARDWARE & EDGE AI SUMMIT

MENTIUM TECHNOLOGIES

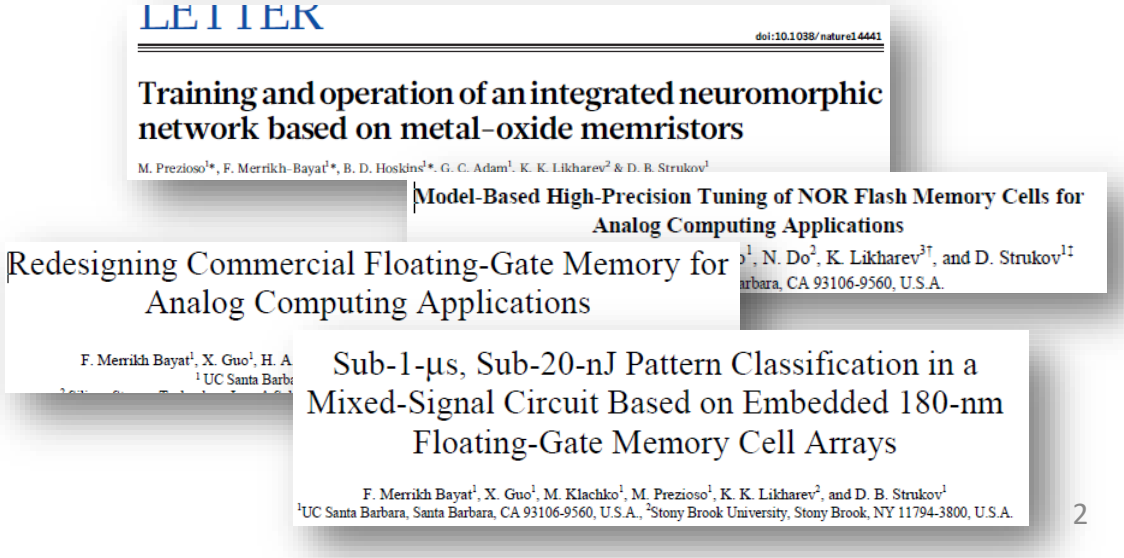
MENTIUM HYBRID COMPUTATION FOR MISSION-CRITICAL EDGE AI, PARTNERING WITH SYNOPSYS

Mirko Prezioso, Co-founder & CEO, Mentium Technologies

September 10, 2024

Mentium Technologies History

- Spin-off from UCSB
- In 2015, we demonstrated the first DNN in-memory computing in an integrated array.
- Mentium, an AI Co-processors Startup, was founded in 2017

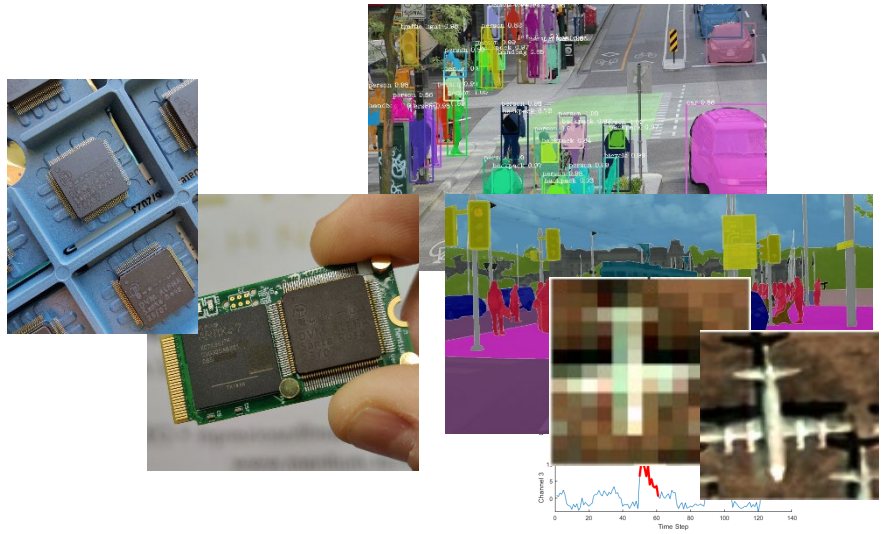


MENTIUM TECHNOLOGIES

Mentium Today

Cloud-level AI on Edge Systems for **mission-critical AI** applications

- **Sub 1W region**
- **Hybrid architecture allows >10x larger** models
  - **AI capabilities** (accuracy, complexity, reliability)
- **Enhancing** existing **systems** by **easy integration**
  - **No competition with SoC manufacturers**
  - **Saves customers millions of dollars to re-use existing system software**
- **Rad-hard** option developed with NASA → no other solution



Space Edge

High-End Smart Security

Military & Police Bodycams

Expanding to

Robotics

AR/VR

Drones

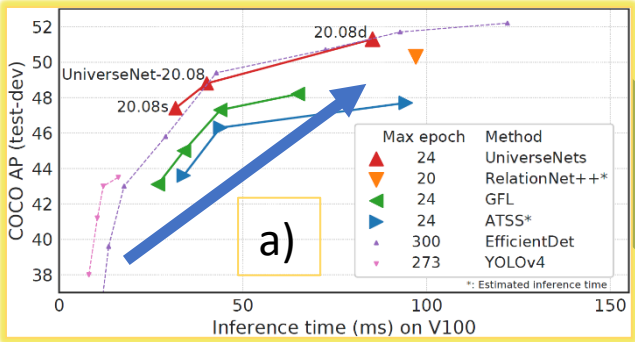
GenAI

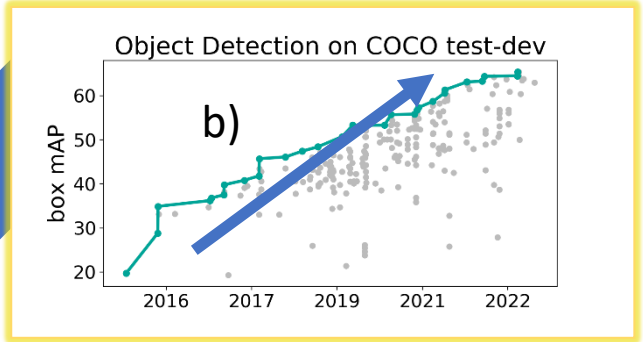
ADAS

Mobile

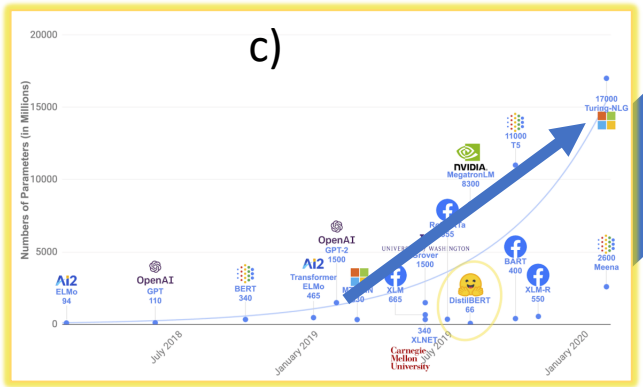
MENTIUM TECHNOLOGIES

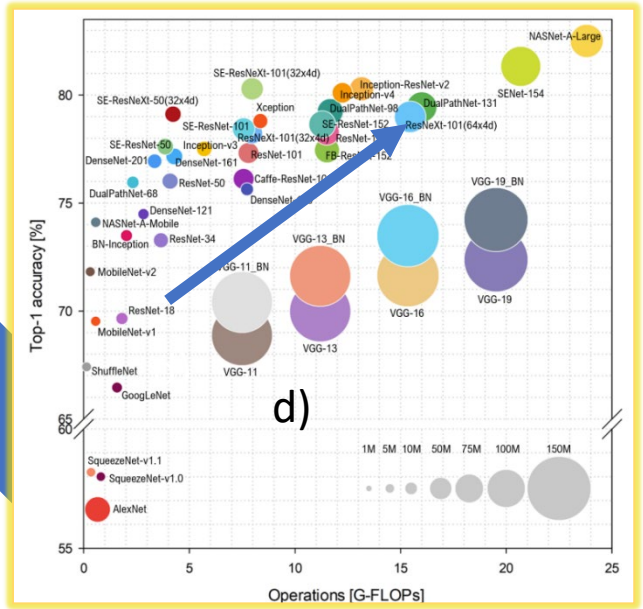
AI Inference is only as good as your Neural Network





Neural Network Capacity
"Inference quality"
is proportional to
Neural Network size

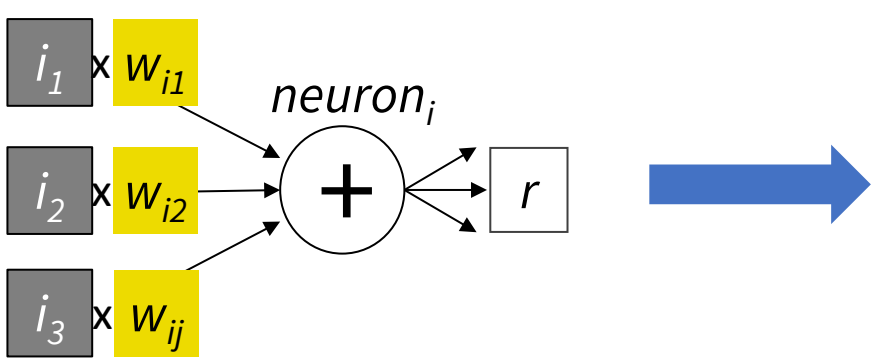


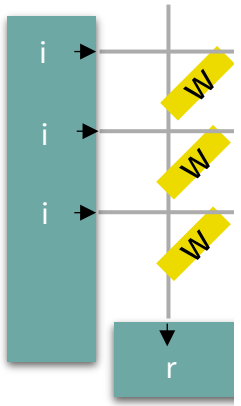


References:

a) COCO test-dev Benchmark (Object Detection) | Papers With Code
b) docs/en · master · weixin_47013828 / myrep · GitCode
c) How Hugging Face achieved a 2x performance boost for Question Answering c) with DistilBERT in Node.js — The TensorFlow Blog
d) 1810.00736 (arxiv.org)

Analog in-memory computing is great…





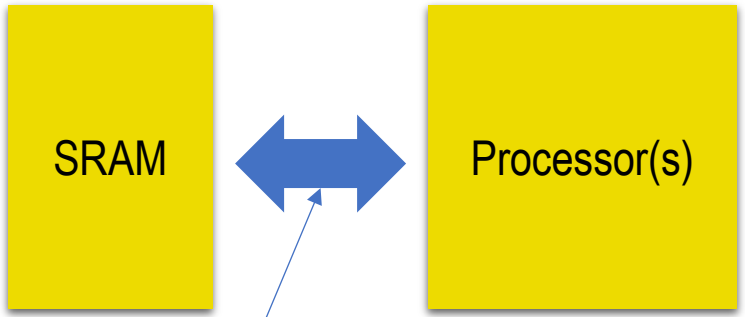
Large DNN model issues	Analog in-memory advantages	Solved?
On-chip storage	Great storage density advantage over SRAM	✓
Memory Bottleneck	Storage and computing in the same physical device	✓
Throughput	Massively parallel	✓

So, problem solved, right? …No!

…but not perfect.

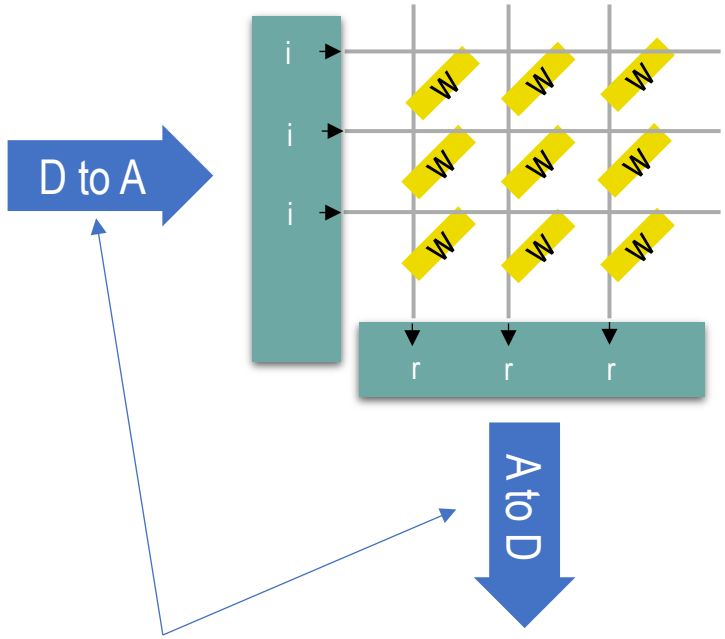
Both Digital and Analog in Memory Computation have Bottlenecks

Digital Processing



Data transfer between memory and processing:
Memory bottleneck

Analog in-Memory Processing

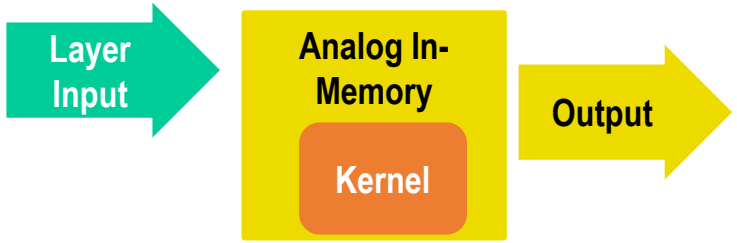


Data conversion between D to A and A to D:
Conversion bottleneck!

hybrid is best!

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

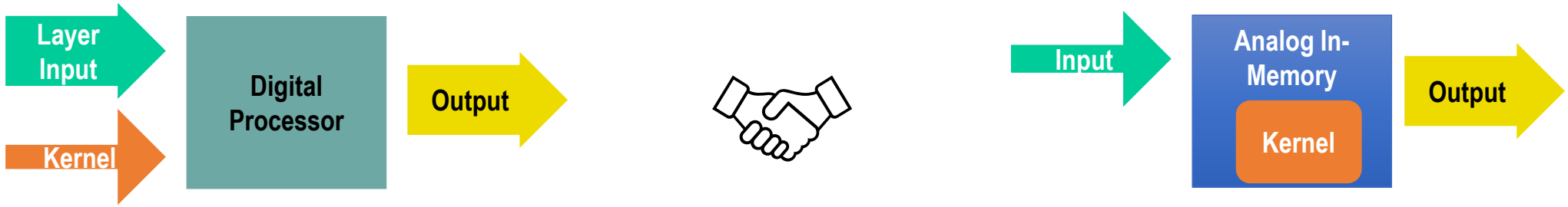
CONS: Lot of time/energy spent on Kernel transfer: Memory Bottleneck
PRO: No ADC and DAC overhead



CONS: Lot of time/energy spent on D/A/D conversion: Conversion Bottleneck
PRO: No time or energy spent on Kernel transfer

hybrid is best!

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



CONS: Lot of time/energy spent on Kernel transfer: Memory Bottleneck
PRO: No ADC and DAC overhead, much higher freq

BEST: Small Kernels, lot of repetitions

CONS: Lot of time/energy spent on D/A/D conversion: Conversion Bottleneck
PRO: No time or energy spent on Kernel transfer

BEST: Large Kernels x Small inputs, few repetitions

We use the best of both worlds!

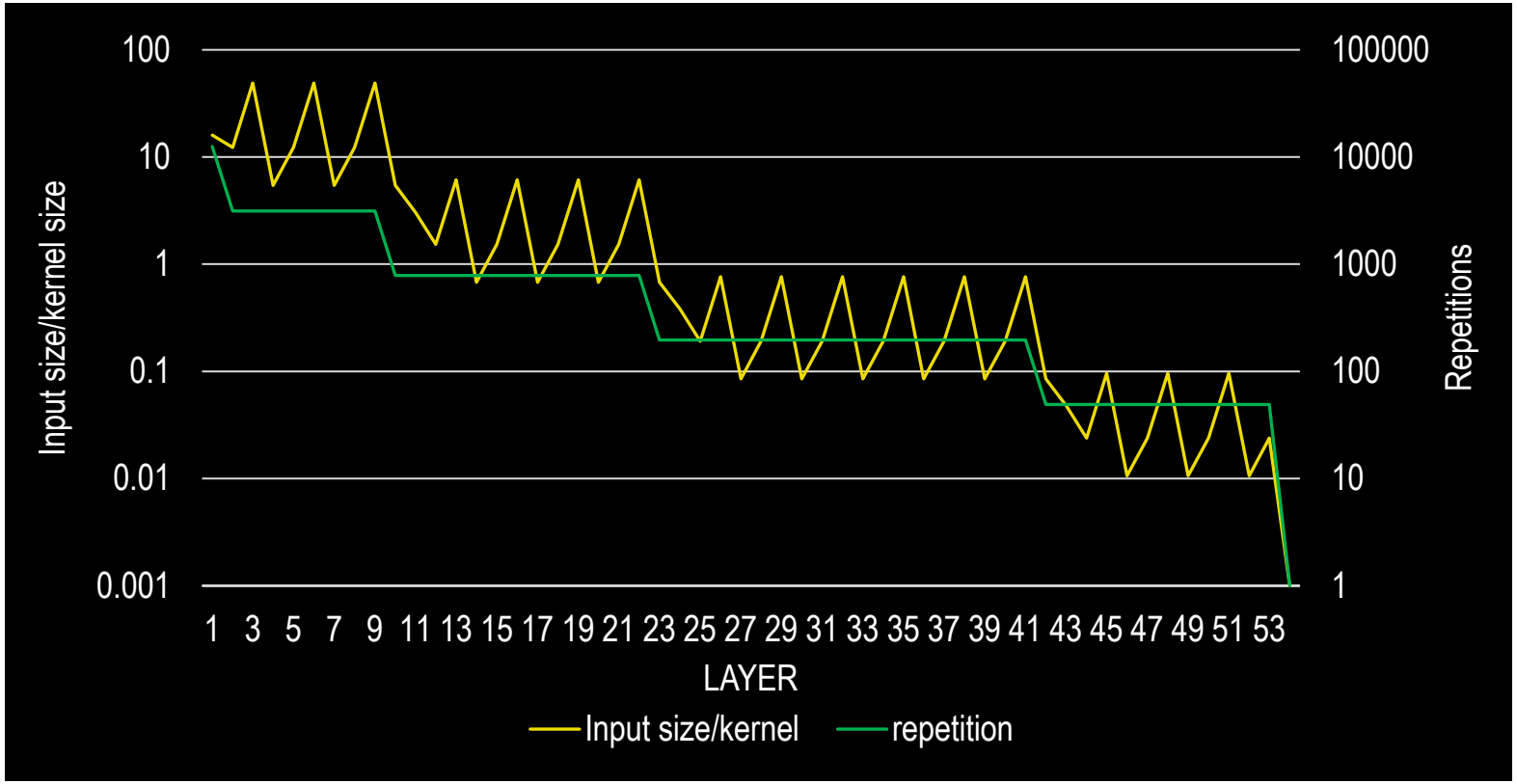
MENTIUM TECHNOLOGIES

Example: Let's look at Resnet 50

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

vs

Layer



Digital Processing
- Small Kernels
- many repetitions

Kernel Size

I/K ratio & Repetitions

Analog In- Memory Processing
- Large Kernels
- few repetitions

MENTIUM TECHNOLOGIES

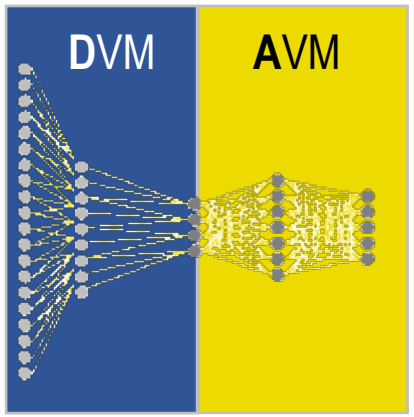
Unique Hybrid Architecture

Fast – Accurate – Efficient

Based on two equally important parts

Digital Computing (DVM Core)

- 10X better than any other digital design
- Co-designed with AVM
- On par with AVM efficiency and speed
- Best solution 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

Result: AIM architecture, a co-designed synergistic system
Delivering best-in-class efficiency at high throughput with large DNNs

Products road-map

Simple **Bolt-on solutions**.
Complete and **expand** AI capabilities of **existing systems** by helping their host processors:

Please feel free to contact Mirko Prezioso, CEO if you are interested in our products and/or our Roadmap
mprezioso@mentium.tech

- ✓ **Easy** hardware and software **integration** across **multiple platforms** and BUs
- ✓ **Re-use** of customer's DNN deployment **toolchain**

Software integration pipeline

Training and compiling DNNs

Dataset

Model Training
PyTorch, Tensorflow…

Dataset subsample

ONNX model

Mentium Compiler

Performance Profiling
Hardware-aware quantization
And optimization

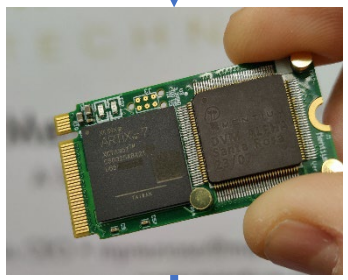
Optimized model

Testing DNNs on Dev kit

m.2 PCB
with Mentium Contributed APIs

+

Mentium APIs



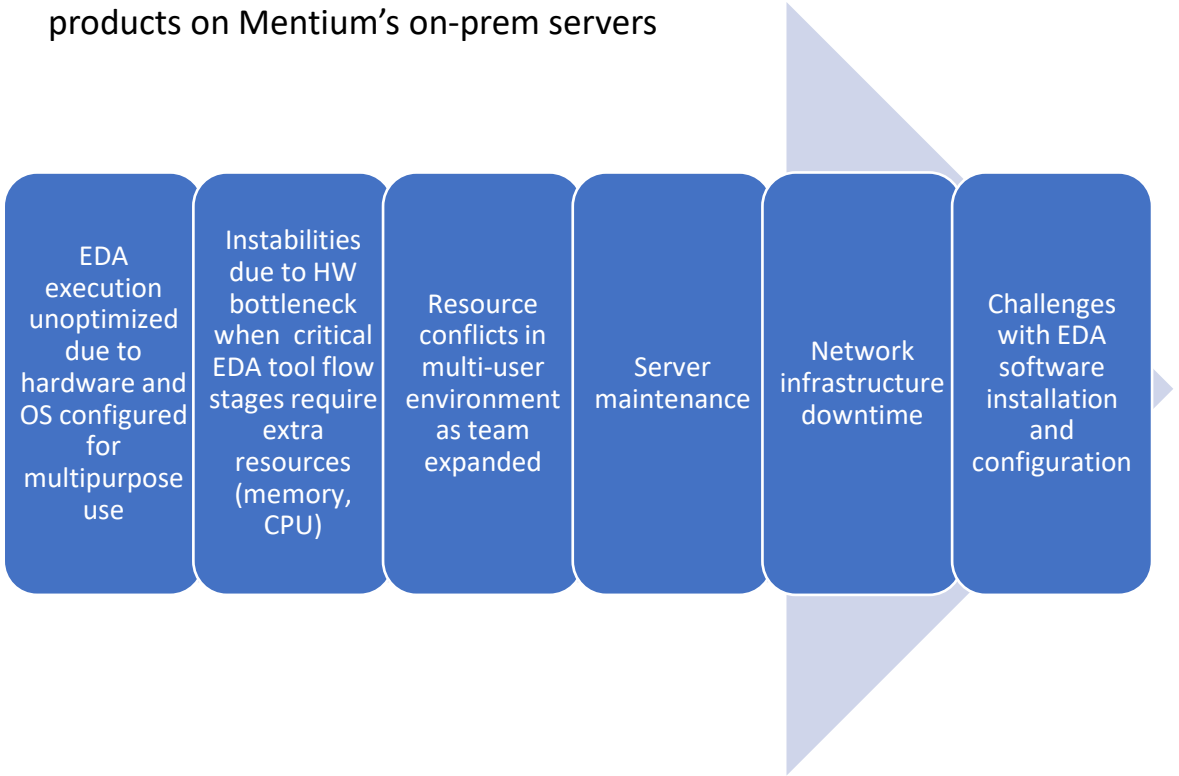
Tested model

Deploying
On Mentium's equipped Products

Synopsys Cloud for Chip Development

Challenges

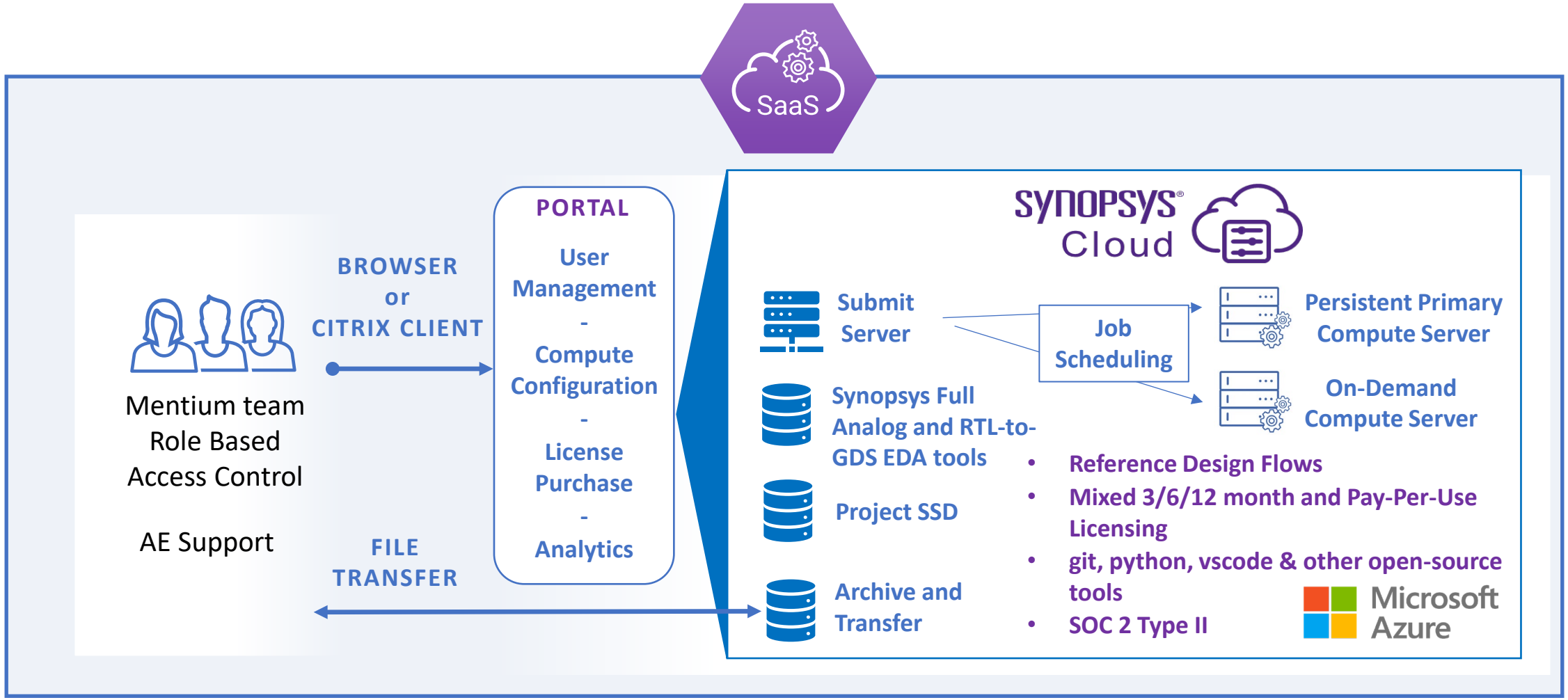
First iteration of Mentium DVM accelerator was done using Synopsys EDA products on Mentium's on-prem servers



Benefits of moving to Synopsys Cloud Saas

- Setup and operational in 2 days
- Optimized environment for EDA workloads (compute/storage)
- Flexible, scalable, self-serve license management (on-demand tool access)
- Streamlined onboarding and EDA tool support
- Built in job scheduler for effective resource management

Synopsys Cloud SaaS: End-to-End SoC Design Environment



MENTIUM TECHNOLOGIES

Development and Production Board Kit

Software SDK

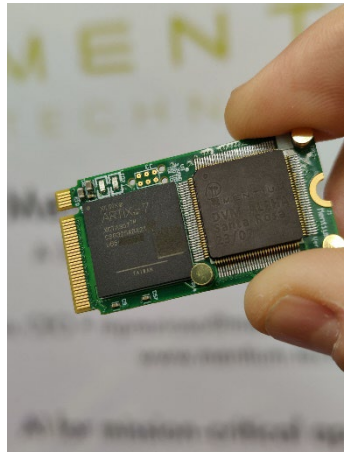
- Mentium Model Optimizer
  - ✓ Neural Net Analysis
  - ✓ Quantizing
  - ✓ Compiling
  - ✓ Profiling
- Python APIs

+

Linux Driver

+

Hardware m.2 PCIe - PCB



- Able to run main neural networks benchmarks and calculate the power consumed.
- Can be connected to any platform running Linux through an m.2 socket or PCIe adaptor.
- Accepts standard AI-model file formats such as ONNX, Tensorflow, PyTorch

To request: products@mentium.tech or follow the link

AI HARDWARE & EDGE AI SUMMIT

Wrap-up

MENTIUM TECHNOLOGIES

UNIQUE HYBRID APPROACH

SIMPLE INTEGRATION, COMPLEMENTARY TO ALL SOCs & MCUs

#1 IN MISSION-CRITICAL AI AT LOWEST EDGE POWER

Mirko Prezioso mprezioso@mentium.tech LinkedIn: