
AWS
re:Invent
NOV. 27 – DEC. 1, 2023 | LAS VEGAS, NV

AIM311-S

SPONSORED BY INTEL

AI acceleration at the edge

Jonathan Lee
Chief Product Officer
ai.io

Alex White
Sales Applications Engineer
Intel

Michael Kleiner
VP Edge AI Solutions
OnLogic

Mohan Potheri
Cloud Solutions Architect
Intel

Allan Gagnon
Senior Solutions Architect
Arduino

Agenda

1. Top Business Outcome for Edge AI
2. Bringing AI Everywhere
3. Compelling Edge AI Use cases
   Edge Computer vision booth demo
4. Real World Successes
   Featuring OnLogic, Arduino & ai.io
5. 4th Generation Intel Xeon for AI
6. The AI Pipeline of processing
7. Start building your AI solutions today

Edge Compute and AI: **Top Trends We're Seeing**

- By year-end 2026, 70% of large enterprises will have a documented strategy for edge computing, compared to fewer than 10% in 2023.
- CEO use of the term "digital" in their top business priorities has roughly doubled from 2018 to 2022.
- Digital data production at the edge is growing exponentially, creating the opportunity for deeper analysis, automation, AI /ML.*

Edge Compute and AI: **Top Trends We're Seeing**

"Enterprises need a strategy so they can overcome and learn from challenges, choose technologies, manage costs and enable successful digital transformation."

Gartner

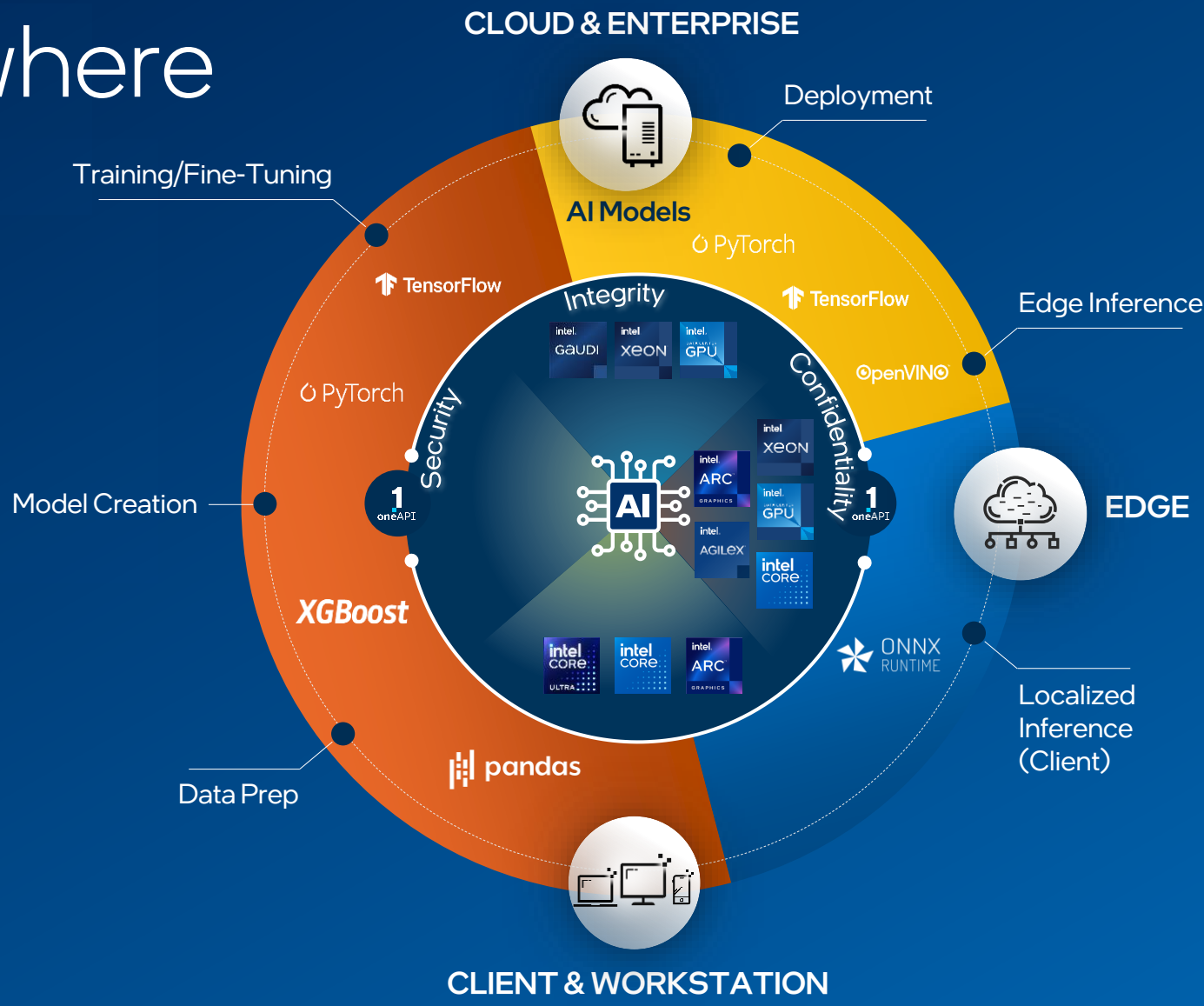
Building an Edge Computing Strategy,
Thomas Bittman, April 12, 2023

Bringing AI to the Edge

it starts with intel

AI Continuum

Bringing AI Everywhere



Note: Intel Core Ultra integrates NPU low power inference engine from Meteor Lake onwards.

Bringing AI Everywhere at the Edge

Move from development to deployment **fast**

Scalable hybrid AI approach

Leverage automatic processing of an AI workload using available/targeted system resources and accelerators on the edge, or in the cloud. OpenVINO and oneAPI allow developers to seamlessly transition between edge-to-cloud with a single code base.

Open tools to speed deployment

The OpenVINO toolkit optimizes deep learning inference deployment for hundreds of pretrained models across multiple Intel platforms. The Intel Geti Platform allows domain experts and data scientists to quickly build and train AI models.

Hundreds of deployment-ready solutions

Intel solutions include defect detection, worker and public safety, robotics, supply chain management, imaging diagnostics, and enhanced service delivery.

Bringing **AI Everywhere** to Unlock New Levels of **Innovation**









Education	Health	Finance	Retail	Government	Energy	Automotive	Manufacturing	Telco
Teacher Assistant Student Study Buddy Parent Chat Portal	Drug Discovery Doctor Co-pilot Patient Family Chatbot	Algorithmic Trading Customer Portfolio Assistant Risk / Credit Assessment	Product Promotion Customer Interface and Sentiment Tool Image Shopping Aid	Gov Services Chatbot Document Search Summarization Live Language Translation	Energy Consumption Forecasting Operational Performance Energy Trading Assistant	Autonomous Car Development Multi-language in car aid Supply Chain Optimization	Factory Automation Predictive Maintenance Precision Agriculture	Personalized Customer Services Network Automation Operational Performance

Secure the Edge

it starts with intel

Layering Security Foundation for Solution Developers

Secure at Boot

- Always boot to safe default configurations
- Secure boot all the way from the HW up to applications
- Periodic monitoring of device integrity through remote attestation

HW RoT and Secure Boot,
Trusted Platform Module (TPM) ,
Intel Platform Trust Technology (Intel PTT)

Seamless to Deploy

- Zero touch onboarding
- Secure configuration updates at runtime
- Lift and shift secure application across multiple SW profiles

FIDO Device Onboard (FDO)

Protected at Runtime

- VM and process memory is isolated via HW mechanisms, and protected by role-based access control
- Interfaces to hardware devices and controlled

Intel Virtualization Technology (Intel VT)
Intel Total Memory Encryption (Intel TME)
Intel Software Guard Extensions (Intel SGX)
Intel Trust Domain Extensions (Intel TDX)

Encrypted at Rest

- Root of trust stored in tamper resistant hardware (TPM, HSMs)
- Storage is encrypted with HW unique keys
- System software sill only decrypt authentic hardware

Trusted Platform Module (TPM)
Intel Platform Trust Technology (Intel PTT)

Edge Computer Vision

it starts with

intel.

Where Will Computer Vision Applications Disrupt and Innovate?

On-demand customer experiences

Preventative maintenance

Automation

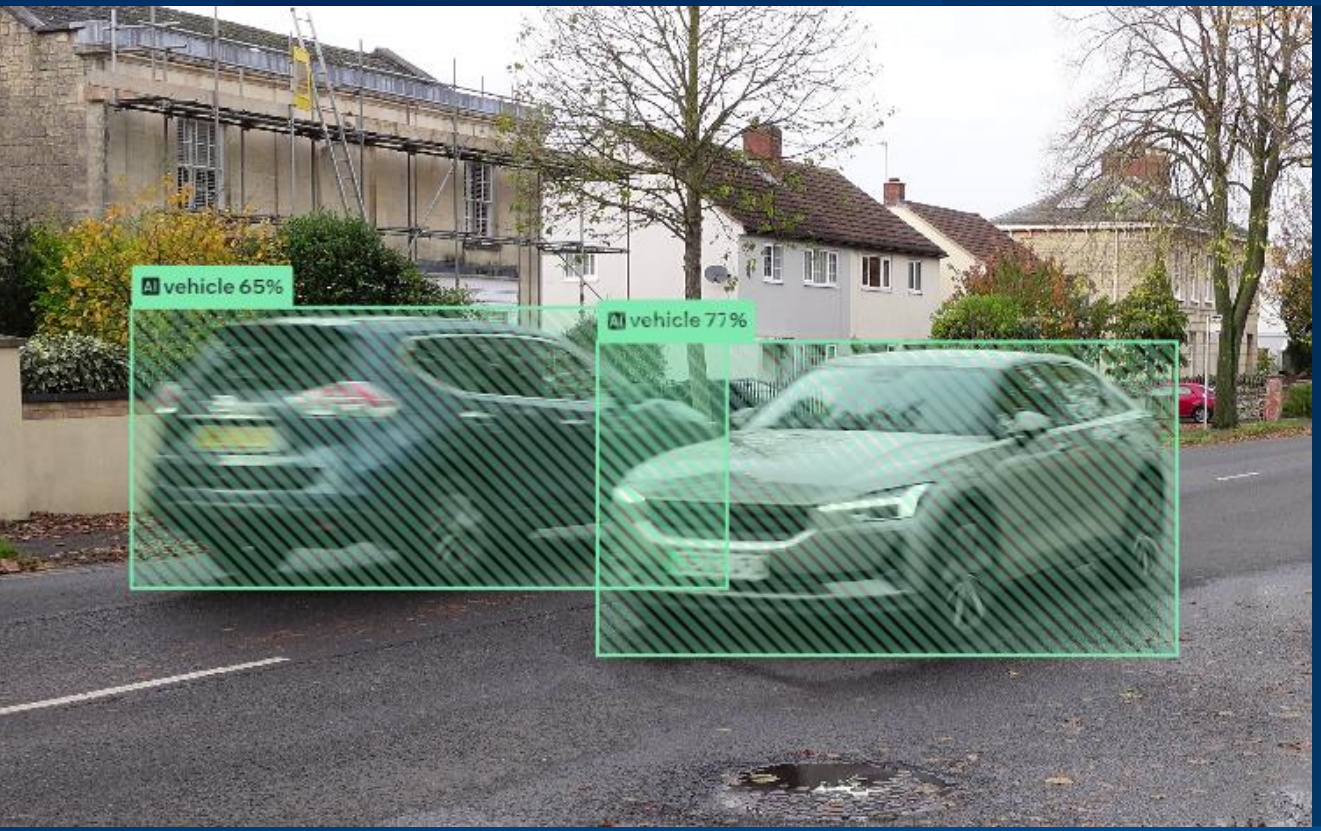
Healthcare

Retail

Scalable asset / inventory tracking

Computer Vision scales into all sectors adding value to businesses and consumers alike

Computer vision at the edge



What is it?

Why does it matter

Where is it used?


Balance
89.9°
Distance
14.6M
30m Sprint
4.18S
Reactions
397MS
Fatigue
88%
Agility
760MS

ai.io

Data Solutions

aiScout

Talent Analysis & Development
App Platform

aiLabs

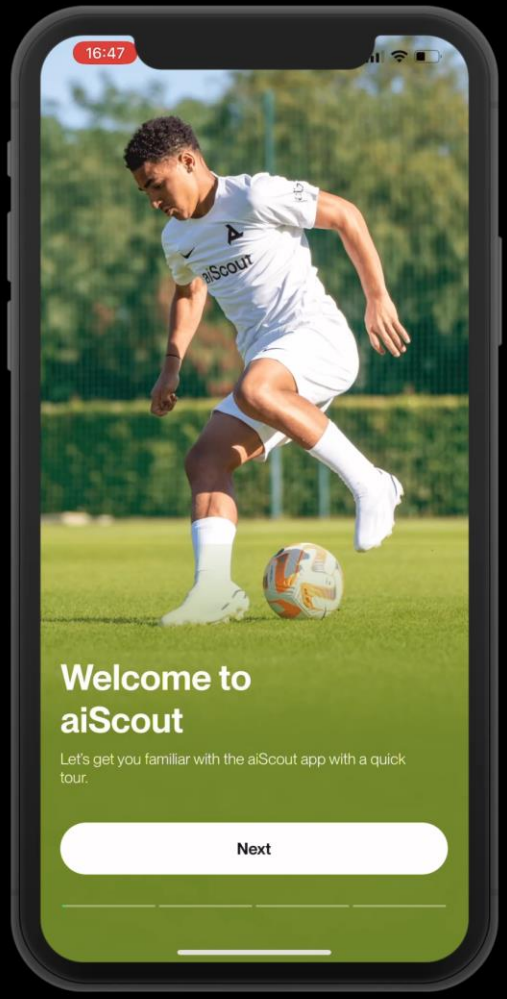
Elite Performance Labs

3DAT

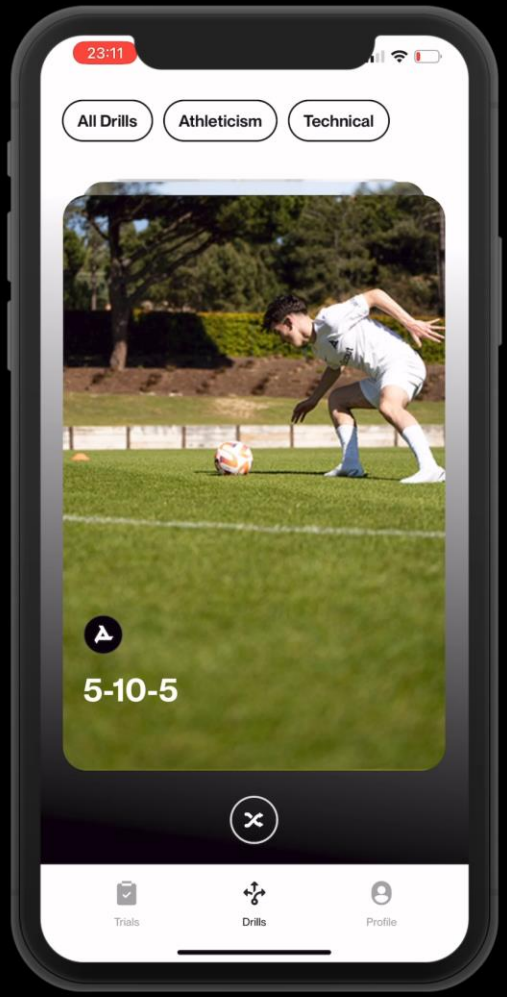
Computer Vision &
Biomechanics Analysis

aiScout

aiScout in action



aiScout in action




TERRITORY
CANONS
CANONS
Kellmatt
Kellmatt
AiSCOUT
aiScout


1

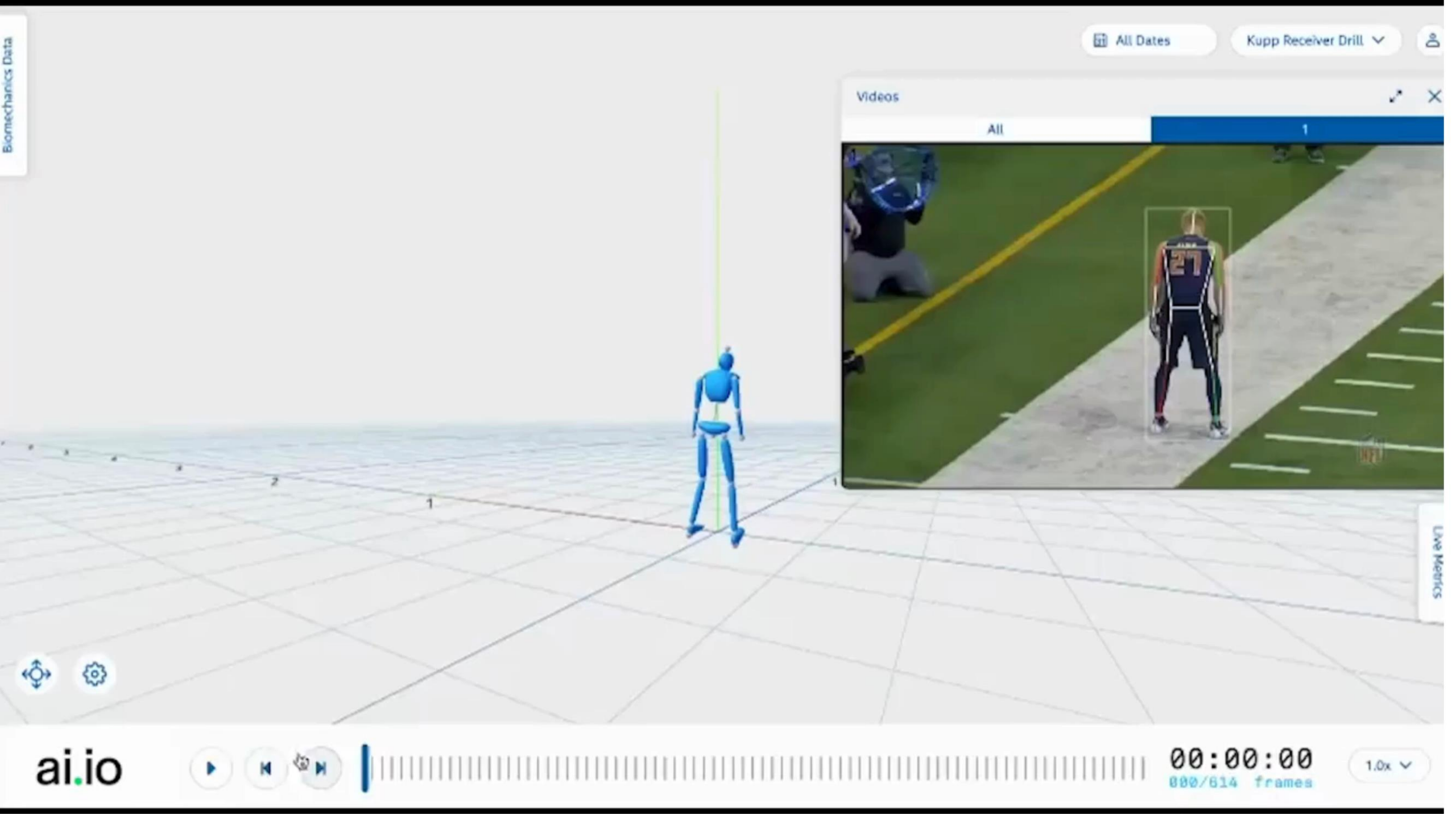

CHELSEA
MLS
ATLANTA UNITED FC
AUSTIN
CHARLOTTE FOOTBALL CLUB
FC CINCINNATI
COLORADO RAPIDS
D.C. UNITED
FC DALLAS
LA GALAXY
HOUSTON DYNAMO FC
LOS ANGELES FC
INTER MIAMI CF
MINNESOTA UNITED FC
CF MONTREAL
NASHVILLE SC
NEW ENGLAND REVOLUTION
RED BULL
Reliance Foundation
YOUNG CHAMPS
NEW YORK CITY FC
ORLANDO CITY
PHILADELPHIA UNION
PORTLAND TIMBERS
REAL SALT LAKE
QUAKES
SEATTLE SOUNDERS FC
ST. LOUIS CITY SC
TORONTO FC
VANCOUVER WHITECAPS FC

3DAT

3D Athlete Tracking

ai.io


OKYO 2020
OKYO 2020
OMEGA
10.8
Men's 200m


Biomechanics Data
All Dates
Kupp Receiver Drill
Videos
All
1
27
Live Metrics
ai.io
00:00:00
000/614 frames
1.0x


aiLabs


aiLabs
Elite Performance Lab

Biomechanics Data

Time Series | Sequence | Metrics

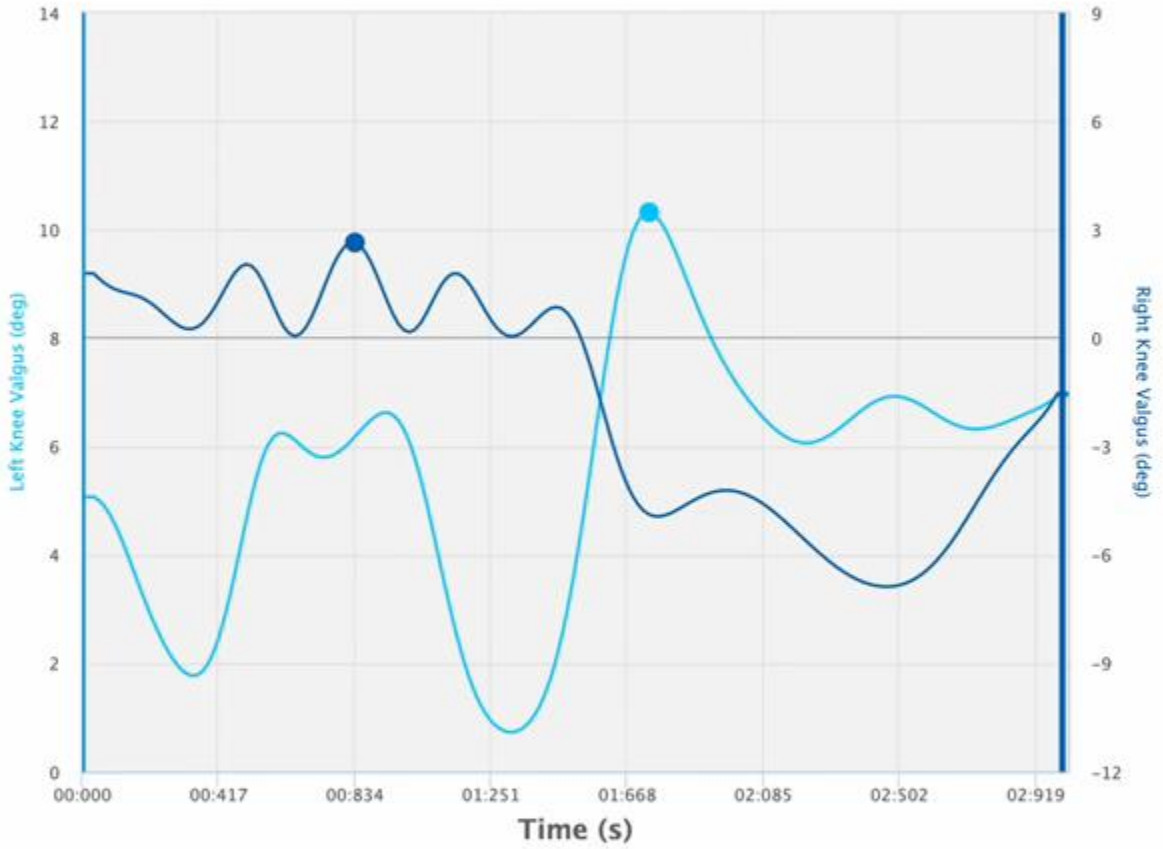
All Dates | CMJ - athlete 2

Left Knee Valgus

5.07

Right Knee Valgus

1.79



Videos

All | 1

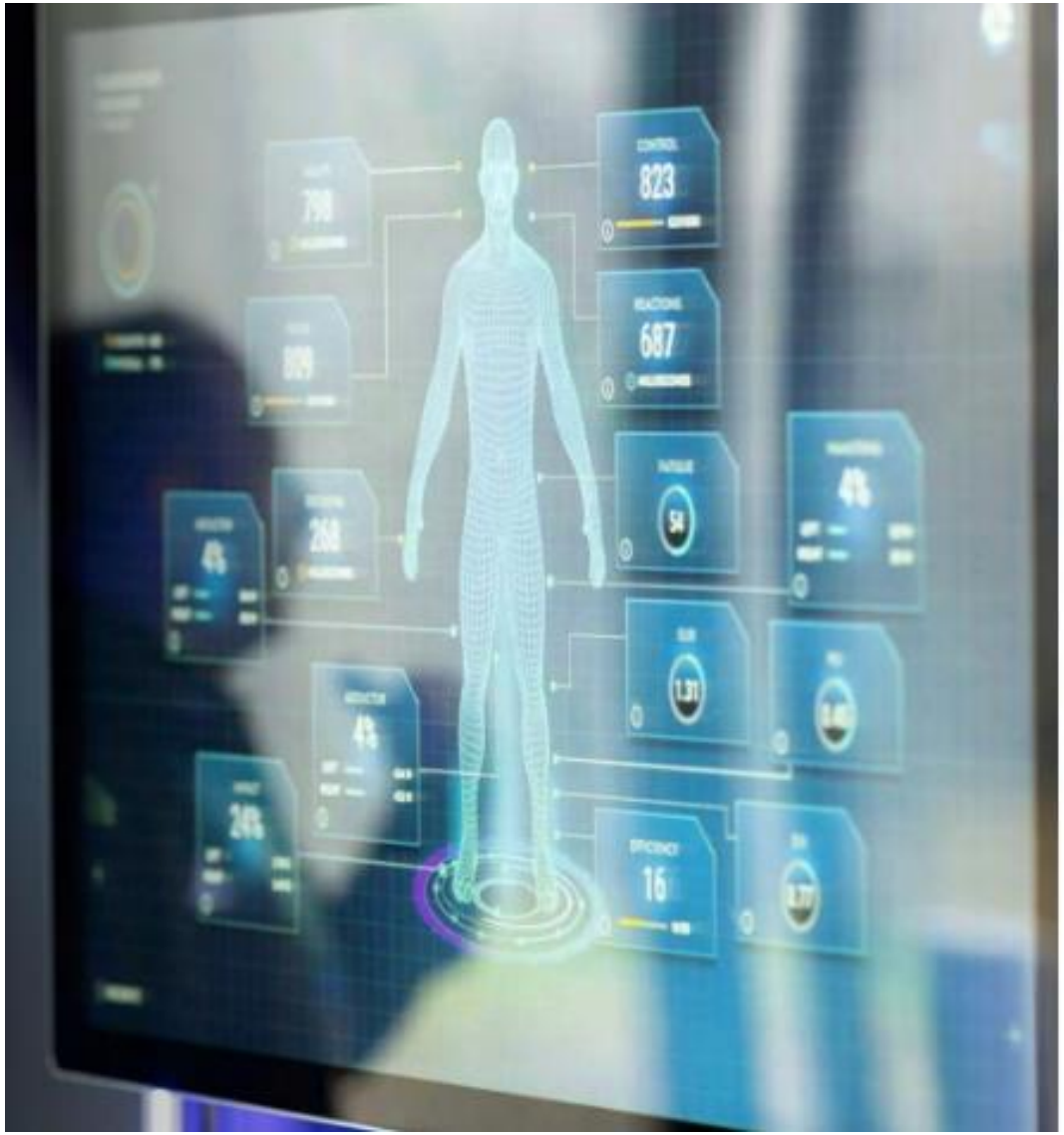
Live Metrics

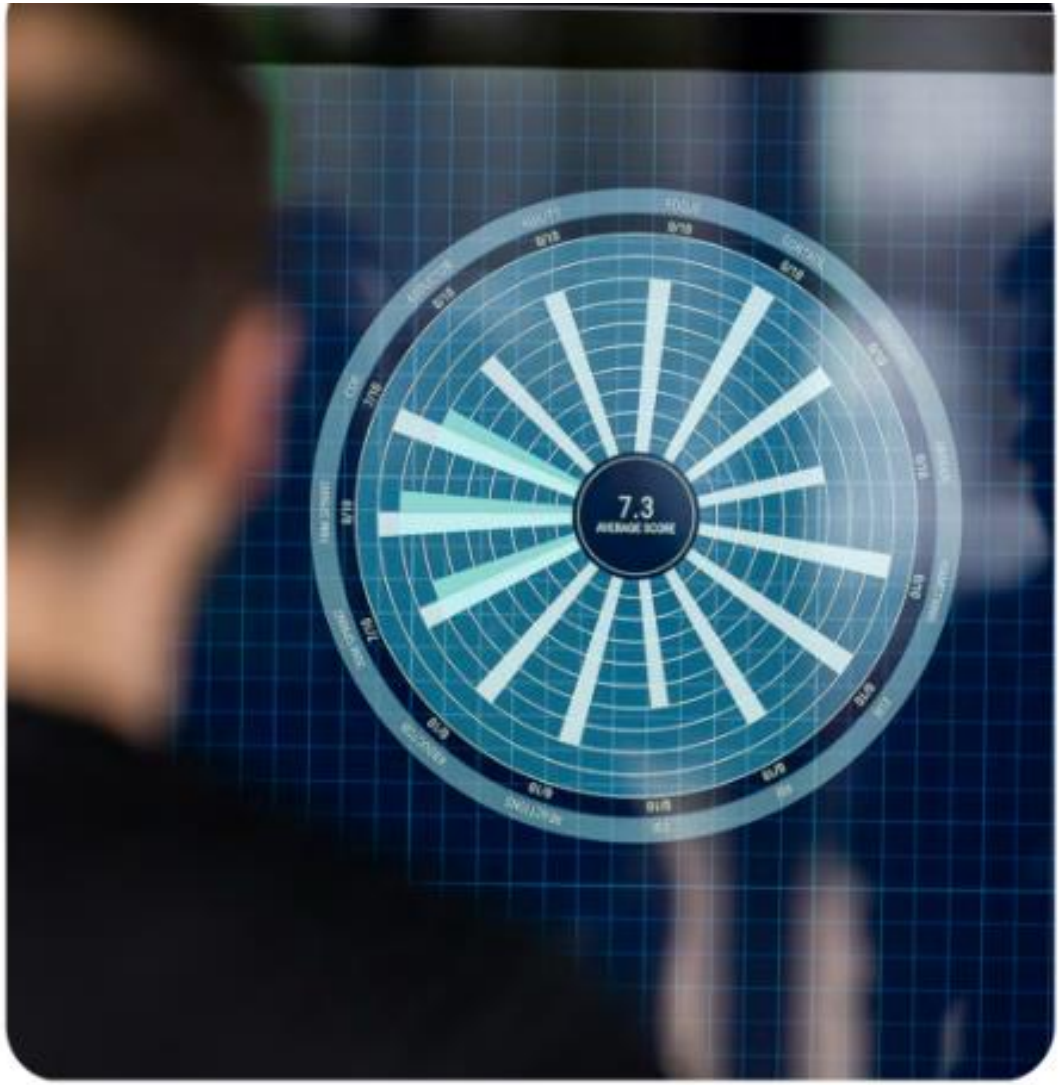
00:00:00 / 00:03:00

frame 000/180

59.940059940059940 fps

1.0x


CONTROL
823
687
268
16


7.3
AVERAGE SCORE


ai.io

Thank You

ai.io

GPU Accelerated Computer Vision

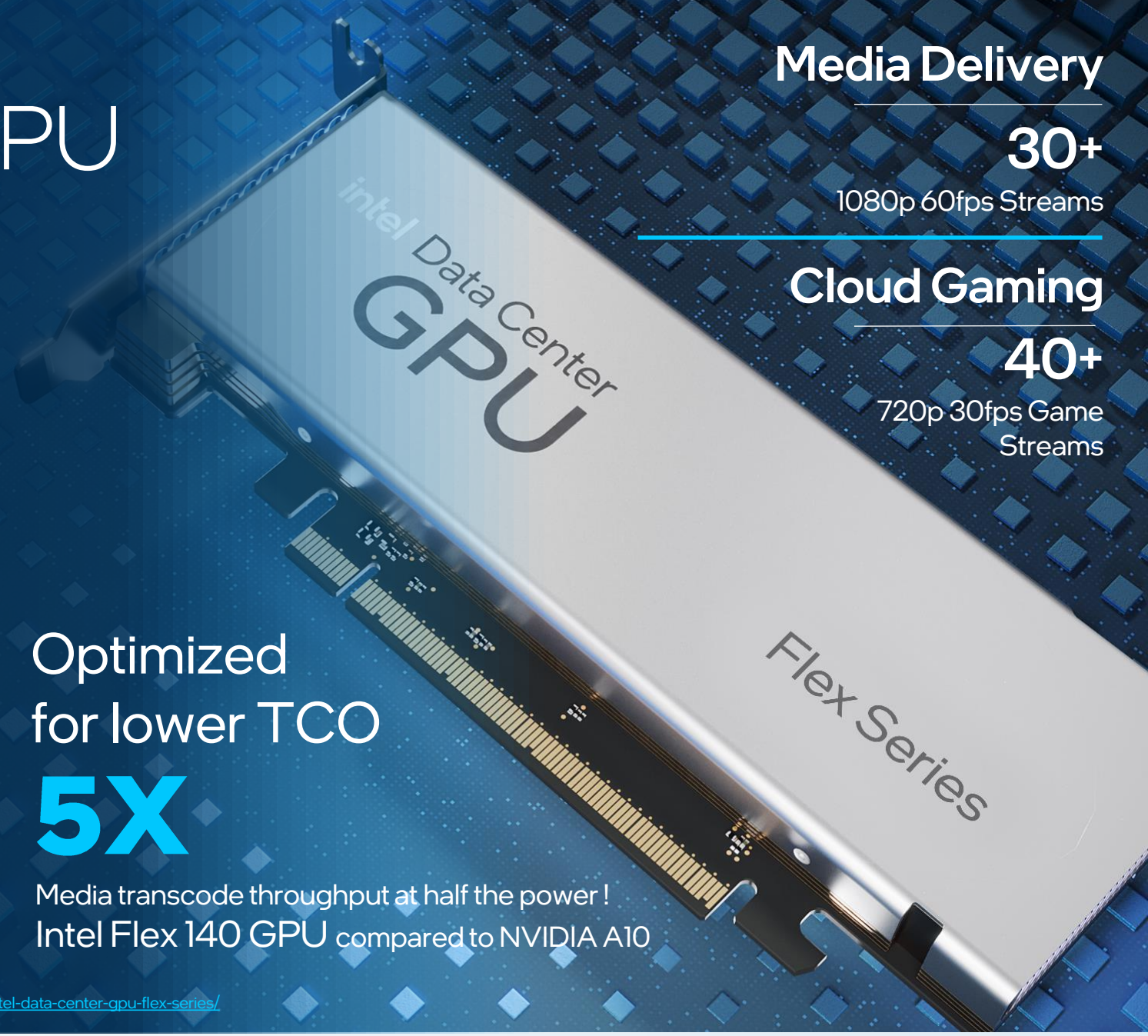
Learn more on the Intel Booth

Visit us at our booth: Intel Booth #750

it starts with

intel

Intel Data Center GPU



Media Delivery

30+

1080p 60fps Streams

Cloud Gaming

40+

720p 30fps Game Streams

FLEX SERIES **140**

4 Media Engines	**75w** Power Envelope	**16** Ray Tracing Units
Xe HPG Architecture	**Half Height** PCIe	**16** Xe cores

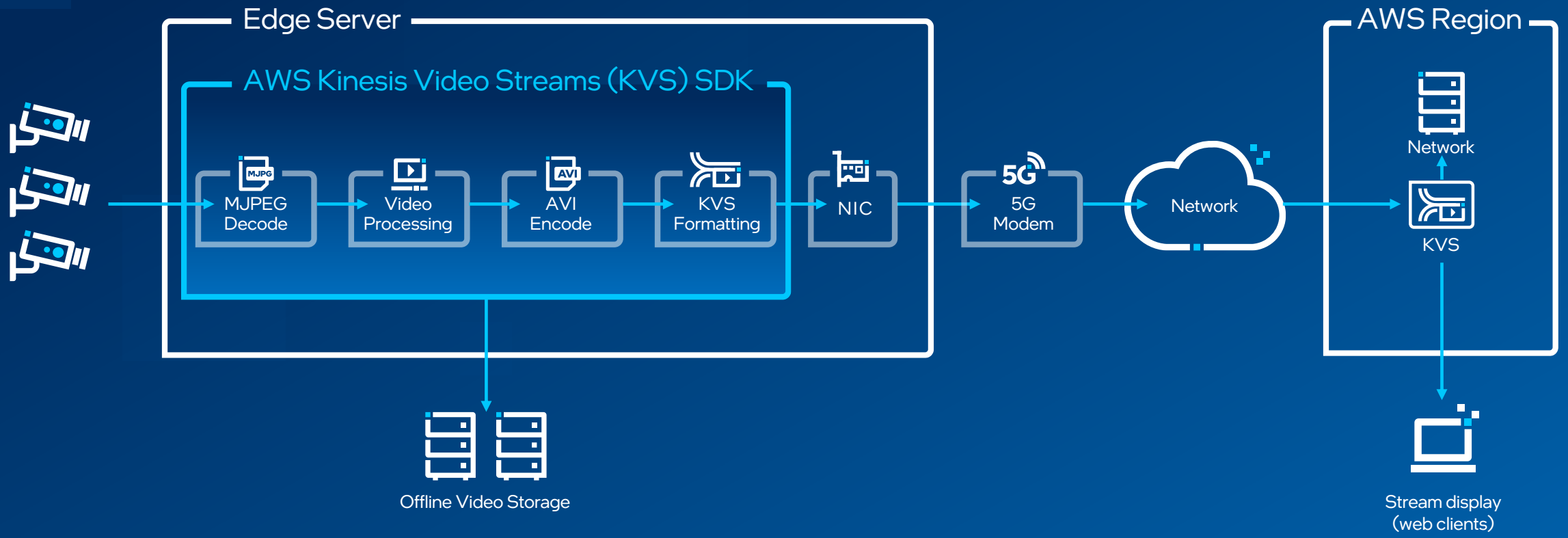
Optimized for lower TCO

5X

Media transcode throughput at half the power !
Intel Flex 140 GPU compared to NVIDIA A10

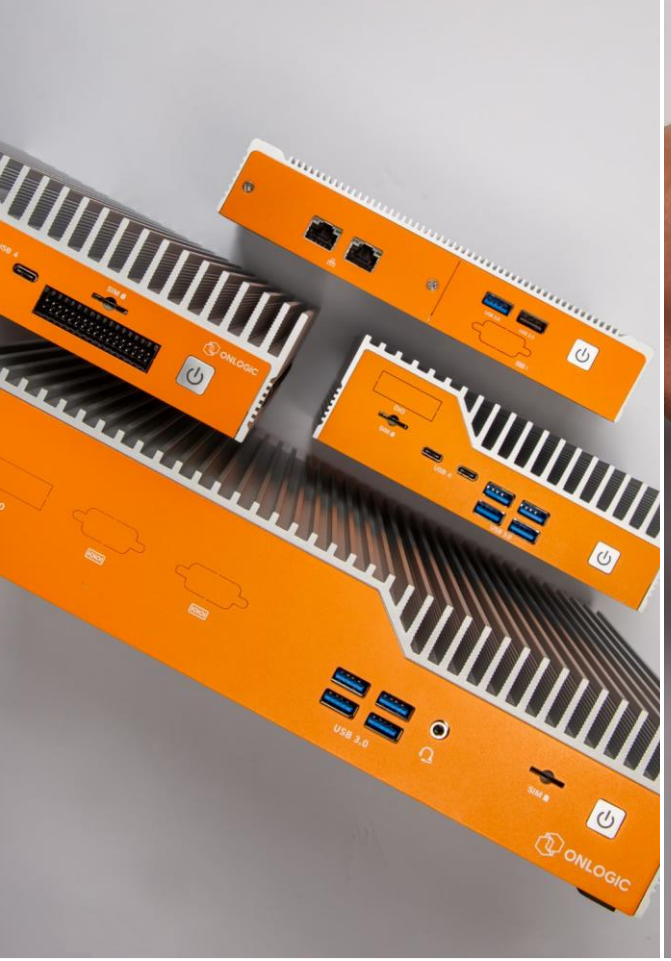
See performance claims https://edc.intel.com/content/www/us/en/products/performance/benchmarks/intel-data-center-gpu-flex-series/

Solution Architecture



ONLOGIC

Let’s Make It Possible



Industrial Computers

- Small Form Factor
- Fanless or Active Cooling
- 0 to 50°C Operating Temp.

Rugged Computers

- Resistant to Shock & Vibration
- Wide Power Input Range
- -40 to 70°C Operating Temp.

Panel PCs

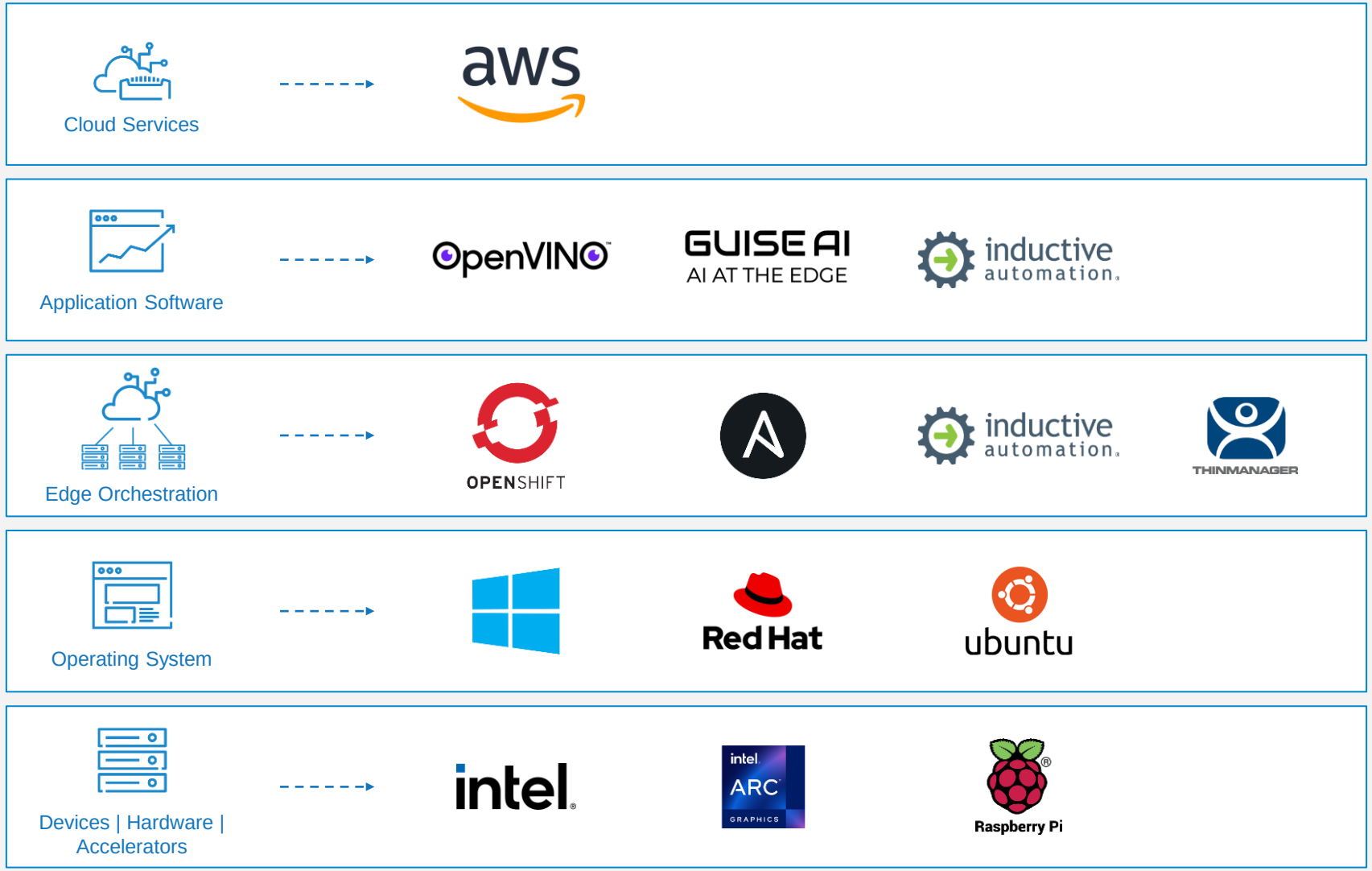
- 8.4” to 24” Screen Sizes
- Resistive or Capacitive Screens
- Available with IP65 Front Bezels

Edge Servers

- 1U to 4U Sizes
- Advanced Intel Processing
- Highly Configurable

ONLOGIC

Collaborations that make **AI at the Edge** easy



*Only a subset of our partners listed

Manufacturing line segment

Raw Data

Inference Data

Deploy, Manage & Monitor

Edge Server

AI Catalog

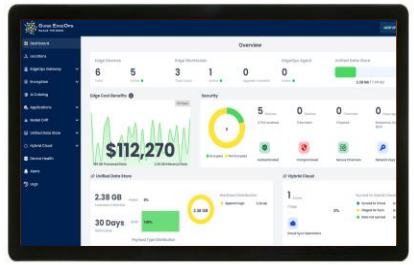
Data Lake

GUISE AI Models

Inference data

GUISE EDGEOPS SCALE THE EDGE

Dashboard



Raw Data

Running on the edge device:

GUISE AI Models + Red Hat Device Edge

ONLOGIC

OnLogic Demos at re:Invent

Intel Booth - KVS Demo:

Expo Hall - Venetian Conference Center, Booth 750

- Running Amazon Kinesis to process HD video streams and send data to the cloud.

Builders' Fair – Find My Ship:

Expo Hall - Venetian Conference Center

- Controlling AWS IoT Shadow updates along with view of Ultra-Wide Band Data visualization in 3D.

AWS IoT Kiosk – AWS IoT DeepRacer Vx Demo:

Expo Hall - Venetian Conference Center

- Showcasing Greengrass Runtime Software at the edge to get car data (via CAN Bus) and visualize it.

AWS Disaster Response DDIL activation:

Venetian Hotel Lobby - Disaster Response Jeep area

- Aggregating the GPS locations and metadata from 60 LoRaWAN GPS trackers deployed to re:Invent staff.

AWS GenAI Chess activation:

Rec Center – Mandalay Bay Convention Center

- Used for AWS IoT Shadow and Jobs update for each Chess move of the robots and to perform a Trust and Verify check between the physical chess board and the output of the GenAI model.

ONLOGIC

Expanding computer vision to the far edge

Cloud Dater Center → Core Network → Network Edge → On-Premise Edge → Far Edge & Things

AWS Region

NFV workloads

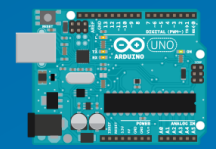
Telco Edge & Colo Edge

RAN Hub & Edge Compute

Edge Compute & Edge Inference

Sensors & Transmitters

Power, form-factor, connectivity and remote management become increasingly challenging



ARDUINO PRO

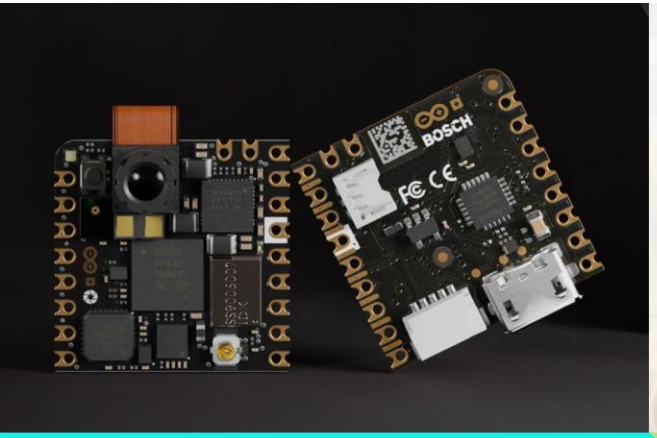
ARDUINO CLOUD

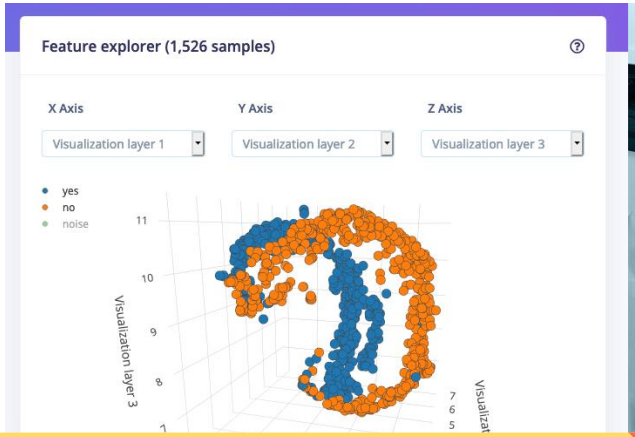
EDGE IMPULSE

Arduino Cloud for Business Machine Learning Tools

AWS re:Invent 2023

Machine Learning Tools: Endless possibilities





Collect

Acquire valuable data securely from your devices and rapidly build custom dataset

Design

Develop algorithms with ready-to-use digital signal processors and machine learning blocks

Test

Validate and train Machine Learning models with real-time data

Deploy

Build optimized embedded inference

ARDUINO PRO

Machine Learning Tools at a Glance

ARDUINO CLOUD

EDGE IMPULSE

Data collection - Build custom datasets by collecting sensor, audio or camera data straight from devices, files or cloud integrations.

ML model design - Start creating your model.

ML model test - Validate your model and picture how it will perform with real-world data.

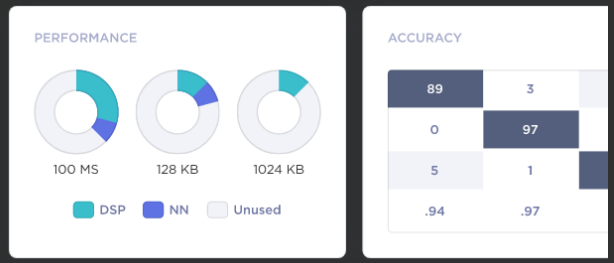
ML model deploy - Deploy your model to any device.

Connected Predictive Maintenance

ARDUINO CLOUD

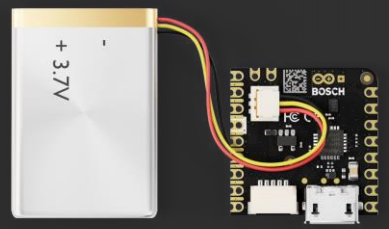
EDGE IMPULSE

ML Model
Sound / object / motion detection



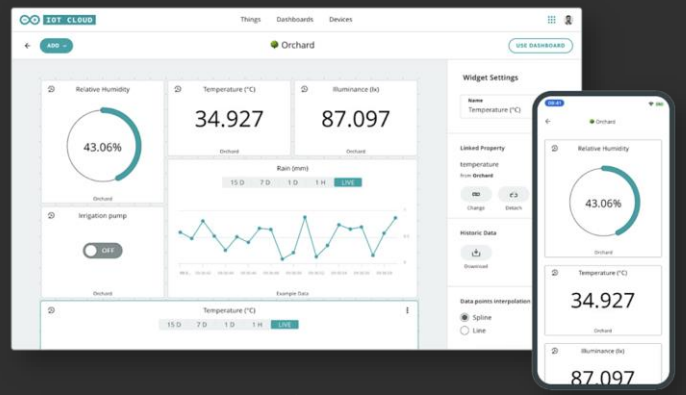
Industrial machine
Cartoning machine

Smart sensing
Nicla board



VIBRATION NOISE VISION

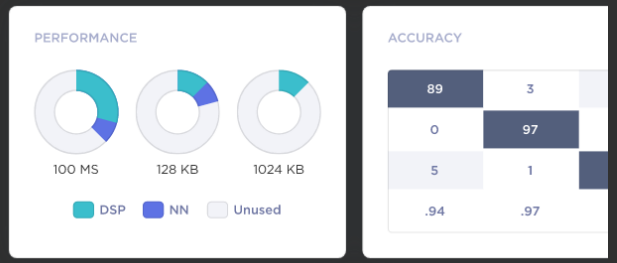
Arduino Cloud
Dashboards + Mobile App



CHARTS ALARMS NOTIFICATIONS

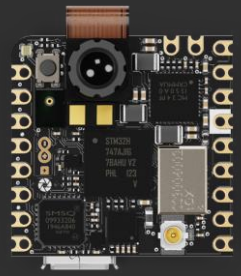
Component defect detection

ML Model
Object detection



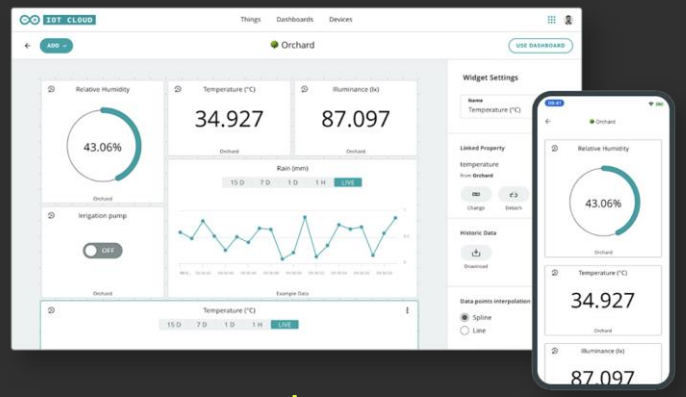
Production line
Juice production

Defect detection
Nicla Vision



VISION

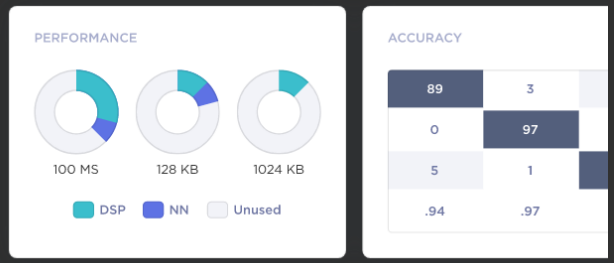
Arduino Cloud
Dashboards + Mobile App



CHARTS ALARMS NOTIFICATIONS

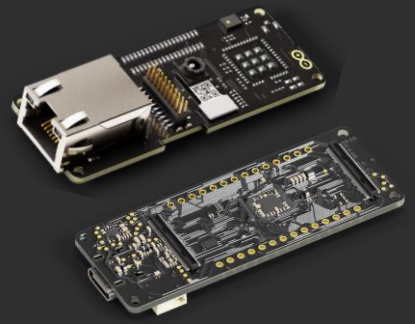
Component defect detection

ML Model
Object detection



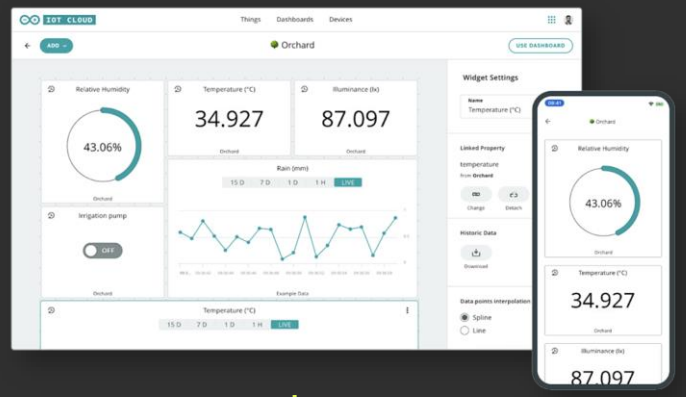
Logistics
Automated inventory management

Amount / defect detection
Portenta H7 + Vision Shield



WiFi

Arduino Cloud
Dashboards + Mobile App

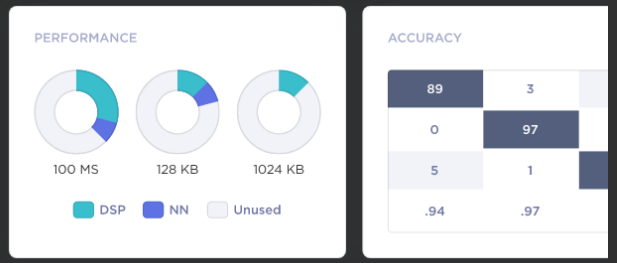


VISION

CHARTS ALARMS NOTIFICATIONS

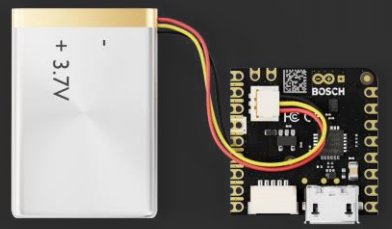
Fall detection

ML Model
Motion detection (IMU)



Person falling

Wearable fall detection
Nicla board



BLE

Smartphone
Help call

MOTION

AI on Xeon

it
starts
with

4th Gen Intel Xeon Processors on AWS

Intel Advanced Matrix Extensions (AMX)
Accelerates Deep Learning Training and Inference

Intel Data Streaming Accelerator (DSA)
Accelerates Data Streaming

Intel In Memory Analytics Accelerator (IAA)
Accelerates in-memory compression and encryption

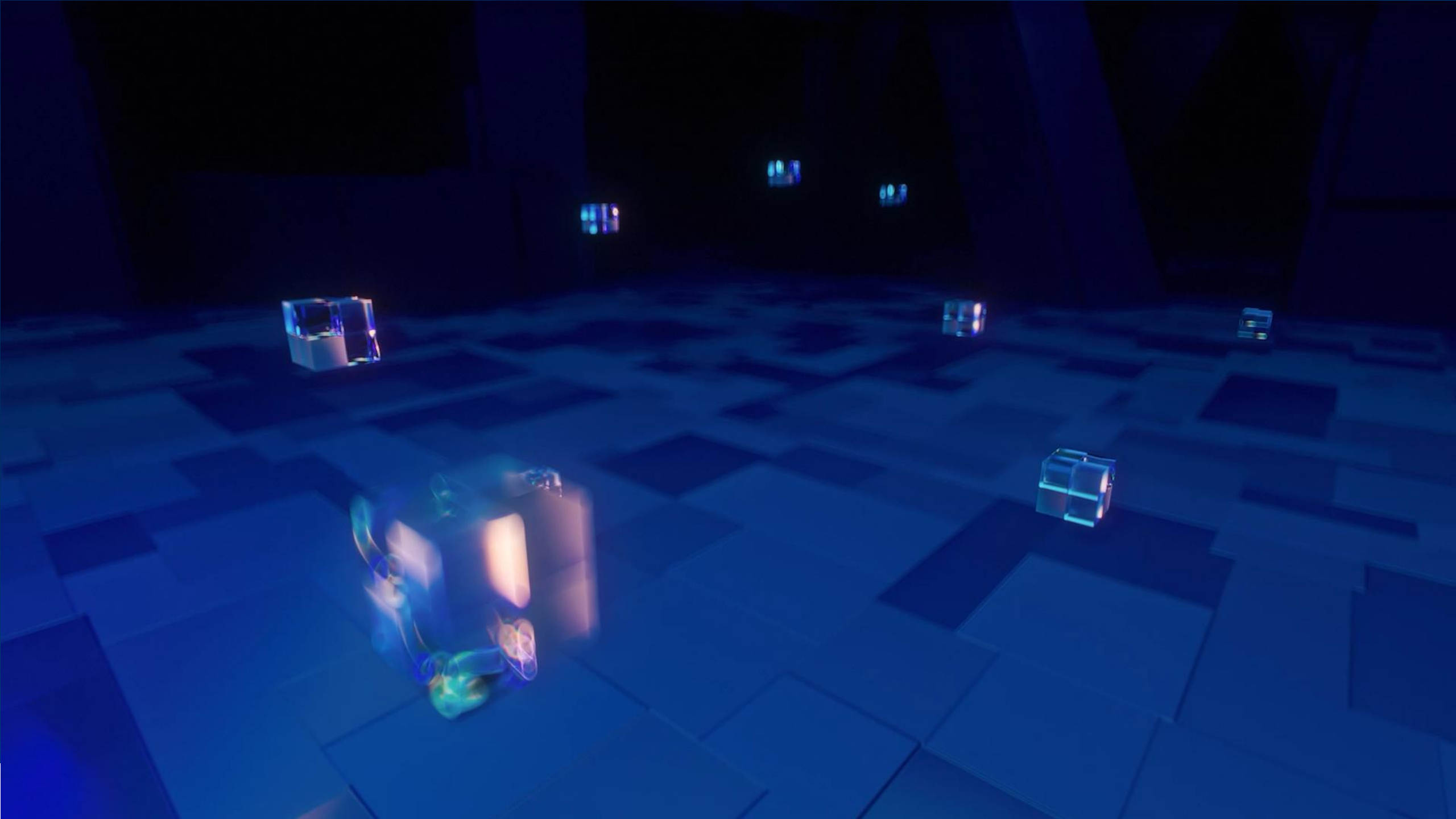
Intel Quick Assist Technology (QAT)
Accelerates data compression and encryption at scale

Amazon EC2 C7i, M7i, R7i and R7iz are available in the US and Europe and growing

M7i-flex offers up to **19% better price performance** versus M6i

Up to **80% Higher AI Inference throughput** over previous generation

1 See backup for config details.
2See intel.com/processor claims: 4th Gen Intel Xeon Scalable processors. Claim E1. Results may vary.
3On select workloads. See intel.com/processor claims: 4th Gen Intel Xeon Scalable processors Claim E6.Results may vary.
4 Intel 2021-22 CSR Report and internal reporting



Intel Optimized Ecosystem for AI

ECOSYSTEM & PARTNERS

DATA | MODEL | DEPLOY

MODIN

TensorFlow PyTorch

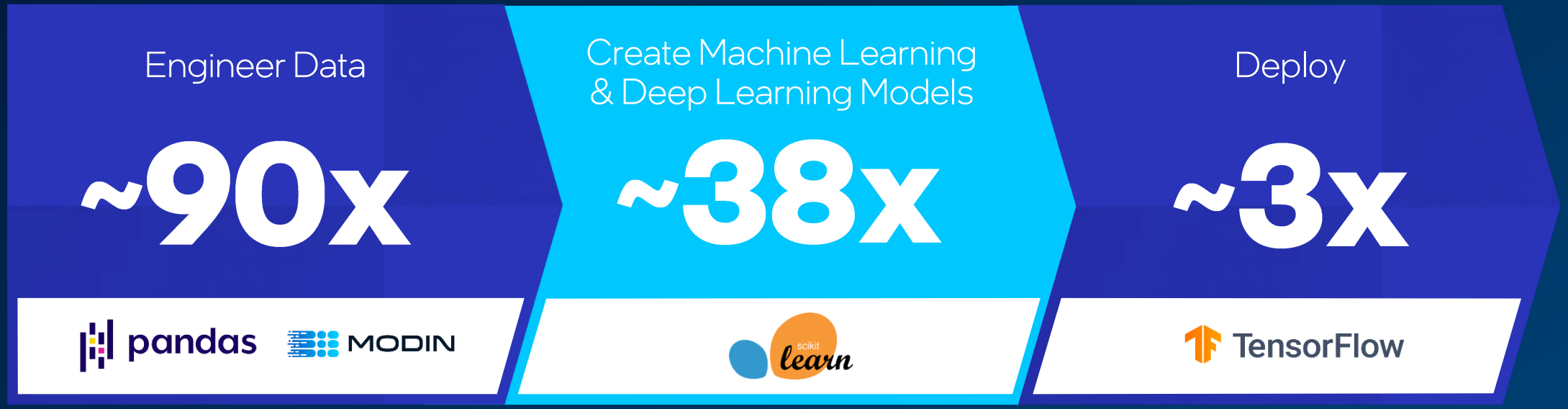
OpenVINO

oneAPI

Open, Standards-Based Programming Model & AI Libraries

intel XEON

One-line Code Change = Huge Performance Gains




```
import modin.pandas as pd
```

```
from sklearnex import patch_sklearn
patch_sklearn()
```

```
TF_ENABLE_ONEDNN_OPTS=1
```

For workloads and configurations see https://www.intel.com/content/www/us/en/developer/tools/oneapi/tech-articles-how-to/overview.html. Results may vary

Xeon AI and "Large" Language Models (LLMs)

Expert LM fine-tuned on a single task can outperform a multi-task model trained on 300 tasks

Race for one model to rule them all

"...we're at the end of the era where it's going to be these, like, giant, giant models," - Sam Altman, OpenAI CEO

"More companies would be better served focusing on smaller, specific models that are cheaper to train and run." - Clement Delangue, HuggingFace CEO

"small optimized models in healthcare are as accurate as large ones while being much more efficient." - David Talby, Spark-NLP CEO

Models to fit every business need

Intel Xeon AI for LLMs

- Enable the most cost effective and ubiquitous approach
- Fine-tune and optimize inference models on Intel Xeon processors
- Leverage hundreds of Intel and 3rd party pretrained models

Summary

- Use Intel solutions to run edge AI efficiently and at scale
- Leverage Intel 4th Generation Xeon based instances with AMX capabilities for AI
- Enable Intel SW optimizations for data processing, training and inference
- Visit us at our booth: **Intel Booth #750**

Connect With Us

Explore and interact with our demo at the **Intel booth #750** for optimization solutions.

Connect With Our Partners

Visit our partners' booths to see how they use Intel technology.

CDW	—	Booth #305
HPE	—	Booth #630
IBM	—	Booth #930
SingleStore	—	Booth #1586
World Wide Technology	—	Booth #131
Nutanix	—	Booth #132
SUSE	—	Booth #501

Notices & Disclaimers

Intel is committed to the continued development of more sustainable products, processes, and supply chain as we strive to prioritize greenhouse gas reduction and improve our global environmental impact. Where applicable, environmental attributes of a product family or specific SKU will be stated with specificity. Refer to the 2022 Corporate Responsibility Report (p. 64) for further information.

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Code names are used by Intel to identify products, technologies, or services that are in development and not publicly available. These are not "commercial" names and not intended to function as trademarks.


it
starts
with
intel

Thank you!

Please complete the session survey in the mobile app

Jonathan Lee
jonathan.lee@ai.io

Michael Kleiner
michael.kleiner@onlogic.com

Allan Gagnon
a.gagnon@arduino.cc

Alex White
alexander.white@intel.com

Mohan Potheri
potheri.mohan@intel.com