9.2 A 1mW Always-on Computer Vision Deep Learning Neural Decision Processor

David Garrett, Youn Sung Park, Seongjong Kim, Jay Sharma, Wenbin Huang, Majid Shaghaghi, Vinay Parthasarathy, Stephen Gibellini, Stephen Bailey, Mallik Moturi, Pieter Vorenkamp, Kurt Busch, Jeremy Holleman, Behrooz Javid, Alireza Yousefi, Mohsen Judy, Atul Gupta

Syntiant, Irvine, CA

Syntiant®'s NDP200 is a specially architected edge AI processor, bringing the ultra-low-power, high performance, deep neural network processing capabilities of Syntiant Core 2™ (SC2) to computer vision (CV) applications. Ideally suited for low-latency inferences on power-constrained devices such as doorbell cameras, the NDP200 is equipped with an 8-bit direct video port (DVP), I2C, and SPI interfaces to interface directly with low-power image sensors, as well as Pulse Density Modulation (PDM) and Inter-IC Sound interface (I2S) to facilitate use cases requiring both video and audio processing.

As seen in Fig. 9.2.1, the largest component in the device is the SC2 neural decision processor. The rest of the device is dedicated to communicating and extracting information from sensors (image, audio, motion) to feed the neural network engine, and then deploy deep learning algorithms. The device includes a HiFi 3 Digital Signal Processing (DSP) block which can be used for feature extraction or image signal processing, as well as a deeply embedded Cortex M0 microcontroller used for supervisory functions, controlling SPI, I2C, GPIO, timers, and UART functionalities of the device. An integrated Phase-Locked Loop (PLL) derives timing from a 32kHz external XTAL. The device can deploy additional security with 128-bit AES-CCM encrypted, authenticated firmware and neural network packages where the encryption key is derived from a 3-level key-derivation ladder using the AES-CMAC algorithm [1]. Using the irreversible AES-CMAC function with a secret key, the device can use a clear-text 128-bit one-time-programmable (OTP) key that does not reveal the actual encryption keys, and allows the device to derive different keys for the bootloader and encrypted neural network packages.

The neural processing is designed with at-memory compute, directly mapping the on-board memories to connect with the computation datapath, minimizing the movement of data. Unlike traditional neural accelerators that work in concert with a microcontroller or DSP, the SC2 neural decision processor is designed to directly execute the graph of the network, with params, data and instruction memories built into the block. Figure 9.2.2 demonstrates the neural decision processor data with 1MB of wide-bandwidth SRAM deployed in the block, directly bus-mapped to the SoC address space using an AHB interface. Internally the memories are configured with wide-memory parallelized interfaces supporting almost 10Gbyte/s memory bandwidth to sustain high efficiency usage of the 32 parallel Multiply-Accumulators (MAC). In every clock cycle, the datapath can retrieve 32 activations and 32 params to feed the MACs, while simultaneously writing back 32 new activations. Separately, the controller executes a linked list of layer instructions, and works to coordinate the movement of data that is optimized for each layer, for example with different strategies of params and activation readout for convolutional 2D and dense layers. The total memory is partitioned between 640kB of params space, and 384kB of neural network data space split over two buffers, with an aliasing option to expand the params space up to 896kB through memory reuse. SC2 supports 32 MAC units per clock with 8-bit by 8-bit performance, or 16 parallel 16-bit by 8-bit MACs. The params for each layer can be stored in memory with a wide range of quantization options, including 1-bit, 2-bit, 4-bit and 8-bit params quantization schemes, as well as specialized 4-bit quantization formats, including a square-root and exponent mapping formats, pushing the maximum neural network sizes up to 7 million, 1-bit weights. The lower precision params are expanded on-the-fly to feed the MAC units in order to reduce power in the memory fetch and datapath computation. On the computation side, SC2 is capable of accommodating almost all of the layers needed to build energy efficient networks at the edge, including dense, convolutional 2D, depth-wise convolutional, avg/max pooling, batch-norm, point-wise multiplication, and point-wise addition. Furthermore, the architecture can support Recurrent Neural Networks (RNN) such as a Gated-Recurrent Unit (GRU) with customized ReLU, tanh and sigmoid activation functions built into the datapath.

The SC2 uses a linked-list execution of layers and in the case of standard networks, like MobileNetV1 [2], can completely finish the entire inference without intervention. As in Fig. 9.2.2, the controller just receives a layer index to start execution, and then provides an interrupt when the entire inference is complete. In this case, the rest of the system can go into a low power state waiting for the neural network inference to finish. For computer vision, SC2 can also support MobileNetV2 that builds on MobileNetV1 layer structure using residual connections [3]. For other neural architectures, the neural datapath can wakeup the DSP using an interrupt to provide additional compute, custom activations, movement of data, or launch independent neural networks. The DSP can access the internal neural core memories (both params and activations) during inference using a memory arbitration scheme. The embedded DSP supports 96kB of instruction and 192kB data memory with 64-bit wide tightly-coupled-memory. It has been built with instruction-set extensions to provide additional support for neural network execution, including 2^x, log2, tanh, and sigmoid functions. The DSP can coordinate closely with SC2 for implementation of networks and can assist in functions like high-precision GRU layers (like 32-bit cell-state accumulators).

The performance of the NDP200 can be seen when analyzing the performance of a 128×128 MobileNet V1 0.25 network running on the NDP200 vs the same network mapped to an Cortex A53. The SC2 in NDP200 achieves 740k cycles per inference, while the A53 requires over 21M cycles to complete an inference on the same network. By using data that is stored channel-first (as seen in Fig. 9.2.2), SC2 can retrieve and write-back 32 parallel activations per cycle, while simultaneously retrieving 32 params to keep feeding the 32 parallel MACs. Once the number of channels in the layer tensors are >= 32, the datapath can routinely execute greater than 90% efficiency on convolutional layers in the neural network. While depth-wise convolutional layers in this region are only 70% efficient, the depth-wise cycle count in the network is only 17% of the total cycle count, so it has a less impact on the overall efficiency. When running this MobileNetV1 0.25 network on NDP200 at 5 frames per second, the device consumes on average only 0.83mW of power when supplied from a 0.9V core supply.

The NDP200 is supported with both Syntiant's Software Development Kit (SDK), which allows for pre-trained models to be deployed on the device, and Syntiant's Training Development Kit (TDK), which enables training and packaging of networks. By directly executing the neural network graphs in the SC2, the TDK offers models of both execution times and energy consumption for all layers, and can be used to estimate the energy performance on the device, providing engineers with a quick feedback loop to optimize their networks and system performance. As seen in Fig. 9.2.3, the NDP200 evaluation system is built around a Raspberry Pi 3B+ and includes a PixArt PAG7920 QVGA image sensor. Figure 9.2.4 shows an exemplary captured image on a 320x240 gray-scale image from the platform while running a person detection model with bounding box detection.

An example of a training reinforcement learning network use-case is shown in Fig. 9.2.5 where the NDP200 was trained to play a game of Doom [4]. Using the VizDoom platform [5] to create a reinforcement learning environment, the SC2 neural core was trained to observe a 4-frame stack of 64×64 pixel images from the game screen and provide game inputs from movement, rotation and shooting actions. The neural network was optimized to achieve the maximum number of monsters killed with a limited amount of ammunition. The neural network architecture was a 5-layer network, with 3 convolutional layers, followed by 2 Dense layers for final decision. With over 606k params, the SC2 consumes 1.0mW in the core network to achieve 35 frames-per-second in the game.

The SC2 power efficiency can be seen in Fig. 9.2.6 which shows the V0.7 results from the TinyML Commons benchmark on the keyword spotting task using an DS-CNN neural network [6]. Using a NDP120 evaluation system (which uses the same neural decision processor as the NDP200), the SC2 device is leading the categories in inference time and energy all while using the lowest operating frequency. The NDP200 is fabricated in an ultra-low power 40nm CMOS technology and is offered in a 3.1mm × 2.5mm package as illustrated in Fig. 9.2.7.

Acknowledgement:
The authors would like to acknowledge all the efforts from the Syntiant teams, from machine learning, software, embedded systems, and operations.

References:

[1] JH Song et. al., "The AES-CMAC Algorithm, *RFC 4493*, June 2006.
[2] Andrew Howard et. al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, arXiv:1704.04861, April 2017.
[3] Mark Sandler et. al., "MobileNetV2: Inverted Residuals and Linear Bottlenecks", arXiv:1801.04381v4, May 21, 2019.
[4] David Garrett, "NDP200 tinyML Vision Processing", *2022 tinyML Summit*, March 2022
[5] M Wydmuch, M Kempka and W Jaśkowski, ViZDoom Competitions: Playing Doom from Pixels, *IEEE Transactions on Games*, arXiv:1809.03470.
[6] ML Commons Tiny Inference V0.7 Results, https://mlcommons.org/en/inference-tiny-07, April 2022.

2023 IEEE International Solid-State Circuits Conference (ISSCC) | 978-1-6654-9016-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISSCC42615.2023.10067588

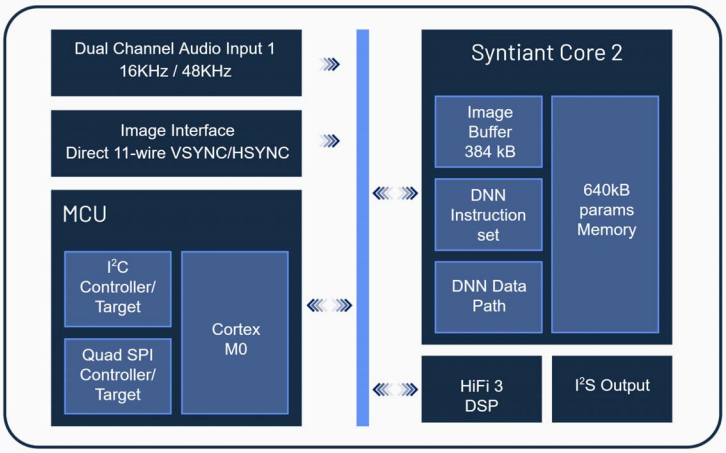

Figure 9.2.1: NDP200 High Level Block Diagram.

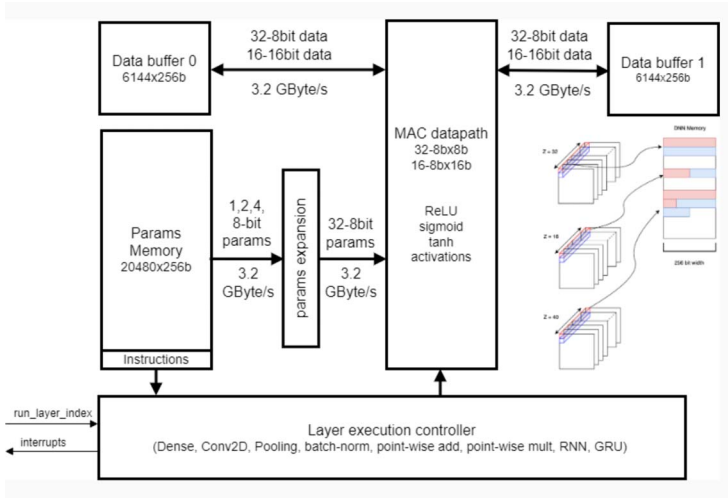

Figure 9.2.2: Syntiant Core 2 Datapath.

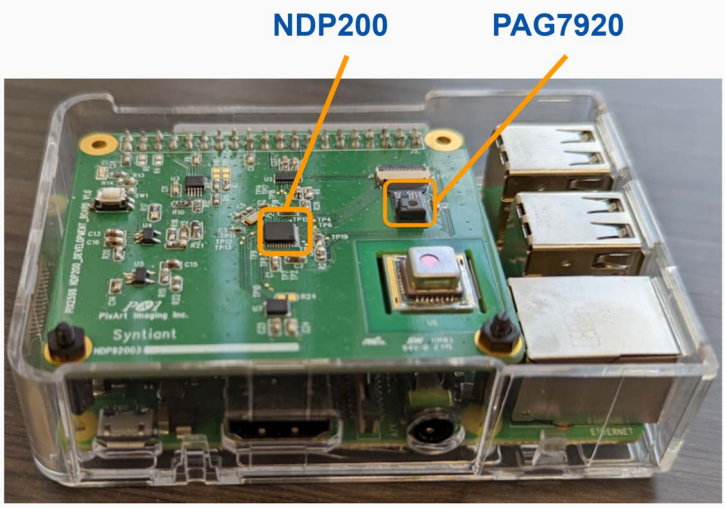

Figure 9.2.3: NDP200 Development Board.

Figure 9.2.4: Person Detection running on NDP200 (320×240 image).

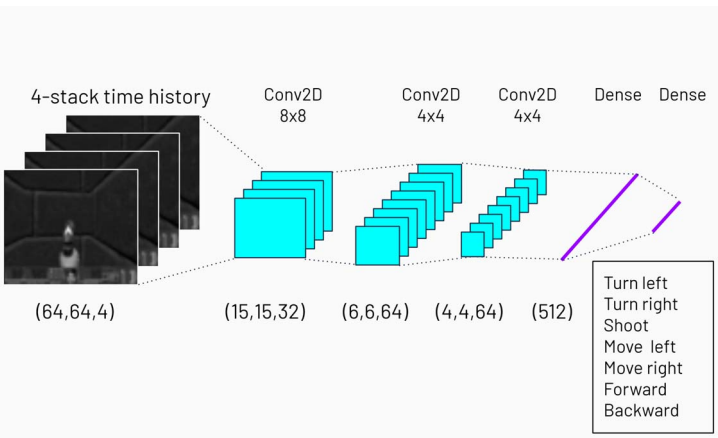

Figure 9.2.5: Neural Architecture to Play Doom.

Platform	Delay (ms)	Energy (uJ)	Energy-Delay Prod (pJ·sec)	Freq (MHz)
ST NUCLEO-L4R5ZI	97.7	4635	452.67	120
Renesas RX65N-Cloud-Kit	81.9	4422	362.00	120
Renesas EK-RA6M4	50.6	3796	192.01	200
ST NUCLEO-H7A3ZI-Q	22.3	3713	82.77	280
ST NUCLEO-U575ZI-Q	54.8	1482	81.25	160
Syntiant NDP9120 1.1V (Syntiant Core 2)	1.8	35	0.15	99
Syntiant NDP9120 0.9V (Syntiant Core 2)	4.3	50	0.09	31

Figure 9.2.6: ML Commons Tiny Inference V0.7 Results for Keyword Spotting.

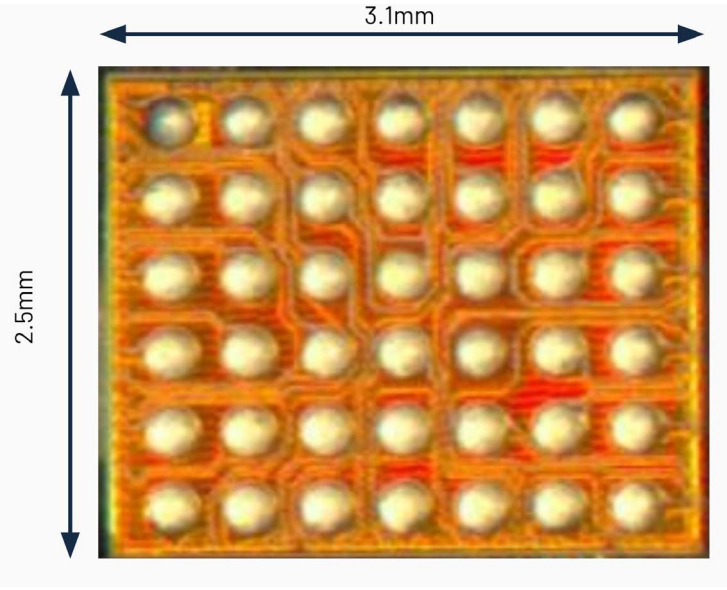

Figure 9.2.7: NDP200 Chip Micrograph.

978-1-6654-9016-0/23/$31.00 ©2023 IEEE