EDGE AI FOR CONNECTED INTELLIGENCE

AMIT MATE , FOUNDER & CEO , GMAC INTELLIGENCE

GMAC MISSION AND VISION

- We are building Connected Intelligent solutions for Physical-Safety / Security

  https://gmacintelligence.com/

- Our mission is to enable "Connected Intelligent" applications on consumer electronic devices (Edge or IoT)

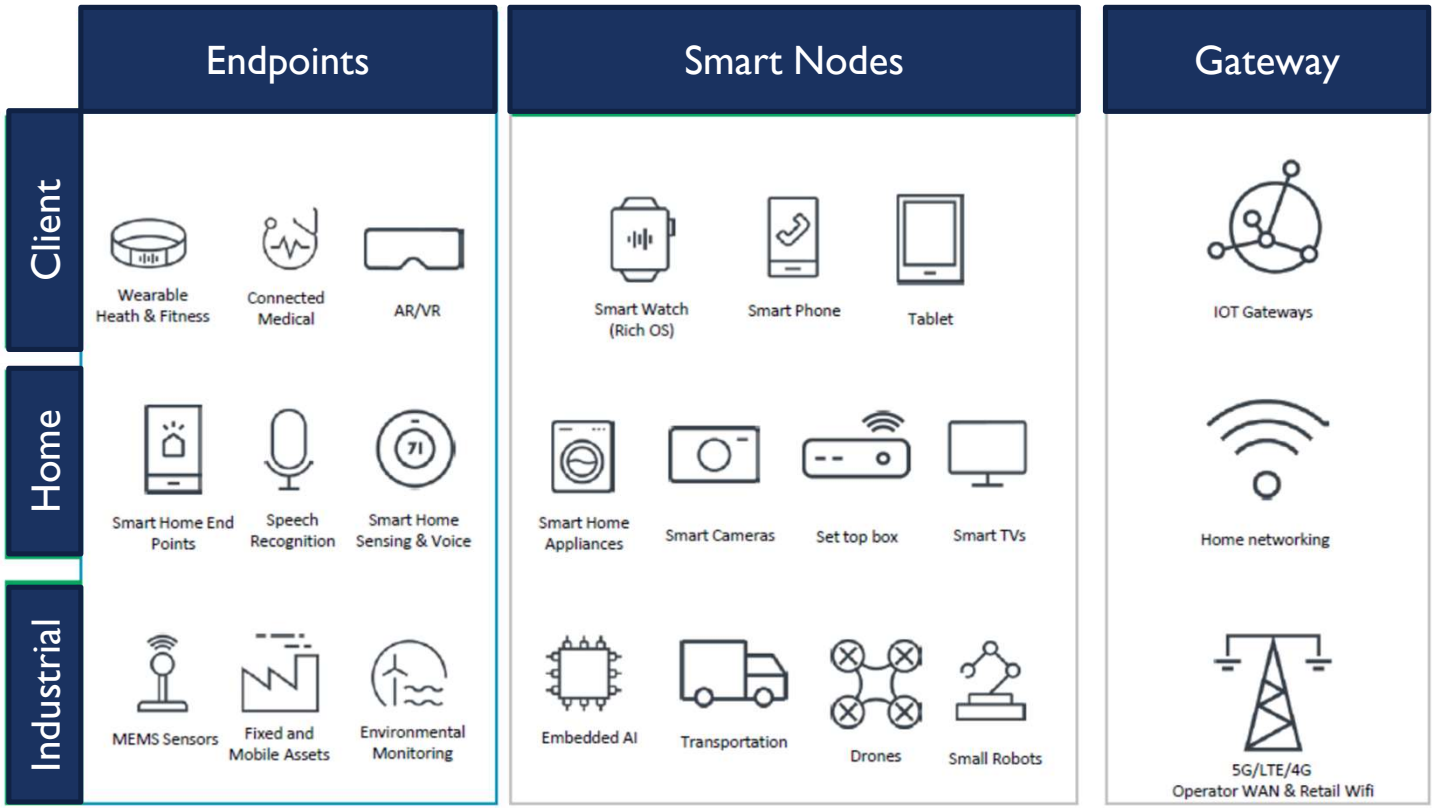
AI/ML software => real-time, on-device implementations of DNN models with state of the art accuracy and power efficiency for Edge devices.

- Qualcomm Smart City Accelerator and Qualcomm Advantage Network member
- NVIDIA Inception cohort & presenter at GTC 2020/2021
- IISc Deep-tech cohort
- 4th globally in Google visual-wake-word challenge-2019
- Recipient of Google TFRC Grant ($100K compute grant)

GMAC EXECUTIVE TEAM

- Founder and CEO : Amit Mate , ME ECE IISc
- Co-Founder : Nagaraj B, ME ECE IISc/IIT-M

EDGE AI SCOPE



By 2025

TAM > 25B devices,
SAM > 6B Edge AI devices

By 2025

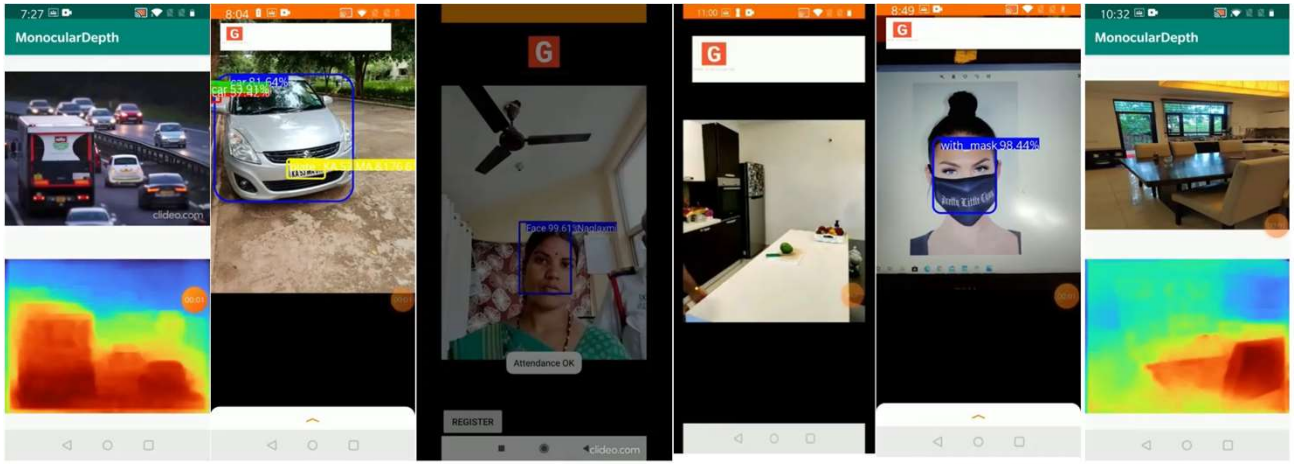
TAM > 25B devices,
SAM > 6B Edge AI devices

Business Model

AIoT-as-a-service
B2B, B2C opportunity

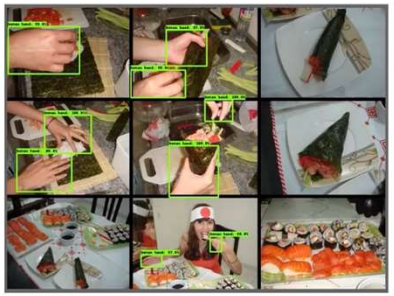
Source: ARM

EDGE AI TECHNOLOGY DEMO VIDEO



https://youtu.be/EctG0wH85Ag?list=PLyrKlKYNwP82NPaA_LkeggZovohEr6FKo

GI4ALL

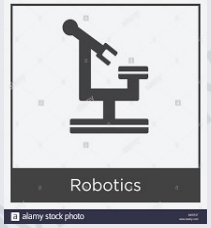


AI/ML SOLUTIONS

License Plate Recognition

Activity recognition

Facial Recognition



Robotics

B100

Automatic License Plate Recognition

D100

Surveillance

FR100

Facial Recognition Attendance

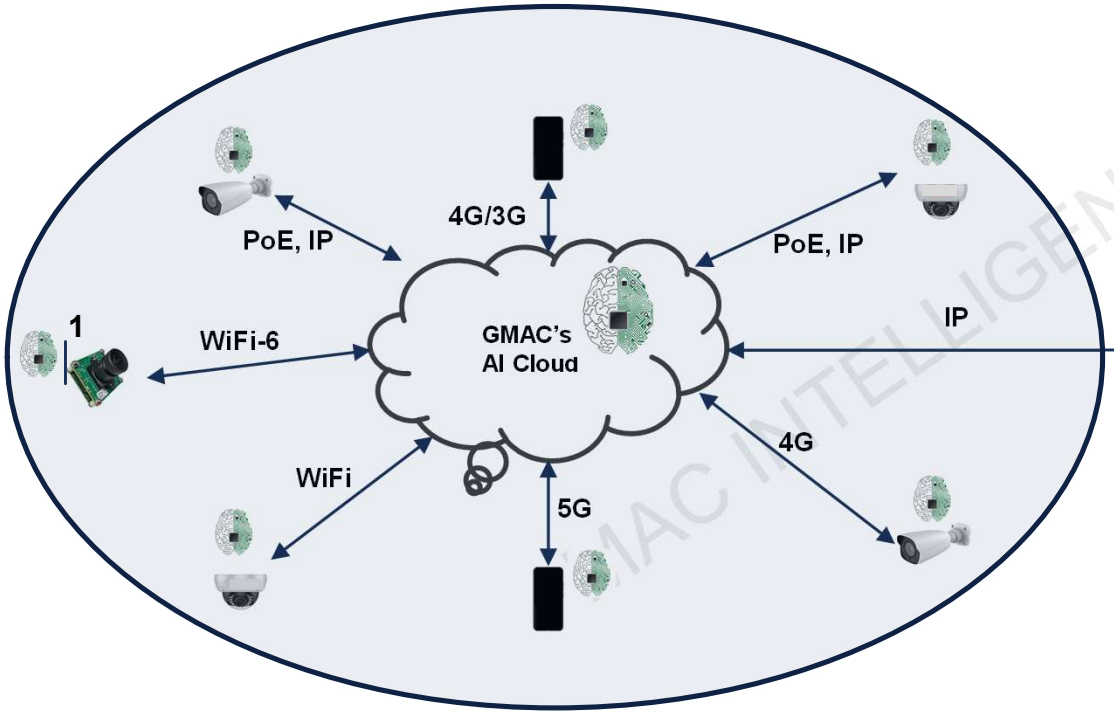
A100

Automotive

Off-the shelf Cameras + GMAC AI/ML Apps + GMAC Intelligent Cloud => Complete AI/ML Solutions

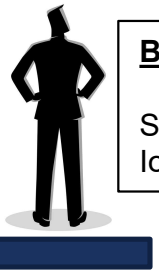
https://play.google.com/store/apps/dev?id=7183583045448282341&hl=en-GB

CONNECTED INTELLIGENCE ARCHITECTURE



B2B Enablers

1 => GMAC's on-device API and library
2 => GMAC's REST API



B2B Options

SaaS aggregators=> enable custom dashboards
IoT device makers => enables AIoT devices

Solution architecture can be leveraged for robots, drones, autonomous vehicles

EDGE AI OPPORTUNITIES AND CHALLENGES

OPPORTUNITIES

- On-device or Edge AI enables new applications and new revenue streams
  - For IoT/Edge device vendors and system integrators, SaaS aggregators (B2B opportunity)
- Cloud-only based AI is often slow, expensive and raises security/privacy concerns

CHALLENGES

- **Constrained environment** - IoT/Edge devices have limited memory (10KB++) , storage (1MB++) and compute (50MHz++)
- **Fragmented technology landscape** – Tensorflow or Pytorch or TensorflowRT – uC or DSP or CPU or GPU or NPU – Android or Yocto linux or Ubuntu

GMAC'S ON-DEVICE LIBRARY AND SOLUTIONS

- Target B2B customer profiles
  - Neuromorphic/AI chip vendors
  - IoT device makers
  - R&D Institutes (Defense)

- GMAC's software solution
  - On-device face-detection/identification or custom library
  - AIBox software solution – multi-stream AI applications (ANPR, Facial Recognition, Activity recognition etc)
  - Custom Edge Solutions –Real-time speech recognition+ translation solution on battery operated Edge devices

GMAC'S E2E SOLUTION FOR SMART CITIES AND SMART CONNECTED SPACES – REST API INTEGRATION FOR 3RD PARTY

- Target B2B partner profiles
  - Facilities Management
  - Airport Management
  - Parking Management
  - Community Management

- GMAC's easy two step Integration
  - API#1 => Sign-in to user account
  - API#2 => Get secure real-time notifications in your front-end or backend

GMAC'S FULL INTEGRATION API – SMART CITY APPLICATIONS

- API: Signin/Signup (POST)
  https://identitytoolkit.googleapis.com/v1/accounts:signInWithPassword?key=XXXXXXXXXXXXXXX
- API: Create New Customer (POST)
  https://us-central1-gi4all.cloudfunctions.net/api/createNewCustomer
- API: Add Employees (POST)
  https://us-central1-gi4all.cloudfunctions.net/api/addEmployees
- API: Add Locations (POST)
  https://us-central1-gi4all.cloudfunctions.net/api/AddNewLocations

- API: Photo Search (POST)
  https://us-central1-gi4all.cloudfunctions.net/api/refimage2
- API: Attendance Data (GET)
  https://us-central1-gi4all.cloudfunctions.net/api/
  + authorized reads from front-end
- API: CRUD Interface (POST)
  https://us-central1-gi4all.cloudfunctions.net/api/create
  https://us-central1-gi4all.cloudfunctions.net/api/read
  https://us-central1-gi4all.cloudfunctions.net/api/update
  https://us-central1-gi4all.cloudfunctions.net/api/delete

CUSTOMER FEEDBACK / CASE STUDIES

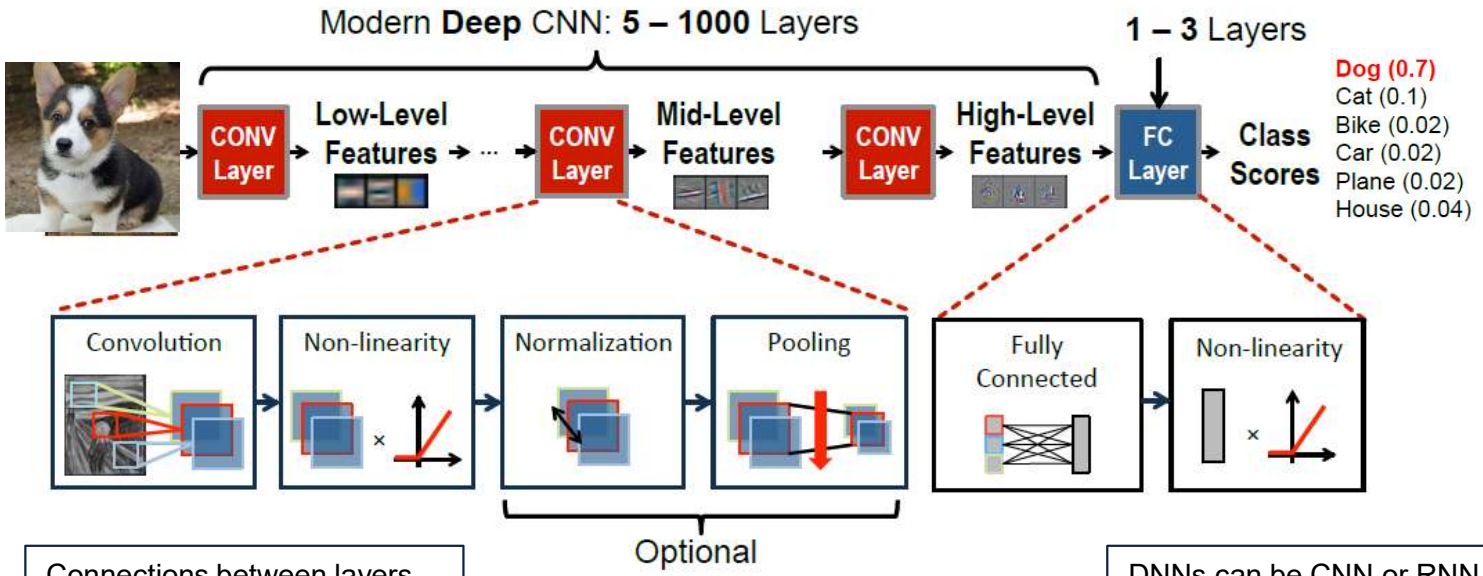
- **CEO Manufacturing Unit , Malur Industrial Area**: "I have 200+ employees, manually tracking their attendance across three shifts was a nightmare. It is humanly impossible for a single person to know all these employees. I didn't know who was coming in and going out. Thanks to GI4ALL-FR100, I now have instantaneous access to attendance data, nicely filtered per department and generates attendance sheets that feed directly to my payroll system.
- **Tier-1 System Integrator, Hyderabad**: "I was paying in lakhs for facial recognition software which was not even accurate, GMAC's product is not only real-time but also accurately recognizes people from a month's worth of data.
- **Head of Enterprise IT company, Bangalore** : "We have our offices at two locations in India, during the pandemic, it was difficult to track people and vehicles coming in and out of our offices. "GMAC's solution was a breeze as it is plug-n-play. Our IT folks love it, absolutely zero maintenance"
- **Medical Care Center, Bangalore:** "We have multiple locations I visit these clinics only during certain hours. I had to rely on other medical staff to keep track of timeliness. Now all that information is at my finger-tips and I didn't even have to install a single computer. Who was going to manage it anyways?
- **Builder and Community Manager, Bangalore**: "We have a security staff of 10+ and several surveillance cameras, but still safety was a concern. It is not humanly possible to keep an eye every second on surveillance cameras. Visitors used to be tired of waiting at security gates to comply with security procedure. After installing GMAC solution, it's a breeze. Its like there is one smart person behind every camera alerting us about events as and when they occur and all visitor compliance procedures now take seconds.
- **Owner** , **Ramada Inn, Ohio, USA**: "I have more than 300+ daily-wage workers on my roll working at different times of the year. My site-manager had to manage username/passwords for all these folks, adjust hours manually to ensure compensation according to Federal laws. When I showed him GI4ALL-FR100, and how smoothly and accurately it works, without eating into his productive time, I could see the smile on his face.

KEY CUSTOMER BENEFITS

- Monthly savings in manual security systems
- Instant and automatic check-ins/check-outs (< 1 sec)
- Instant text search of old recordings/footage (< 5 sec retrieval)
- Server-less application – $0 maintenance in cloud/onsite servers
- Instant provisioning from portal (Add 500 people/cars in 2 sec)
- Automatic and instant analytics (work hours, attendance and more)
- Plug-n-play (instant addition of new devices in sync with existing data)

EDGE AI – UNDER THE HOOD

MODERN DNNs

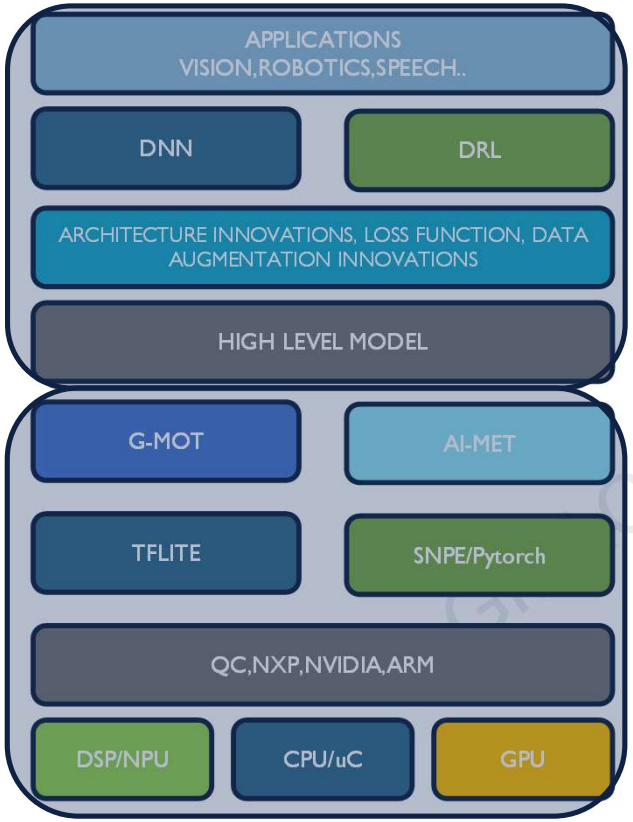


Connections between layers could be dense or sparse OR feed-forward, skip or recurrent

DNNs can be CNN or RNN or combined or just FC

From Accelerators's (e.g., ARA=1) workload perspective – inferencing is the focus – training usually done on high end servers/GPUs and not covered here

EDGE AI TECH STACK – KEY DIFFERENTIATORS



Differentiators

- **On device acceleration API :** Supports heterogenous compute
- **On device training API :** Supports training on the Edge
- **Distributed Intelligence :** Real-time AI sync across Edge devices
- **Hybrid Edge/Cloud AI :** Facial recognition on Edge, Analytics/Search in cloud
- **Plug-n-play :** Provision new devices under a minute
- **Common Dashboard**: Single dashboard for ANPR, FR and Activity recognition

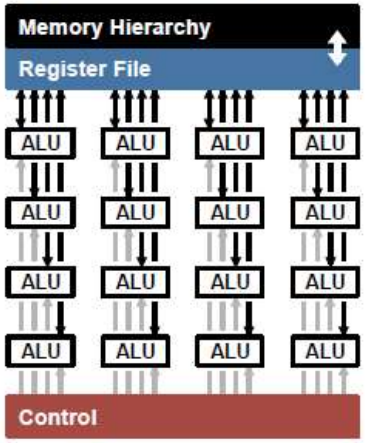
B2B Application Enabler

- Rest API to enable custom dashboards
  (Create/Delete customer/device/employee/visitor, Read/Update attendance)
- On-device API to enable new applications and devices (Register models, Infer input)

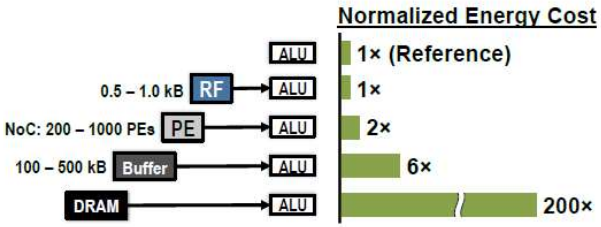
DNN ACCELERATOR HIGH LEVEL ARCHITECTURE

1 MAC operation requires 1 filter coefficient read, 1 input read, 1 partial sum read, 1 partial sum write
724M MAC for Alexnet requires ~3000 M DRAM reads !! (worst case)

Temporal Architecture (SIMD/SIMT)

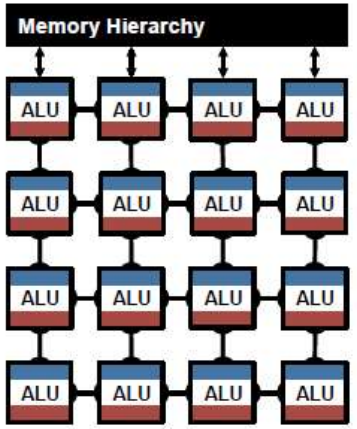


Main metric to minimize => pJ/Inference
Main bottleneck => Memory access



Memory read relative cost

Spatial Architecture (Dataflow Processing)



- Key architectural difference => Interconnect between ALUs (MAC array elements)
- Multicast support ,
- More ALUs and more local register/buffer per ALU
- Hardware-algorithm co-design necessary to exploit the new features