Workshop

Accelerating AI 2023 – Challenges and Opportunities in Cloud and Edge Computing

May 4th, 2023 (1pm-5pm EDT)

Virtual

Host: Yassine Hariri Senior Staff Scientist – AI/ML
CMC Microsystems

CMC MICROSYSTEMS

Accelerating AI 2023 – Challenges and Opportunities in Cloud and Edge Computing

Goal: Bring together experts from industry and academia to:

- Share the lasted trends and innovations
- Identify challenges and opportunities
- Explore collaboration opportunities
- Identify common infrastructure requirements

CMC

Topics of Accelerating AI 2023

- ML applications: Computer Vision, NLP, EDA and CAD…
- Novel AI HW: GPUs, FPGAs and Custom Accelerators
- Software stack: libraries, compilers, and ML frameworks
- ML Benchmarking on Emerging Hardware
- AI Latest trends in chip design and commercialization.

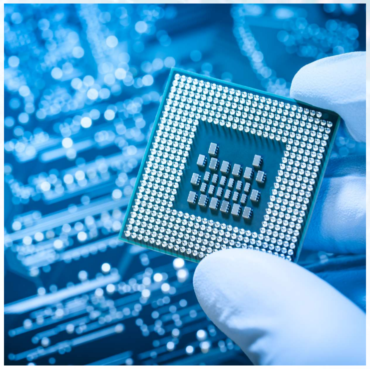
Agenda

Time	Presenter	Organization	Title
1:00 – 1:10	**Yassine Hariri**	**CMC Microsystems**	Welcome and opening remarks
1:10 – 1:30	**Rick O'Connor**	**OpenHW Group**	CORE-V Cores: Open-Source RISC-V Cores for Industry & Academia
1:30 – 1:50	**Gaurav Singh**	**Untether AI**	Energy Efficient AI Inference Acceleration with Untether AI
1:50 – 2:10	**Griffin Lacey**	**Nvidia**	Accelerating Transformers with FP8
2:10 – 2:30	**Davis Sawyer**	**Deeplite**	Running 2bit Quantized CNNs on Arm CPUs
2:30 – 2:40	**Break**		
2:40 – 3:00	**Andreas Moshovos**	**University of Toronto**	Capitalizing on a Decade of Machine Learning Accelerators: SW/HW Assists for Training and Inference
3:00 – 3:20	**Warren Gross**	**McGill University**	Standard Deviation-Based Quantization for Deep Neural Networks
3:20 – 3:40	**Nizar El Zarif**	**Polytechnique Montréal**	Polara: a RISCV multicore vector processor
3:40 – 4:00	**François Leduc-Primeau**	**Polytechnique Montréal**	Designing Robust DNN Models That Exploit Energy-Reliability Tradeoffs
4:00 – 4:30	**Open Discussion**		
4:30 PM	**Closing**		

CMC Microsystems

CMC Microsystems

- CMC provides services to simplify access and reduce cost to advanced technologies:
  - Microelectronics
  - Photonics
  - IoT and Edge AI
  - MEMS, Nanofabrication and Integration
  - Quantum Technologies



- Academic and Industrial Support
  - Not-for-profit founded in 1984
  - Enabling innovation in a network involving more than 10,000 academic and industry participants.

Fueling Innovation and Competitiveness Across Strategic Sectors

CMC User Network Technology Drivers

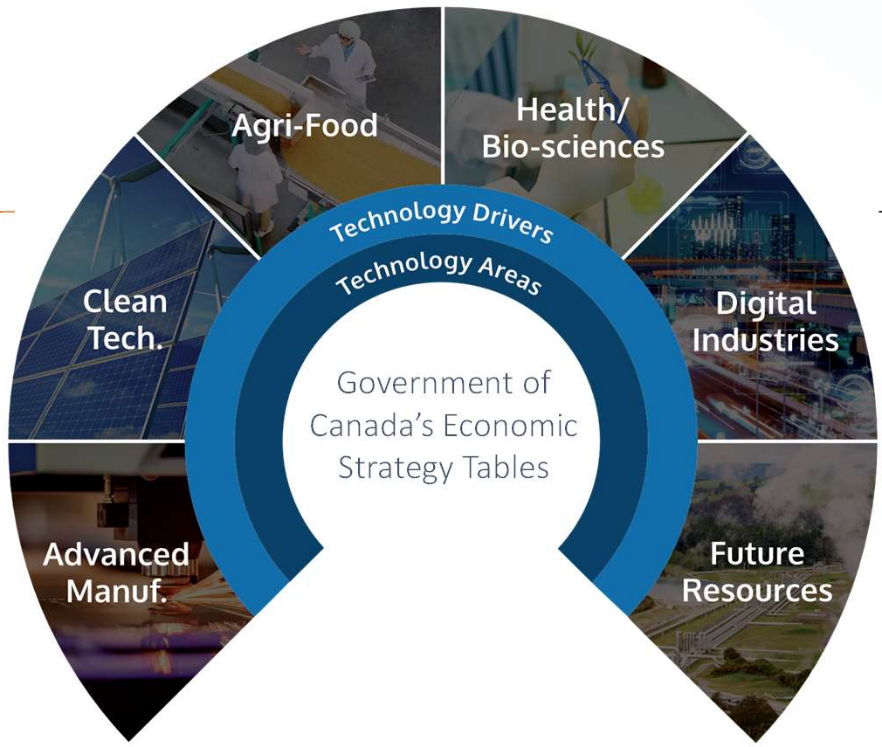
Pervasive Computing & AI

Ultra-High Speed Communication

Energy Management

Biophotonics, Bioelectronics

Industry 4.0



CMC User Network Technology Areas

Microelectronics

Photonics

IoT & Edge AI

MEMS, Nanofabrication and Integration

Quantum Technologies

Advanced Technologies Across all Strategic Sectors

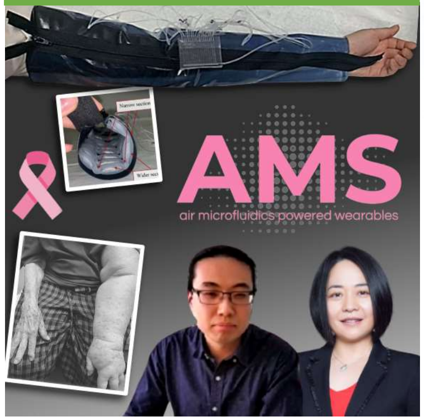
Clean Technology

Simon Fraser University
Dr. Byron Gates
Dr. Michael Paul

Advanced Manufacturing & Digital Industries

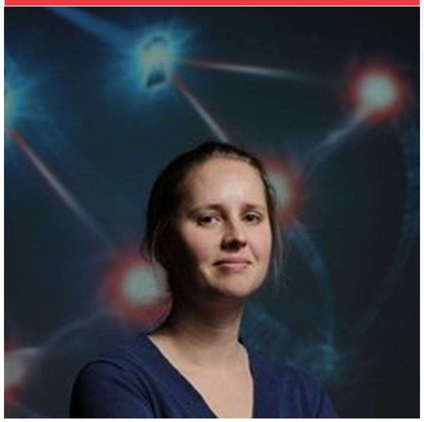
Université Laval
Dr. Leslie Rusch
Dr. Wei Shi

Health/Bio-sciences



University of Waterloo
Mr. Run Ze Gao
Dr. Carolyn Ren

A Quantum Leap in Cybersecurity

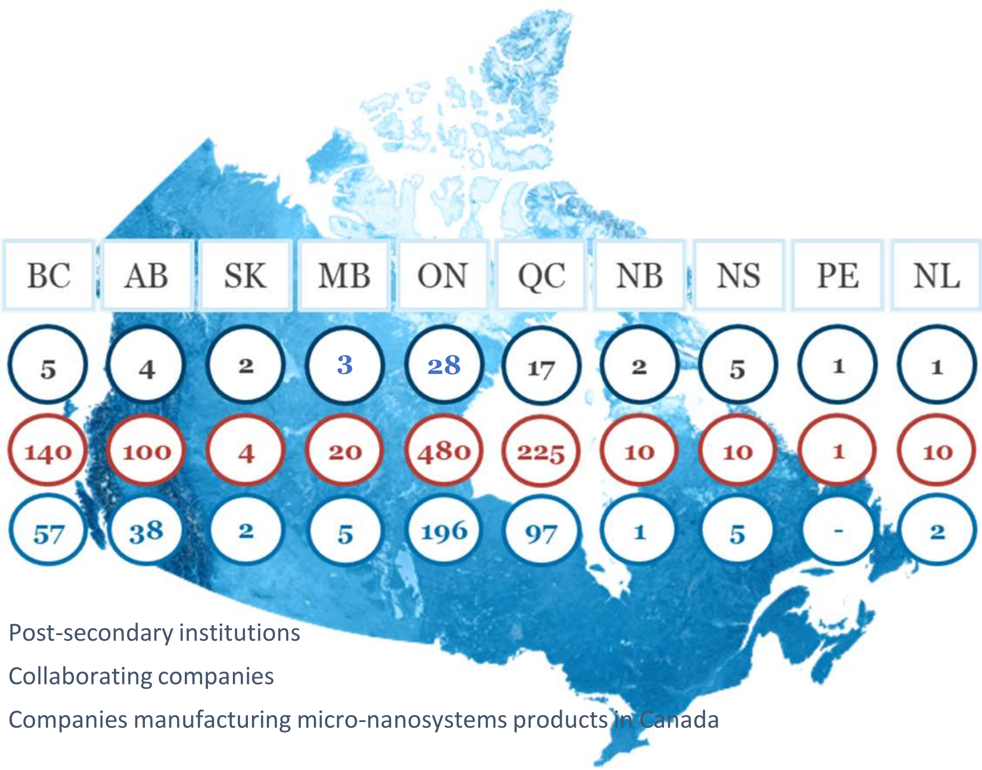


University of Ottawa
Dr. Anne Broadbent

www.CMC.ca/SuccessStories

Accelerating innovation across Canada

A Canada-wide collaboration between 68 universities/colleges to connect 10,000 academic participants with 1,000 companies to design, make and test micro-nanosystem prototypes.



	BC	AB	SK	MB	ON	QC	NB	NS	PE	NL
Post-secondary institutions	5	4	2	3	28	17	2	5	1	1
Collaborating companies	140	100	4	20	480	225	10	10	1	10
Companies manufacturing micro-nanosystems products in Canada	57	38	2	5	196	97	1	5	-	2

Enabling innovation and HQP skills development

1,370 connected professors

9,790 HQP benefitting

400 collaborations with industry

3,550 publications

125 awards

65 patents awarded

8 new startups

820 trained HQP moved to industry

CMC

CAD

State-of-the-art software for successful design

- Computer-Aided Design tools and design environments
- A secure, distributed private cloud for hosting
- User guides process design kits (PDKs), application notes, training materials, courses

FAB

Simple access and reduced cost for working prototypes

- Multi-Project Wafer (MPW) services through a global supply chain for
  - Microelectronics down to 12nm
  - Silicon photonics
  - MicroElectroMechanical Systems (MEMS)
  - Nanofabrication
- Expert assistance for first time right designs
- Packaging and assembly services

LAB

Tools for test and demonstration

- Platform technologies to speed your research
- Test equipment loans for short term needs
- Technical contract services including quantum coding
- Constructing research networks
- International partnerships for unique needs

CMC

CMC's advanced technology supply chain

Over 100 alliances in 17 countries

North America

CANADA
15 CAD · 15 FAB · 16 LAB
40 University MNT Labs

USA
18 CAD · 8 FAB · 6 LAB

Europe

EUROPE
6* Collaborative Organizations

AUSTRIA
1 FAB

IRELAND*
1 FAB

BELGIUM*
1 CAD · 2 FAB

NETHERLANDS
3 FAB

FINLAND
1 FAB

SWEDEN
1 CAD

FRANCE*
3 FAB

SWITZERLAND
1 FAB.

GERMANY*
1 CAD · 2 FAB

UK*
1 CAD

Asia

JAPAN
1 Collaborative Organization

SINGAPORE
2 FAB

SOUTH KOREA
1 Collaborative Organization

TAIWAN
2 FAB · 2 LAB
1 Collaborative Organization

Australia

AUSTRALIA
1 FAB
1 Collaborative Organization

- Collaborative organizations have similar mandates to accelerate advanced technology research and innovation.

CMC

Building a first of its kind national ecosystem to create critically needed semiconductor capability in Canada

https://www.cmc.ca/fabric/

One project, five activities:

1. Create capacity for the fabrication of semiconductor devices in Canada
   - Manufacturers enhance or develop new processes (a new product/service)
2. Accelerate R&D of IoT products and services by SMEs operating in all verticals
   - Growth activities for both supply and demand of semiconductors
3. Develop skills needed by industry
   - HQP training and reskilling for Canada's tech industry
4. Take quantum technologies to market
   - Enabling SMEs to assess quantum technologies and accelerate their adoption
5. Grow Canada's semiconductor ecosystem
   - Leveraging each other's strengths, developing ecosystem IP, attracting investment

AI Key Trends

Focus on edge computing

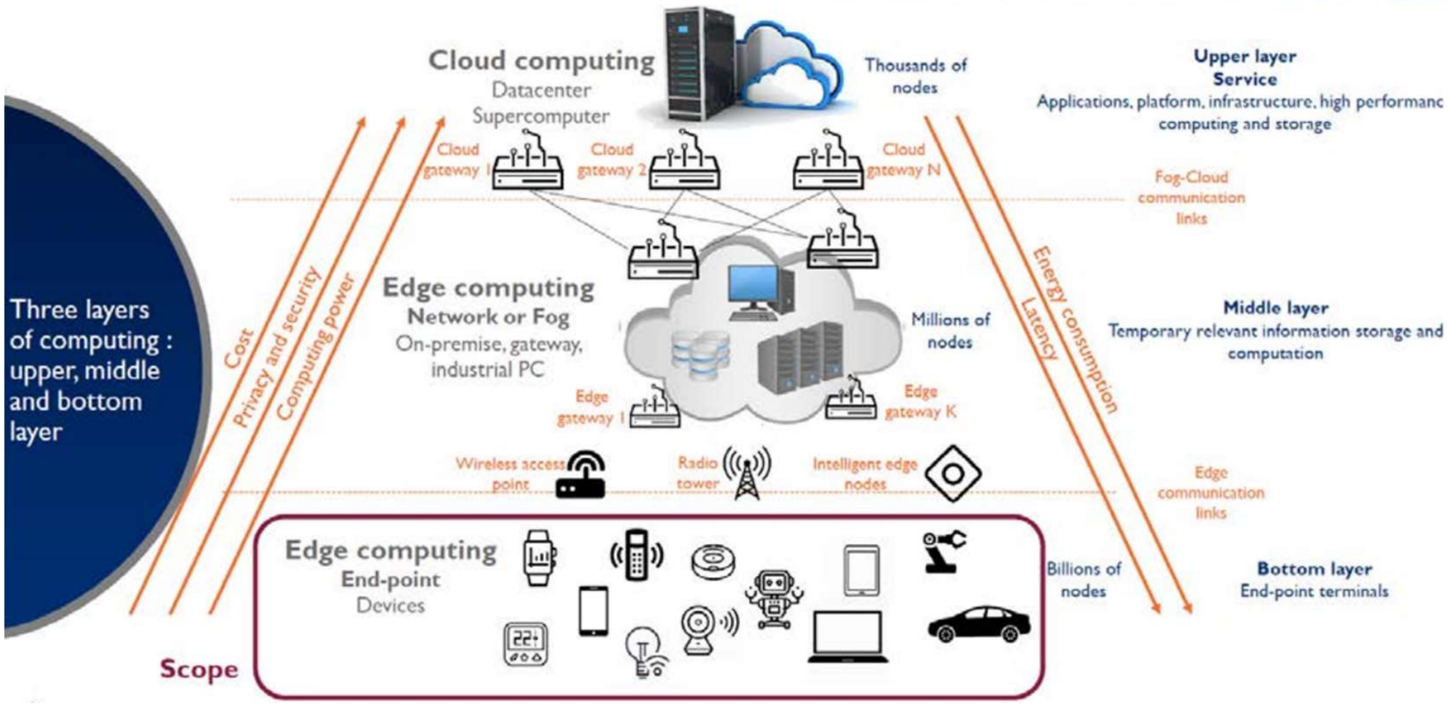
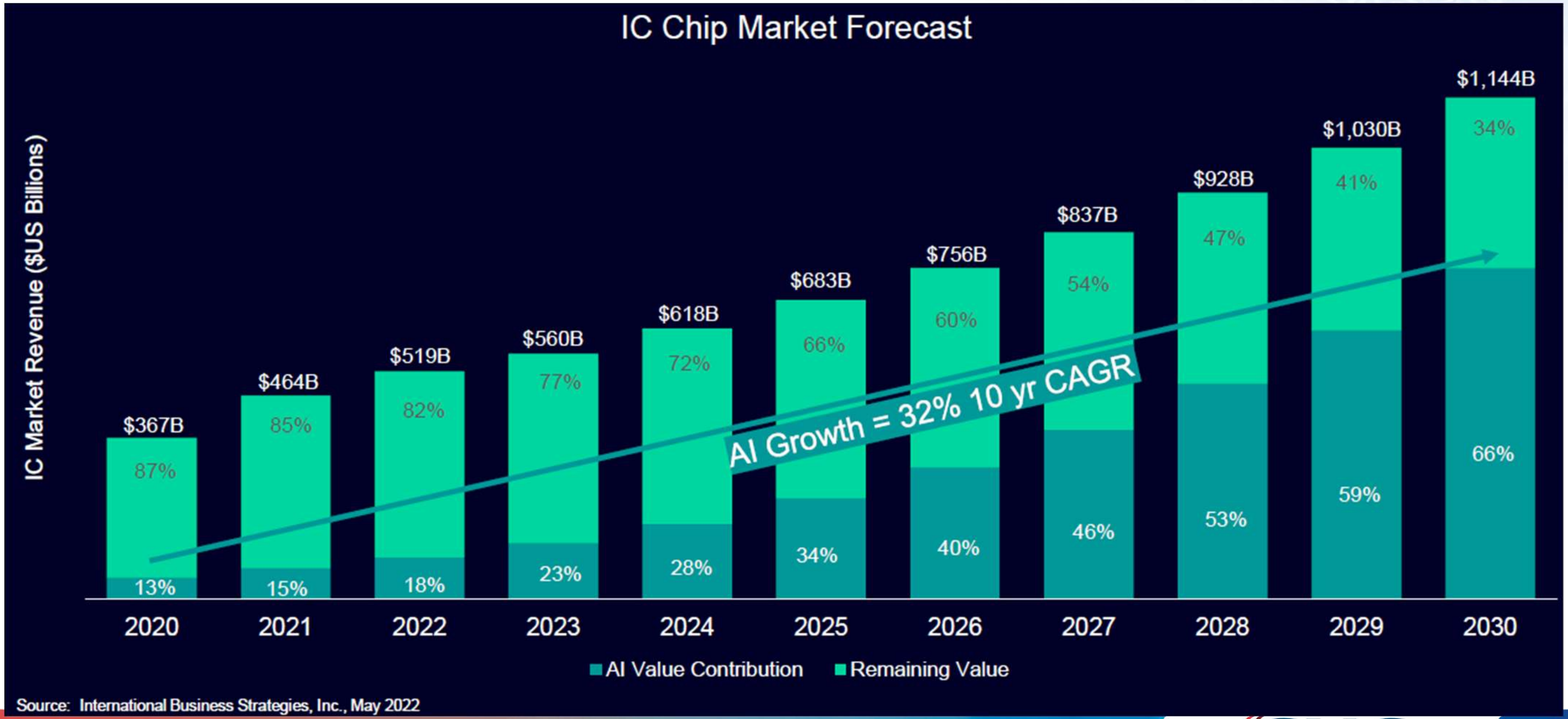


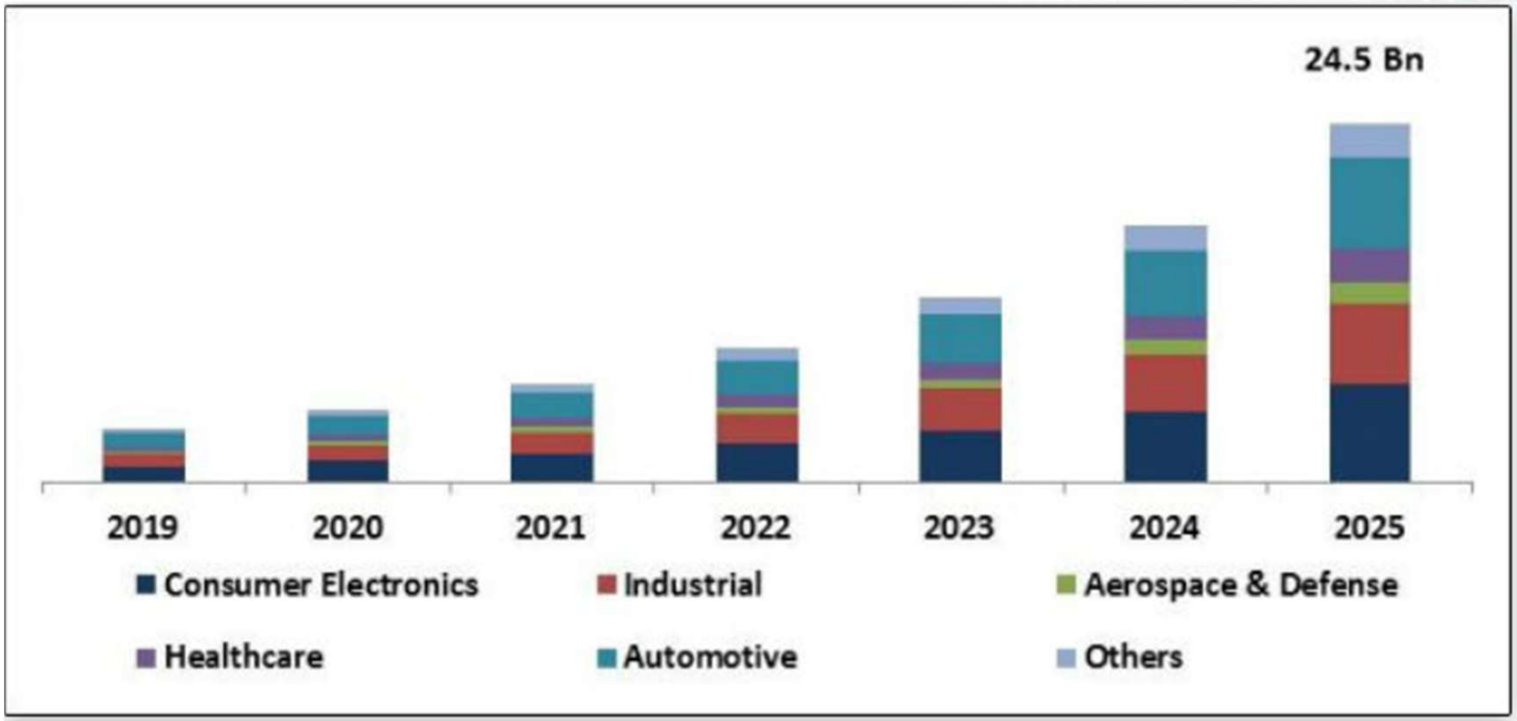
FIGURE 10: Focus on Edge Computing

Source: Computing and AI technologies for mobile and consumer applications, Yole Développement | www.yole.fr | ©2021

AI Value contribution to IC Chip market forecast to Grow 32% through 2030

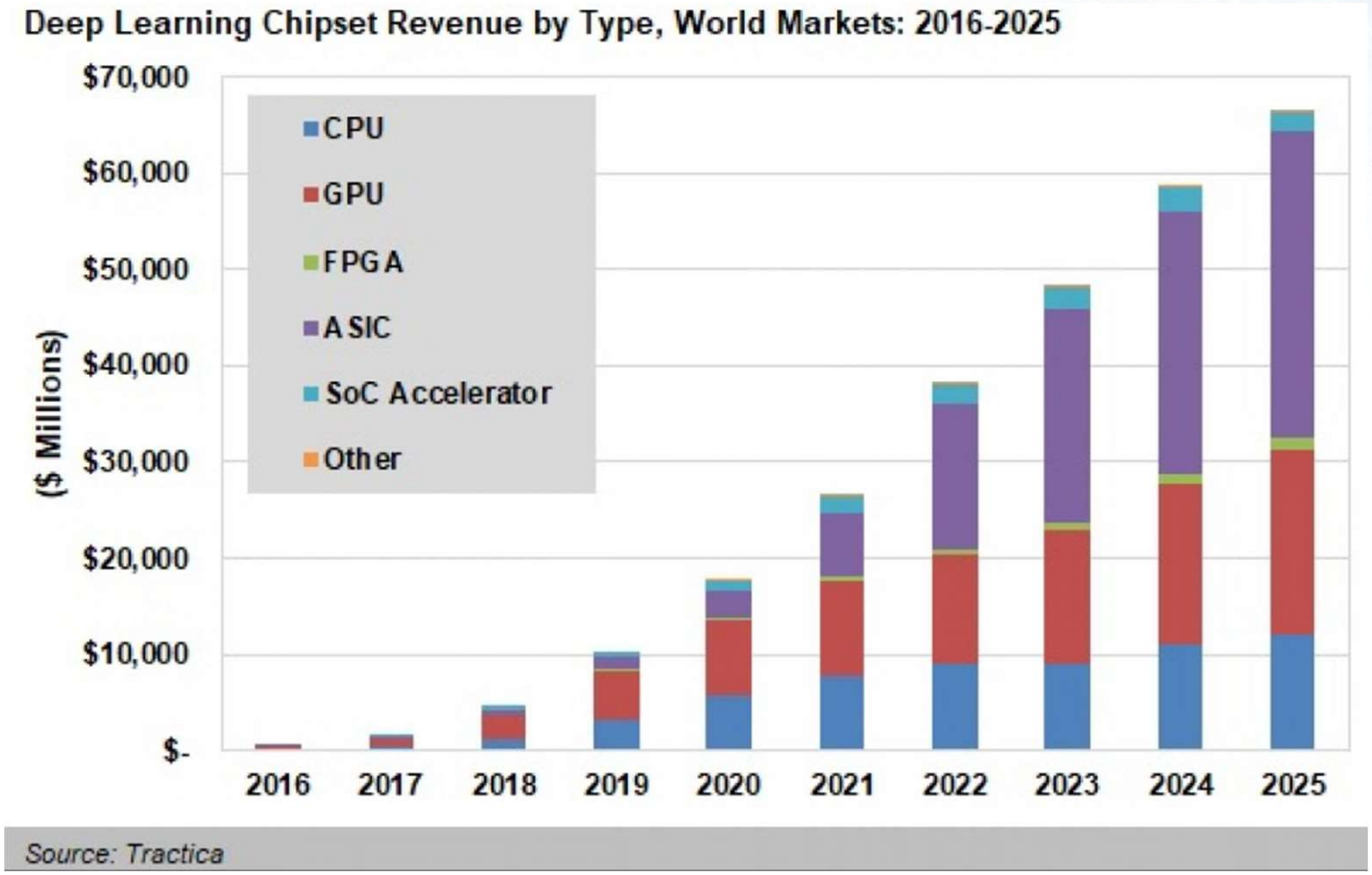


AI Chipset revenue by *market sector*

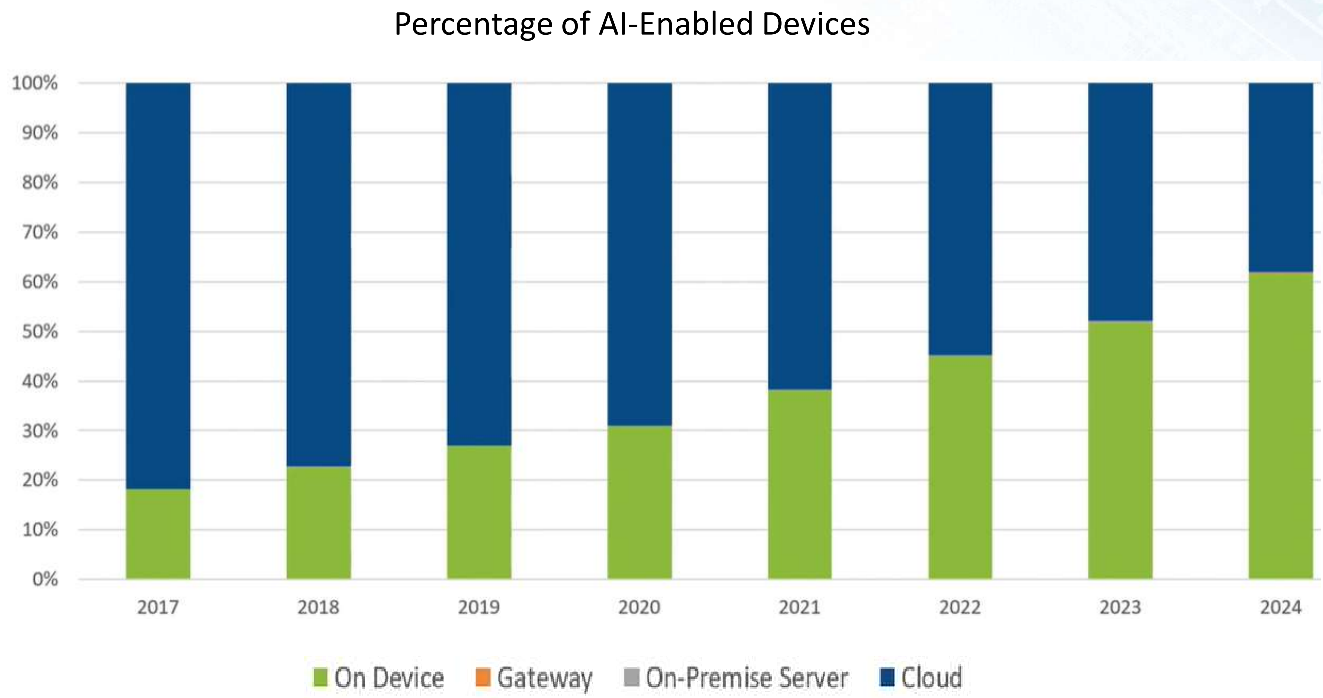


Source: kbv research

AI Chipset revenue by *type*



AI-Enabled Devices



Source: ABI Research – AI and ML

Neural Network Market (Edge Devices)

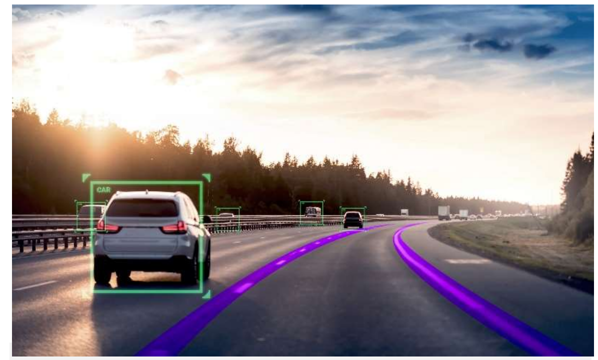
- Performance Requirements per Application are Increasing

- Facial detection
- Human activity recognition
- Always-on IoT / Smart Home
- Games/toys
- Voice control

<1 TOPS

- Augmented reality
- Surveillance
- Digital still cameras
- Facial recognition
- Automotive infotainment
- High-end smartphones
- Robotics / Drones
- Automotive RADAR / LiDAR

1 to 50 TOPS



- ADAS Vision/LiDAR/RADAR
- High end surveillance
- DTV
- High End Gaming
- Next gen augmented reality
- Microservers (inference)
- Data center (inference)

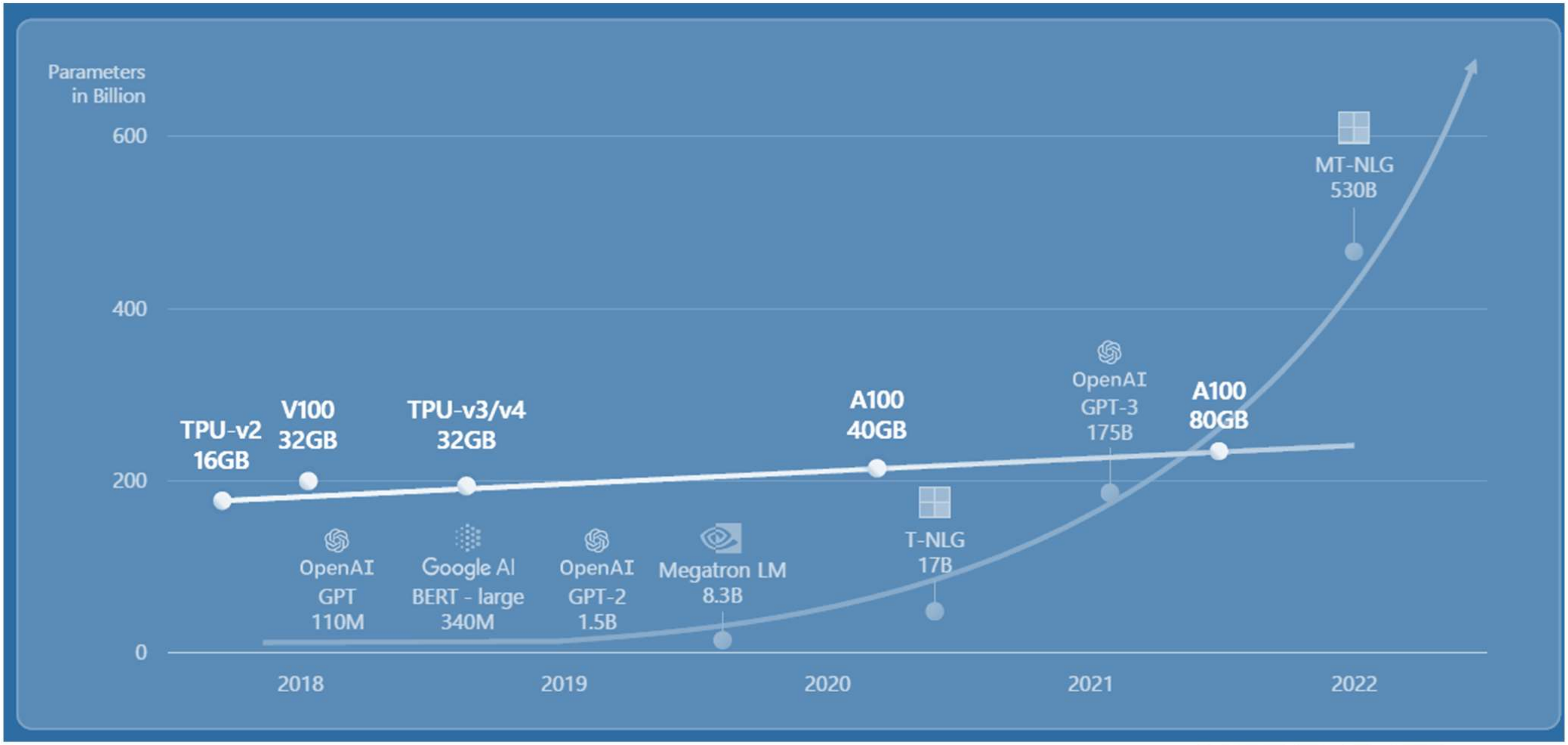
50 to 1000+ TOPS

Pierre Paulin, Synopsys, MPSoC 2022

Attributes of Edge Layers

Attribute	Cloud	CDN Edge	MEC Edge	Edge Gateway	Vehicle Edge	Smart Device
Latency to Sensor	100s of ms	10s of ms	10s of ms	1-5ms	100s of µs	10s of µs
Bandwidth to Sensor	10s of Mbps	100s of Mbps	~1Gbps	~1Gbps	10s of Gbps	10s of Gbps
Bandwidth to Cloud	100s of Gbps	10s of Gbps	~1Gbps	100s of Mbps	100s of Mbps	10s of Mbps
Available Energy	Megawatts	100s of KW	Killowatts	100s of Watts	100-10W	10-0.1W
Available Storage	Petabytes+	Petabytes	1-10 Terabytes	Terabytes	100s of GB	10s of GB
Security	Cloud	Cloud-like	Telco-like	Telco-like	Automotive?	Lightweight
Reliability / Resilience	Good	Good	Excellent	Good	Fair	Simplex
Scalability	Excellent	Very good	OK	Limited	Very limited	Constrained
Data Gravity	Remote	Regional	Neighborhood	Local	Local	Colocated
Programming Environment	Rich&Familiar	Familiar	Specialized	Specialized	Constrained	Primitive

Data bound - > Compute bound



Edge AI Environment

Macro Scan	Risks/Uncertainties	Challenges	Desired Figures of Merits FOM

Compute intensive AI workloads are migrating to the edge:
• Lower latency, bandwidth saving, connectivity
• Data privacy and security,
• Autonomy and reliability
Increasing demand for edge AI HW that balances performance, power, efficiency, and cost creating opportunities for:
• Domain specific GP neural processors unit
    • Complex multi level memory hierarchy
    • Massive multi level parallelism
    • Highly efficient ML mapping tools
Non-conventional architectures such as:
    • Quantum computing
    • Spiking neural processors (Neuromorphic computing)
    • Analog computing
	Canada excels at AI software but risks lagging on innovative Edge AI Hardware and trained HQP	• Increasing complexity and continued innovation in AI algorithms
		• High Computational Complexity
		• Small Memory Footprint and DRAM bandwidth
		• Low power consumption
		• Need highly efficient ML mapping tools
		• Trustworthiness: Safety, Security, resiliency, reliability and privacy
			• Smart Device
			    • 0.1W-10W
			    • 1-10 TOPS
			• Vehicle Edge
			    • 10W-100W
			    • 10 to 50 TOPS
			• Edge gateway
			    • 100s of Watts
			    • 50 to 1000+ TOPS
			
			We need to focus on meaningful FOM:
			• Inferences/sec/mm2
			• Inference/sec/W
			• Bandwidth/infer

CMC Services in Artificial Intelligence: Platforms

Build

AI Training

FPGA/GPU CLUSTER

AI Training
- Cerebro: 6 Alveo FPGAs
- Genisys: 6 V100 GPUs
- Synergy: 2 GPU 2 FPGA
- HW emulation, acceleration, validation
- FPGA Prototyping and SW bring-up

ATLAS 800 CLUSTER

AI Training
- 32 Ascend 910
- 16 Kunpeng 920
- 8 PFLOPS@FP16
- CANN computing architecture
- SW stack allowing 32 simultaneous users

Optimize

AI Optimization

AI-DRIVEN OPTIMIZER

AI Optimization
Deep Neural Networks by making them faster, smaller and energy-efficient from cloud to the edge

Deploy

AI inference

UNTETHER AI

- 2 PetaOps of Performance in a Single Card

CORE-V

32- and 64-bit open-source RISC-V cores
- Vector scalar processor
- On-chip eFPGA
- Barrel RISC-V

Atlas 200 DK AI Developer Kit
- Custom AI acc.

CMC

Workshop Agenda

Agenda

Time	Presenter	Organization	Title
1:00 – 1:10	**Yassine Hariri**	**CMC Microsystems**	Welcome and opening remarks
1:10 – 1:30	**Rick O'Connor**	**OpenHW Group**	CORE-V Cores: Open-Source RISC-V Cores for Industry & Academia
1:30 – 1:50	**Gaurav Singh**	**Untether AI**	Energy Efficient AI Inference Acceleration with Untether AI
1:50 – 2:10	**Griffin Lacey**	**Nvidia**	Accelerating Transformers with FP8
2:10 – 2:30	**Davis Sawyer**	**Deeplite**	Running 2bit Quantized CNNs on Arm CPUs
2:30 – 2:40	**Break**		
2:40 – 3:00	**Andreas Moshovos**	**University of Toronto**	Capitalizing on a Decade of Machine Learning Accelerators: SW/HW Assists for Training and Inference
3:00 – 3:20	**Warren Gross**	**McGill University**	Standard Deviation-Based Quantization for Deep Neural Networks
3:20 – 3:40	**Nizar El Zarif**	**Polytechnique Montréal**	Polara: a RISCV multicore vector processor
3:40 – 4:00	**François Leduc-Primeau**	**Polytechnique Montréal**	Designing Robust DNN Models That Exploit Energy-Reliability Tradeoffs
4:00 – 4:30	**Open Discussion**		
4:30 PM	**Closing**		

CMC

Agenda

Time	Presenter	Organization	Title
1:00 – 1:10	Yassine Hariri	CMC Microsystems	Welcome and opening remarks
1:10 – 1:30	Rick O'Connor	OpenHW Group	CORE-V Cores: Open-Source RISC-V Cores for Industry & Academia
1:30 – 1:50	Gaurav Singh	Untether AI	Energy Efficient AI Inference Acceleration with Untether AI
1:50 – 2:10	Griffin Lacey	Nvidia	Accelerating Transformers with FP8
2:10 – 2:30	Davis Sawyer	Deeplite	Running 2bit Quantized CNNs on Arm CPUs
2:30 – 2:40	Break		
2:40 – 3:00	Andreas Moshovos	University of Toronto	Capitalizing on a Decade of Machine Learning Accelerators: SW/HW Assists for Training and Inference
3:00 – 3:20	Warren Gross	McGill University	Standard Deviation-Based Quantization for Deep Neural Networks
3:20 – 3:40	Nizar El Zarif	Polytechnique Montréal	Polara: a RISCV multicore vector processor
3:40 – 4:00	François Leduc-Primeau	Polytechnique Montréal	Designing Robust DNN Models That Exploit Energy-Reliability Tradeoffs
4:00 – 4:30	Open Discussion		
4:30 PM	Closing		

Agenda

Time	Presenter	Organization	Title
1:00 – 1:10	**Yassine Hariri**	**CMC Microsystems**	Welcome and opening remarks
1:10 – 1:30	**Rick O'Connor**	**OpenHW Group**	CORE-V Cores: Open-Source RISC-V Cores for Industry & Academia
1:30 – 1:50	**Gaurav Singh**	**Untether AI**	Energy Efficient AI Inference Acceleration with Untether AI
1:50 – 2:10	**Griffin Lacey**	**Nvidia**	Accelerating Transformers with FP8
2:10 – 2:30	**Davis Sawyer**	**Deeplite**	Running 2bit Quantized CNNs on Arm CPUs
2:30 – 2:40	**Break**		
2:40 – 3:00	**Andreas Moshovos**	**University of Toronto**	Capitalizing on a Decade of Machine Learning Accelerators: SW/HW Assists for Training and Inference
3:00 – 3:20	**Warren Gross**	**McGill University**	Standard Deviation-Based Quantization for Deep Neural Networks
3:20 – 3:40	**Nizar El Zarif**	**Polytechnique Montréal**	Polara: a RISCV multicore vector processor
3:40 – 4:00	**François Leduc-Primeau**	**Polytechnique Montréal**	Designing Robust DNN Models That Exploit Energy-Reliability Tradeoffs
4:00 – 4:30	**Open Discussion**		
4:30 PM	**Closing**		

CMC

Agenda

Time	Presenter	Organization	Title
1:00 – 1:10	**Yassine Hariri**	**CMC Microsystems**	Welcome and opening remarks
1:10 – 1:30	**Rick O'Connor**	**OpenHW Group**	CORE-V Cores: Open-Source RISC-V Cores for Industry & Academia
1:30 – 1:50	**Gaurav Singh**	**Untether AI**	Energy Efficient AI Inference Acceleration with Untether AI
1:50 – 2:10	**Griffin Lacey**	**Nvidia**	Accelerating Transformers with FP8
2:10 – 2:30	**Davis Sawyer**	**Deeplite**	Running 2bit Quantized CNNs on Arm CPUs
2:30 – 2:40	**Break**		
2:40 – 3:00	**Andreas Moshovos**	**University of Toronto**	Capitalizing on a Decade of Machine Learning Accelerators: SW/HW Assists for Training and Inference
3:00 – 3:20	**Warren Gross**	**McGill University**	Standard Deviation-Based Quantization for Deep Neural Networks
3:20 – 3:40	**Nizar El Zarif**	**Polytechnique Montréal**	Polara: a RISCV multicore vector processor
3:40 – 4:00	**François Leduc-Primeau**	**Polytechnique Montréal**	Designing Robust DNN Models That Exploit Energy-Reliability Tradeoffs
4:00 – 4:30	**Open Discussion**		
4:30 PM	**Closing**		

Agenda

Time	Presenter	Organization	Title
1:00 – 1:10	**Yassine Hariri**	**CMC Microsystems**	Welcome and opening remarks
1:10 – 1:30	**Rick O'Connor**	**OpenHW Group**	CORE-V Cores: Open-Source RISC-V Cores for Industry & Academia
1:30 – 1:50	**Gaurav Singh**	**Untether AI**	Energy Efficient AI Inference Acceleration with Untether AI
1:50 – 2:10	**Griffin Lacey**	**Nvidia**	Accelerating Transformers with FP8
2:10 – 2:30	**Davis Sawyer**	**Deeplite**	Running 2bit Quantized CNNs on Arm CPUs
2:30 – 2:40	Break		
2:40 – 3:00	**Andreas Moshovos**	**University of Toronto**	Capitalizing on a Decade of Machine Learning Accelerators: SW/HW Assists for Training and Inference
3:00 – 3:20	**Warren Gross**	**McGill University**	Standard Deviation-Based Quantization for Deep Neural Networks
3:20 – 3:40	**Nizar El Zarif**	**Polytechnique Montréal**	Polara: a RISCV multicore vector processor
3:40 – 4:00	**François Leduc-Primeau**	**Polytechnique Montréal**	Designing Robust DNN Models That Exploit Energy-Reliability Tradeoffs
4:00 – 4:30	**Open Discussion**		
4:30 PM	**Closing**		

Agenda

Time	Presenter	Organization	Title
1:00 – 1:10	Yassine Hariri	CMC Microsystems	Welcome and opening remarks
1:10 – 1:30	Rick O'Connor	OpenHW Group	CORE-V Cores: Open-Source RISC-V Cores for Industry & Academia
1:30 – 1:50	Gaurav Singh	Untether AI	Energy Efficient AI Inference Acceleration with Untether AI
1:50 – 2:10	Griffin Lacey	Nvidia	Accelerating Transformers with FP8
2:10 – 2:30	Davis Sawyer	Deeplite	Running 2bit Quantized CNNs on Arm CPUs
2:30 – 2:40	Break		
2:40 – 3:00	Andreas Moshovos	University of Toronto	Capitalizing on a Decade of Machine Learning Accelerators: SW/HW Assists for Training and Inference
3:00 – 3:20	Warren Gross	McGill University	Standard Deviation-Based Quantization for Deep Neural Networks
3:20 – 3:40	Nizar El Zarif	Polytechnique Montréal	Polara: a RISCV multicore vector processor
3:40 – 4:00	François Leduc-Primeau	Polytechnique Montréal	Designing Robust DNN Models That Exploit Energy-Reliability Tradeoffs
4:00 – 4:30	Open Discussion		
4:30 PM	Closing		

CMC

Agenda

Time	Presenter	Organization	Title
1:00 – 1:10	**Yassine Hariri**	**CMC Microsystems**	Welcome and opening remarks
1:10 – 1:30	**Rick O'Connor**	**OpenHW Group**	CORE-V Cores: Open-Source RISC-V Cores for Industry & Academia
1:30 – 1:50	**Gaurav Singh**	**Untether AI**	Energy Efficient AI Inference Acceleration with Untether AI
1:50 – 2:10	**Griffin Lacey**	**Nvidia**	Accelerating Transformers with FP8
2:10 – 2:30	**Davis Sawyer**	**Deeplite**	Running 2bit Quantized CNNs on Arm CPUs
2:30 – 2:40	**Break**		
2:40 – 3:00	**Andreas Moshovos**	**University of Toronto**	Capitalizing on a Decade of Machine Learning Accelerators: SW/HW Assists for Training and Inference
3:00 – 3:20	**Warren Gross**	**McGill University**	Standard Deviation-Based Quantization for Deep Neural Networks
3:20 – 3:40	**Nizar El Zarif**	**Polytechnique Montréal**	Polara: a RISCV multicore vector processor
3:40 – 4:00	**François Leduc-Primeau**	**Polytechnique Montréal**	Designing Robust DNN Models That Exploit Energy-Reliability Tradeoffs
4:00 – 4:30	**Open Discussion**		
4:30 PM	**Closing**		

CMC

Agenda

Time	Presenter	Organization	Title
1:00 – 1:10	**Yassine Hariri**	**CMC Microsystems**	Welcome and opening remarks
1:10 – 1:30	**Rick O'Connor**	**OpenHW Group**	CORE-V Cores: Open-Source RISC-V Cores for Industry & Academia
1:30 – 1:50	**Gaurav Singh**	**Untether AI**	Energy Efficient AI Inference Acceleration with Untether AI
1:50 – 2:10	**Griffin Lacey**	**Nvidia**	Accelerating Transformers with FP8
2:10 – 2:30	**Davis Sawyer**	**Deeplite**	Running 2bit Quantized CNNs on Arm CPUs
2:30 – 2:40	**Break**		
2:40 – 3:00	**Andreas Moshovos**	**University of Toronto**	Capitalizing on a Decade of Machine Learning Accelerators: SW/HW Assists for Training and Inference
3:00 – 3:20	**Warren Gross**	**McGill University**	Standard Deviation-Based Quantization for Deep Neural Networks
3:20 – 3:40	**Nizar El Zarif**	**Polytechnique Montréal**	Polara: a RISCV multicore vector processor
3:40 – 4:00	**François Leduc-Primeau**	**Polytechnique Montréal**	Designing Robust DNN Models That Exploit Energy-Reliability Tradeoffs
4:00 – 4:30	**Open Discussion**		
4:30 PM	**Closing**		

Agenda

Time	Presenter	Organization	Title
1:00 – 1:10	**Yassine Hariri**	**CMC Microsystems**	Welcome and opening remarks
1:10 – 1:30	**Rick O'Connor**	**OpenHW Group**	CORE-V Cores: Open-Source RISC-V Cores for Industry & Academia
1:30 – 1:50	**Gaurav Singh**	**Untether AI**	Energy Efficient AI Inference Acceleration with Untether AI
1:50 – 2:10	**Griffin Lacey**	**Nvidia**	Accelerating Transformers with FP8
2:10 – 2:30	**Davis Sawyer**	**Deeplite**	Running 2bit Quantized CNNs on Arm CPUs
2:30 – 2:40	**Break**		
2:40 – 3:00	**Andreas Moshovos**	**University of Toronto**	Capitalizing on a Decade of Machine Learning Accelerators: SW/HW Assists for Training and Inference
3:00 – 3:20	**Warren Gross**	**McGill University**	Standard Deviation-Based Quantization for Deep Neural Networks
3:20 – 3:40	**Nizar El Zarif**	**Polytechnique Montréal**	Polara: a RISCV multicore vector processor
3:40 – 4:00	**François Leduc-Primeau**	**Polytechnique Montréal**	Designing Robust DNN Models That Exploit Energy-Reliability Tradeoffs
4:00 – 4:30	**Open Discussion**		
4:30 PM	**Closing**		

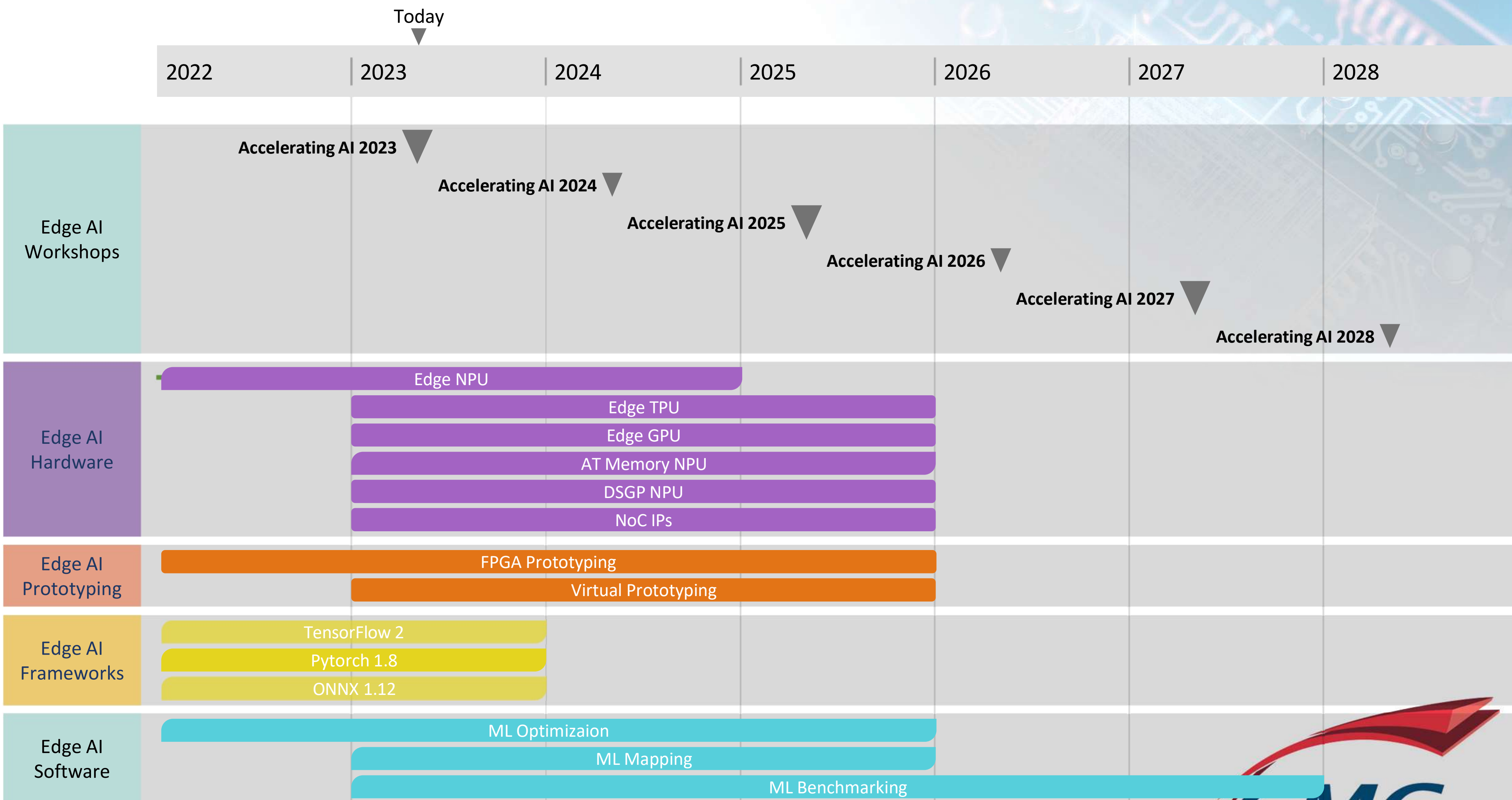
Agenda

Time	Presenter	Organization	Title
1:00 – 1:10	**Yassine Hariri**	**CMC Microsystems**	Welcome and opening remarks
1:10 – 1:30	**Rick O'Connor**	**OpenHW Group**	CORE-V Cores: Open-Source RISC-V Cores for Industry & Academia
1:30 – 1:50	**Gaurav Singh**	**Untether AI**	Energy Efficient AI Inference Acceleration with Untether AI
1:50 – 2:10	**Griffin Lacey**	**Nvidia**	Accelerating Transformers with FP8
2:10 – 2:30	**Davis Sawyer**	**Deeplite**	Running 2bit Quantized CNNs on Arm CPUs
2:30 – 2:40	**Break**		
2:40 – 3:00	**Andreas Moshovos**	**University of Toronto**	Capitalizing on a Decade of Machine Learning Accelerators: SW/HW Assists for Training and Inference
3:00 – 3:20	**Warren Gross**	**McGill University**	Standard Deviation-Based Quantization for Deep Neural Networks
3:20 – 3:40	**Nizar El Zarif**	**Polytechnique Montréal**	Polara: a RISCV multicore vector processor
3:40 – 4:00	**François Leduc-Primeau**	**Polytechnique Montréal**	Designing Robust DNN Models That Exploit Energy-Reliability Tradeoffs
4:00 – 4:30	**Open Discussion**		
4:30 PM	**Closing**		

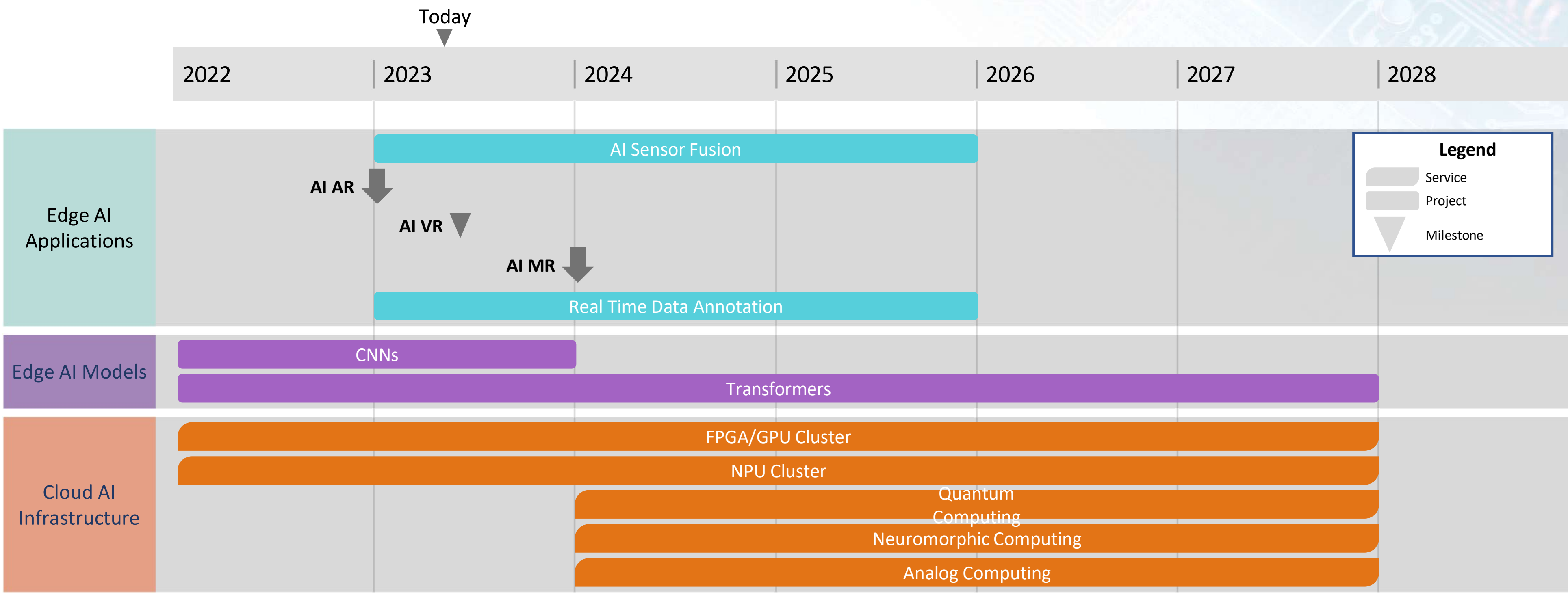
Panel Session

Panel Session: Challenges and Opportunities in Cloud and Edge Computing

What do you see as the most important challenges and opportunities in Cloud and Edge Computing?



CMC MICROSYSTEMS



Today
2022
2023
2024
2025
2026
2027
2028
Edge AI Applications
AI Sensor Fusion
AI AR
AI VR
AI MR
Real Time Data Annotation
Legend
Service
Project
Milestone
Edge AI Models
CNNs
Transformers
Cloud AI Infrastructure
FPGA/GPU Cluster
NPU Cluster
Quantum Computing
Neuromorphic Computing
Analog Computing