Leveraging sparsity to drive fast response times at the edge

Orlando Moreira, Chief Architect and Fellow
GLOBAL TEAM

Silicon Valley
- Product Marketing & Sales, CEO
- Customer Solutions

Paris
- Science Center
- System & Applications Engineering

Eindhoven
- Silicon Design Center
- SDK Engineering
Latency and Throughput

On an edge device, coupled to a camera sensor, data batching cannot be used, as it would push latency way up. (image) Data batching is used to achieve high throughput, it does not help with latency.

And do we need high throughput on an edge device?
At the Edge, Response Latency matters

- Responsive smart devices
- Closely-coupled feedback loops
- Autonomous systems

Input data streams are continuous

- Video feeds
- Audio feeds
- Industrial sensor ensembles
- Bio signals (EEG, EKG, movement)
Two components to response latency

Response latency = Inference latency + Sampling Latency

Inference latency: the time it takes to process the network from the start of an input frame. Sampling latency: the time it takes between something happening in the environment and the sensor being able capturing it.

#Insert image that shows event, sensor, inference machine, with temporal diagram.
Sampling latency

• The worst case time between an event happening and being “seen” by the sensor.
• It is equal to the sampling rate, or the inverse of the FPS.
  • By doubling FPS we halve the sampling Latency.

For 30 FPS -> 33.3 ms
For 120 FPS -> 8.3 ms
For 240 FPS -> 4.2 ms
If our inference latency is close to 10s of ms, the sampling latency will be dominant for < 120 fps #Express this in a table.

Caveat emptor:
• In traditional inference, 2x FPS -> 2x energy consumption and computation load.
• High FPS is normally achieved by pipelining, and pipelining can be bad for latency...
High throughput in single batches

High throughput with single batches is often achieved with pipelining, on a layer basis.

- reduces transactions to load weights.

Slowest stage determines the pipeline period, which will define the inference latency:

Inference latency = pipeline period x number of stages

With each stage being a layer: inference latency = period x depth

At high fps, the pipeline period will be bound by the frame period (1/fps)

Inference latency = depth / fps

Now recall that Sampling latency = 1/fps, which leads to conclude that with single batch pipelining

1) Fps is closely related to the latency;
2) Inference latency becomes dominant
3) Network depth becomes dominant for Total Latency.

So, how does one reduce inference latency?

Enter mini-batch pipelining.
Mini-batch pipelining

• **Idea:** each pipeline stage processes a mini-batch (stripe) of the input image, and the whole pipe executes in parallel.
  • There are some issues to fix, so truth is a bit more complex...

• the depth of the pipeline is still given by the layers, but the stage length is now the worst case processing time **per stripe**.

• Inference latency (worst case)= depth . (1/(fps * number of mini-batches))
  • How low can we go in terms of batches?
    • Two constraints: synchronization overhead & data dependencies (depends on kernel size).

• Thus mini-batching dramatically reduces latency while preserving high throughput.

#show the GANTT chart for mini-batch pipelining
What about Sampling Latency

Response latency = Inference latency + Sampling Latency
Mini-batch pipelining dramatically reduces inference latency, potentially making Sampling Latency dominant.
And recall that both Sampling and Inference Latency depend on FPS.
This, ultimately, truly low response latency requires high FPS.

But
Can we increase FPS without a linear increase in computation and energy requirements?
Yes. The solution is TIME SPARSITY!
Sparsity in video

- **Sparsity in structure**
  - Pruning of needless weights and kernels in network.

- **Sparsity in space**
  - Most pixels in an image have no relevant feature data.
  - Results in 0-valued activations.

- **Sparsity in time**
  - Image changes little from instant to instant
    - *why should we always re-process the whole frame?*
Transform Response Latency

...with NeuronFlow
NEURONFLOW TECHNOLOGY

Architecture Pillars

or

100x Opportunities

DYNAMIC DATAFLOW
Exploits data-dependent sparsity @Edge:
Compute only what changes, reducing load.

DIGITAL NEUROMORPHICS
Enables scalable, cost-efficient silicon design

NEAR-MEMORY COMPUTE
Allows for Mini-Batching to reduce latency
How Neuronflow supports mini-batch pipelining

So, through mini-batch pipelining, we can dramatically reduce inference latency. GML sNeuronflow architecture is designed to leverage mini-batch pipelining:

• Near-memory compute keeps parameters local to the processing cores;
• Event-based processing and push execution model:
  • simplified mini-batch self-timed synchronization
  • while exploiting **sparse execution**
NeuronFlow: Sparse and Event-Based execution model

- Exploit time-sparsity in a time series;
- Convert frame-based network to event-based inference;
- Event-based: change is sent sporadically, so no frame structure to input data;
- Only propagate changes => less work to do;
- Requires resilient neuron state;
- Threshold: how much change in neuron output warrants propagation.
- Convert from CNN by setting thresholds.

Red = active links and activated neurons
PilotNet in SparNet
Time sparsity enables low response Latency

Executing PilotNet with SparNet dramatically reduces the number of operations.

Effect becomes dramatic at high fps:
- within a time interval, a fixed amount of change occurs;
- but for **frame-based processing**, load increases linearly with frame rate;
- for **Neuronflow** time-sparse computation, load increases modestly with fps
  - And 2x fps means $\approx \frac{1}{2}$ sampling latency!
We thank the authors for their presentations and everyone who participated in the tinyML Summit 2021.

Along with a special thank you to the sponsors who made this event possible!
Executive Sponsors
Arm: The Software and Hardware Foundation for tinyML

Connect to high-level frameworks

Profiling and debugging tooling such as Arm Keil MDK

Supported by end-to-end tooling

Connect to Runtime

Optimized models for embedded

Runtime (e.g. TensorFlow Lite Micro)

Optimized low-level NN libraries (i.e. CMSIS-NN)

RTOS such as Mbed OS

Arm Cortex-M CPUs and microNPUs

Stay Connected

@ArmSoftwareDevelopers

@ArmSoftwareDev

Resources: developer.arm.com/solutions/machine-learning-on-arm
Advancing AI research to make efficient AI ubiquitous

Power efficiency
- Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization
- Continuous learning, contextual, always-on, privacy-preserved, distributed learning

Efficient learning
- Robust learning through minimal data, unsupervised learning, on-device learning

Perception
- Object detection, speech recognition, contextual fusion

Reasoning
- Scene understanding, language understanding, behavior prediction

Action
- Reinforcement learning for decision making

A platform to scale AI across the industry

Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
Samsung brings AI in the hands of everyone, with >300M Galaxy phones per year. Fingerprint ID, speech recognition, voice assistant, machine translation, face recognition, AI camera; the application list goes on and on.

In the heart of AI applications is the NPU, the neural processor that efficiently calculates AI workloads. Samsung NPU is a home grown IP that was employed since 2018 inside Samsung Exynos SoC.

Samsung NPU is brought by global R&D ecosystem that encompasses US, Korea, Russia, India, and China. In US, we are the fore-runner to guide the future directions of Samsung NPU, by identifying major AI workloads that Samsung’s NPU needs to accelerate in 3-5 years. For this, we collaborate with world-renowned academia research groups in AI and NPU.
Platinum Sponsors
Eta Compute creates energy-efficient AI endpoint solutions that enable sensing devices to make autonomous decisions in energy-constrained environments in smart infrastructure and buildings, consumer, medical, retail, and a diverse range of IoT applications.

www.etacompute.com
Lattice Semiconductor (NASDAQ: LSCC) is the low power programmable leader. We solve customer problems across the network, from the Edge to the Cloud, in the growing communications, computing, industrial, automotive and consumer markets. Our technology, relationships, and commitment to support lets our customers unleash their innovation to create a smart, secure and connected world. www.Latticesemi.com.
Gold Sponsors
AKIDA™ Neuromorphic Technology: Inspired by the Spiking Nature of the Human Brain

- Supports ultra-low power applications (microwatts to milliwatts)
- Edge capabilities: on-chip training, learning, and inference
- Designed for AI Edge applications: vision, audio, olfactory, and smart transducer applications
- Licensed as IP to be designed into SoC or as silicon
- Sensor inputs are analyzed at the point of acquisition rather than through transmission via the cloud to the data center. Enables real time response for power-efficient systems
- Software Development Platform
BabbleLabs AI speech wizardry in Cisco Webex

AI meets speech - deep experience in speech science, AI/ML, embedded systems

Massive compute

Novel deep neural networks

Massive data corpus

Silicon-optimized software

300 TFLOPS per engineer

40K hours of speech
15K hours of music
10K hour of noise
100K room models

Speech enhancement

Speech recognition

Conferencing
Call centers
Digital Assistants
Calling
DSP Group, Inc. develops wireless communications and voice processing chipsets, algorithms, and software solutions for converged communications and smart-enabled devices. Core competencies include, but are not limited to, voice processing. Its technology supports the development and integration of voice user interfaces (VUIs) for applications ranging from smartphones to the smart home. Its Ultra-Low Energy (ULE, per the ULE Alliance) wireless solutions enable low-power, long-range, secure communication applications for the IoT and are distinguished by their native support of two-way voice communication. On-going development efforts include the application of machine learning (ML) and artificial intelligence (AI) hardware and algorithms to address the need for accurate AI solutions at the edge for applications such as sound detection, proximity detection, and acoustic beacons.
TinyML for all developers

**Dataset**
- Acquire valuable training data securely
- Enrich data and train ML algorithms

**Edge Device**
- Real sensors in real time
- Open source SDK
- Embedded and edge compute deployment options

**Impulse**
- Test impulse with real-time device data flows

**Test**

www.edgeimpulse.com
The Eye in IoT

Edge AI Visual Sensors

CMOS Imaging Sensor
- Ultra Low power CMOS imager
- AI + IR capable

Computer Vision Algorithms
- Machine Learning algorithm
- <1MB memory footprint
- Microcontrollers computing power
- Trained algorithm
- Processing of low-res images
- Human detection and other classifiers

IoT System on Chip
- Machine Learning edge computing silicon
- <1mW always-on power consumption
- Computer Vision hardware accelerators

info@emza-vs.com
GrAI Matter Labs has created an AI Processor for use in edge devices like drones, robots, surveillance cameras, and more that require real-time intelligent response at low power. Inspired by the biological brain, its computing architecture utilizes sparsity to enable a design which scales from tiny to large-scale machine learning applications.

www.graimatterlabs.ai
Enabling the next generation of **Sensor and Hearable products** to process rich data with energy efficiency

- Visible Image
- Sound
- IR Image
- Radar
- Bio-sensor
- Gyro/Accel

**Wearables / Hearables**

**Battery-powered consumer electronics**

**IoT Sensors**
Himax Technologies, Inc. provides semiconductor solutions specialized in computer vision. Himax’s WE-I Plus, an AI accelerator-embedded ASIC platform for ultra-low power applications, is designed to deploy CNN-based machine learning (ML) models on battery-powered AIoT devices. These end-point AI platforms can be always watching, always sensing, and always listening with on-device event recognition.

Imagimob AI SaaS

• End-to-end development of tinyML applications
• Guides and empowers users through the process
• Support for high accuracy applications requiring low power and small memory
• Imagimob AI have been used in 25+ tinyML customer projects
• Gesture control
Adaptive AI for the Intelligent Edge

LatentAI

Latentai.com
Sensors and Signal Conditioning

Health sensors measure PPG and ECG signals critical to understanding vital signs. Signal chain products enable measuring even the most sensitive signals.

www.maximintegrated.com/sensors

Low Power Cortex M4 Micros

Large (3MB flash + 1MB SRAM) and small (256KB flash + 96KB SRAM, 1.6mm x 1.6mm) Cortex M4 microcontrollers enable algorithms and neural networks to run at wearable power levels.

www.maximintegrated.com/microcontrollers

Advanced AI Acceleration IC

The new MAX78000 implements AI inferences at low energy levels, enabling complex audio and video inferencing to run on small batteries. Now the edge can see and hear like never before.

www.maximintegrated.com/MAX78000
Qeexo AutoML

Automated Machine Learning Platform that builds tinyML solutions for the Edge using sensor data

Key Features

- Supports 17 ML methods:
  - Multi-class algorithms: GBM, XGBoost, Random Forest, Logistic Regression, Gaussian Naive Bayes, Decision Tree, Polynomial SVM, RBF SVM, SVM, CNN, RNN, CRNN, ANN
  - Single-class algorithms: Local Outlier Factor, One Class SVM, One Class Random Forest, Isolation Forest

- Labels, records, validates, and visualizes time-series sensor data

- On-device inference optimized for low latency, low power consumption, and small memory footprint applications

- Supports Arm® Cortex™-M0 to M4 class MCUs

End-to-End Machine Learning Platform

Target Markets/Applications

- Industrial Predictive Maintenance
- Automotive
- Smart Home
- Mobile
- Wearables
- IoT

For more information, visit: www.qeexo.com
Add Advanced Sensing to your Product with Edge AI / TinyML

Pre-built Edge AI sensing modules, plus tools to build your own

Reality AI solutions
- Prebuilt sound recognition models for indoor and outdoor use cases
- Solution for industrial anomaly detection
- Pre-built automotive solution that lets cars “see with sound”

Reality AI Tools® software
- Build prototypes, then turn them into real products
- Explain ML models and relate the function to the physics
- Optimize the hardware, including sensor selection and placement

https://reality.ai  info@reality.ai  @SensorAI  Reality AI
Build Smart IoT Sensor Devices From Data

SensiML pioneered TinyML software tools that auto generate AI code for the intelligent edge.

- End-to-end AI workflow
- Multi-user auto-labeling of time-series data
- Code transparency and customization at each step in the pipeline

We enable the creation of production-grade smart sensor devices.
Silicon Labs (NASDAQ: SLAB) provides silicon, software and solutions for a smarter, more connected world. Our technologies are shaping the future of the Internet of Things, Internet infrastructure, industrial automation, consumer and automotive markets. Our engineering team creates products focused on performance, energy savings, connectivity, and simplicity.  
silabs.com
Syntiant Corp. is moving artificial intelligence and machine learning from the cloud to edge devices. Syntiant’s chip solutions merge deep learning with semiconductor design to produce ultra-low-power, high performance, deep neural network processors. These network processors enable always-on applications in battery-powered devices, such as smartphones, smart speakers, earbuds, hearing aids, and laptops. Syntiant’s Neural Decision Processors™ offer wake word, command word, and event detection in a chip for always-on voice and sensor applications.

Founded in 2017 and headquartered in Irvine, California, the company is backed by Amazon, Applied Materials, Atlantic Bridge Capital, Bosch, Intel Capital, Microsoft, Motorola, and others. Syntiant was recently named a CES® 2021 Best of Innovation Awards Honoree, shipped over 10M units worldwide, and unveiled the NDP120 part of the NDP10x family of inference engines for low-power applications.

www.syntiant.com

@Syntiantcorp
TensorFlow is an end-to-end open source platform for machine learning. Our ecosystem of tools, libraries, and community resources help users push the state-of-the-art in building and deploying ML powered applications.
A DEEP TECH COMPANY AT THE LEADING EDGE OF THE AIOT

JOIN OUR SESSIONS DURING THE TINYML SUMMIT

Performing inference on BNNs with xcore.ai
Tuesday, March 23 at 12pm (PST)

TinyML: The power/cost conundrum
Thursday, March 25 at 12pm (PST)

VISIT XMOS.AI TO FIND OUT MORE
Silver Sponsors

EDGE
cortix

HotG

SynSense
Copyright Notice

The presentation(s) in this publication comprise the proceedings of tinyML® Summit 2021. The content reflects the opinion of the authors and their respective companies. This version of the presentation may differ from the version that was presented at the tinyML Summit. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyML.org