Exynos 2100
Neural Processing Unit

Samsung Semiconductor, Inc.
Neural Processing Lab
San Jose, CA

Joseph Hassoun
3/22/2021
Introducing Exynos 2100

• Started sales worldwide from January in Samsung Galaxy S20, S21+, S21 Ultra

• 5nm EUV process
  – +10% Performance
  – -20% power
Exynos 2100

- Exynos offers perfect balance among several high Performance components.
Exynos 2100 Neural Processing Unit, NPU

- 2018: 1.9 TOPS
- 2019: 15 TOPS
- Exynos 2100: 26 TOPS
Exynos 2100 NPU Applications

Object Eraser

Blur / Bokey

Hi/Lo key Mono

Before

After
Commercial Impact

- Samsung Galaxy sales are over 300M phones per year.
- SoC’s in Samsung Galaxy
  - SnapDragon : US premium
  - Mediatek : Low cost phones
  - Exynos : All the rest, usually 25% ~ 50% of Galaxy phones.
- Exynos 2100 sales estimation : 30M ~ 70M chips
- In addition, features of the Exynos 2100 NPU technology will be applied to future Exynos chips for mid level Galaxy A series, and possibly for budget Galaxy M series.
Motivation - Flexibility

- NPU needs to support a comprehensive range of NN
  - Diverse Kernel sizes, dilation rates and strides.
NPU Architecture
Convolutional Engines (CE)

- CE executes 16 dim. Data in parallel along the channel
- The smallest unit of compute is $1 \times 1 \times 16$. So utilization can be maximized with variety of kernel sizes.
Key Points

- **Challenge:**
  - To maintain a high utilization factor for these diverse convolution while maintaining high energy efficiency

- **Solution:**
  - Serialize the convolutional operation along the spatial direction
  - Skip redundant operations by exploiting sparseness.
Dilated Convolution
Feature-map-aware Zero skipping

Benefits of feature-map-aware zero skipping

- Effective performance
- Energy efficiency
- HW Utilization
Measurement Results

- NPU achieves 623 inferences/s at 1196 MHz in multi-thread mode
- NPU achieves energy efficiency of 0.84 mJ/inference (1196 Inf/J)
  - Corresponds to 13.6 TOPs /W for Inception-V3 including DMA power.
Die Photo

<table>
<thead>
<tr>
<th>Process</th>
<th>5nm CMOS technology (Samsung)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area</td>
<td>5.46 mm²</td>
</tr>
<tr>
<td>Voltage</td>
<td>0.55-to-0.9V</td>
</tr>
<tr>
<td>Frequency</td>
<td>332-to-1196-MHz</td>
</tr>
<tr>
<td>Best Peak Performance</td>
<td>623 inferences/s @ 0.9V (Inception V3)</td>
</tr>
<tr>
<td>Best Energy Efficiency</td>
<td>13.6 TOPS/W @ 0.6V (Inception V3)</td>
</tr>
</tbody>
</table>
Summary

- Adder-tree-based datapath and serialized convolutional operations achieves high utilization in MAC array.
- Feature-map-aware zero-skipping for high performance and energy efficiency.
- Reduced memory footprint and bandwidth via weight and Feature map compression.
- Parallelization of DMA and MAC compute time by fast resource scheduling.
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
Optimized models for embedded
Application

Profiling and debugging tooling such as Arm Keil MDK

Supported by end-to-end tooling

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

Connect to high-level frameworks

Connect to Runtime

Runtime (e.g. TensorFlow Lite Micro)

AI Ecosystem Partners

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

IoT/IoT
- Edge cloud

Automotive

Mobile

Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
NEURAL PROCESSING

- 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.
**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](http://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
  - 300 TFLOPS per engineer
- Novel deep neural networks
- Silicon-optimized software
- Massive data corpus
  - 40K hours of speech
  - 15K hours of music
  - 10K hours 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

Acquire valuable training data securely

Real sensors in real time
Open source SDK

Embedded and edge compute deployment options

Enrich data and train ML algorithms

Test impulse with real-time device data flows

www.edgeimpulse.com
The Eye in IoT

Edge AI Visual Sensors

- Ultra Low power CMOS imager
- AI + IR capable
- Machine Learning edge computing silicon
- <1mW always-on power consumption
- Computer Vision hardware accelerators

- Machine Learning algorithm
- <1MB memory footprint
- Microcontrollers computing power
- Trained algorithm
- Processing of low-res images
- Human detection and other classifiers
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.
Enabling the next generation of **Sensor and Hearable products** to process rich data with energy efficiency.
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
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
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
- Smart Home
- Wearables
- Automotive
- Mobile
- IoT

For more information, visit: www.qeexo.com
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.

sensiml.com
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](http://www.syntiant.com)  
[@Syntiantcorp](https://twitter.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.
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
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