Title: Neural processing unit

URL: https://en.wikipedia.org/wiki/Neural_processing_unit

PageID: 50827978

Categories: Category:Application-specific integrated circuits, Category:Computer optimization, Category:Coprocessors, Category:Deep learning, Category:Gate arrays, Category:Neural processing units

---

A neural processing unit (NPU), also known as AI accelerator or deep learning processor , is a class of specialized hardware accelerator [1] or computer system [2] [3] designed to accelerate artificial intelligence (AI) and machine learning applications, including artificial neural networks and computer vision .

Use

Their purpose is either to efficiently execute already trained AI models (inference) or to train AI models. Their applications include algorithms for robotics , Internet of things , and data -intensive or sensor-driven tasks. [4] They are often manycore or spatial designs and focus on low-precision arithmetic, novel dataflow architectures , or in-memory computing capability. As of 2024 [update] , a typical datacenter-grade AI integrated circuit chip, the H100 GPU, contains tens of billions of MOSFETs . [5]

Consumer devices

AI accelerators are used in mobile devices such as Apple iPhones , AMD AI engines [6] in Versal and NPUs, Huawei , and Google Pixel smartphones, [7] and seen in many Apple silicon , Qualcomm , Samsung , and Google Tensor smartphone processors. [8]

It is more recently (circa 2022) added to computer processors from Intel , [9] AMD , [10] and Apple silicon. [11] All models of Intel Meteor Lake processors have a built-in versatile processor unit (VPU) for accelerating inference for computer vision and deep learning. [12]

On consumer devices, the NPU is intended to be small, power-efficient, but reasonably fast when used to run small models. To do this they are designed to support low-bitwidth operations using data types such as INT4, INT8, FP8, and FP16. A common metric is trillions of operations per second (TOPS), though this metric alone does not quantify which kind of operations are being done. [13]

Datacenters

Accelerators are used in cloud computing servers, including tensor processing units (TPU) in Google Cloud Platform [14] and Trainium and Inferentia chips in Amazon Web Services . [15] Many vendor-specific terms exist for devices in this category, and it is an emerging technology without a dominant design .

Graphics processing units designed by companies such as Nvidia and AMD often include AI-specific hardware, and are commonly used as AI accelerators, both for training and inference . [16]

Programming

Mobile NPU vendors typically provide their own application programming interface such as the Snapdragon Neural Processing Engine. An operating system or a higher-level library may provide a more generic interface such as TensorFlow Lite with LiteRT Next (Android) or CoreML (iOS, macOS).

Consumer CPU-integrated NPUs are accessible through vendor-specific APIs. AMD (Ryzen AI), Intel (OpenVINO), Apple silicon (CoreML) [a] each have their own APIs, which can be built upon by a higher-level library.

GPUs generally use existing GPGPU pipelines such as CUDA and OpenCL adapted for lower precisions. Custom-built systems such as the Google TPU use private interfaces.

Notes

References

External links

Nvidia Puts The Accelerator To The Metal With Pascal , The Next Platform

Eyeriss Project , MIT

v

t

e

Universal Turing machine

Parallel computing

Distributed computing

GPU GPGPU software DirectX

GPGPU software

DirectX

Audio

Digital signal processing

Hardware random number generation

Neural processing unit

Cryptography TLS

TLS

Machine vision

Custom hardware attack scrypt

scrypt

Networking

Data

High-level synthesis C to HDL

C to HDL

FPGA

ASIC

CPLD

System on a chip Network on a chip

Network on a chip

Dataflow

Transport triggered

Multicore

Manycore

Heterogeneous

In-memory computing

Systolic array

Neuromorphic

Programmable logic

Processor design chronology

design

chronology

Digital electronics

Virtualization Hardware emulation

Hardware emulation

Logic synthesis

Embedded systems