

# On-device NPU terminology cheat sheet

\* Reverse-engineered Apple M1 findings. A dash means no validated mapping in the reviewed sources.

| Concept & meaning                                                                                                                               | MediaTek                                                     | Qualcomm                      | Apple                         | Intel                        | AMD                         |
|-------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|-------------------------------|-------------------------------|------------------------------|-----------------------------|
|                                                                                                                                                 | Genio MT8370                                                 | Hexagon                       | M1 Neural Engine              | Core Ultra 200S/HX           | XDNA / XDNA 2               |
| 01 Hardware                                                                                                                                     | Compute, conversion, memory, data movement and interconnects |                               |                               |                              |                             |
| Accelerator                                                                                                                                     | APU / APUSYS                                                 | Hexagon NPU                   | Neural Engine                 | Intel NPU                    | XDNA NPU                    |
| The neural-network accelerator as a whole. Family names identify hardware, not a common microarchitecture.                                      | AI Processor Unit                                            |                               |                               |                              |                             |
|                                                                                                                                                 | 1                                                            | 2                             | 5                             | 7                            | 9                           |
| Compute grouping                                                                                                                                | MDLA + VP6                                                   | Scalar + vector + tensor      | Neural Engine core            | NCE tile                     | AI Engine tile              |
| How computing resources are grouped. A core, tile or engine has a different scope on each platform.                                             | Specialized engines                                          | Fused accelerators            |                               | Neural Compute Engine        | Spatial array               |
|                                                                                                                                                 | 1                                                            | 2                             | 5                             | 7                            | 9 11                        |
| Matrix / tensor engine                                                                                                                          | MDLA convolution engine                                      | HMX                           | MAC array*                    | DPU                          | AI Engine vector unit       |
| Hardware that performs many multiply-accumulate operations for matrix products and convolutions.                                                | Within the MDLA accelerator                                  | Matrix eXtension              | Multiply-accumulate datapath  | Data Processing Unit         | Matrix MAC operations       |
|                                                                                                                                                 | 1                                                            | 4                             | 6                             | 15                           | 9 10                        |
| Compute precision                                                                                                                               | Fix8 / Fix16 / FP16 / BF16                                   | FP16 (HMX V81 example)        | FP16 operands*                | INT8 / FP16                  | INT8 / BF16                 |
| Numeric formats used for multiply operands. Documented examples, not full support lists; accumulation and stored weights may use other formats. | MDLA; Fix = fixed point                                      | 37-bit float accumulation     | Wider accumulation reported   | NPU 3 MAC arrays             | XDNA 2; also Block FP16     |
|                                                                                                                                                 | 1                                                            | 4                             | 6                             | 16                           | 10 17 27                    |
| Vector processor                                                                                                                                | VP6                                                          | HVX                           | -                             | ACT-SHAVE                    | AI Engine vector processor  |
| A programmable processor that applies one instruction across multiple data elements (SIMD).                                                     | Vision-P6 DSP                                                | Vector eXtensions             |                               | Activation SHAVE DSP         | SIMD                        |
|                                                                                                                                                 | 1                                                            | 3                             |                               | 8                            | 9                           |
| Activation / elementwise                                                                                                                        | Element-wise engine                                          | HMX convert state             | Elementwise / planar*         | Activation Function          | AI Engine vector processor  |
| Resources for pointwise arithmetic and activation functions. Fixed-function stages and programmable vector processors differ.                   | Within MDLA                                                  | Activation operations         | Activation lookup table       | Inference-pipeline stage     | Elementwise kernels         |
|                                                                                                                                                 | 1                                                            | 4                             | 6                             | 16                           | 9 18                        |
| Numeric conversion                                                                                                                              | EDMA numeric conversion                                      | HMX convert transfer          | FP16 output rounding*         | Data Conversion              | SRS path / to_vector        |
| Changes numeric representation or range through format conversion, rounding or saturation. Available operations differ.                         | Fi6 / F32                                                    | Scale / round for FP16 output | After accumulation            | Inference-pipeline stage     | Shift-round-saturate        |
|                                                                                                                                                 | 1                                                            | 4                             | 6                             | 16                           | 24 25                       |
| Control & scheduling                                                                                                                            | CMDE                                                         | Hexagon scalar core           | Firmware scheduler*           | LEON scheduler               | RISC scalar processor       |
| Logic that decodes or schedules work. These examples range from command engines to scalar processors.                                           | MDLA command engine                                          |                               | Task queues                   | DPU / ACT-SHAVE dispatch     | Inside each AI Engine tile  |
|                                                                                                                                                 | 1                                                            | 2 3                           | 6                             | 8                            | 9                           |
| Local data memory                                                                                                                               | VP6 data RAM                                                 | VTCM                          | On-chip SRAM*                 | CMX SRAM                     | Local data memory           |
| Storage near the compute units for operands or intermediate data. Caches and scratchpads have different management rules.                       | Local to the vector DSP                                      | Vector tightly coupled memory | Compiler-managed scratchpad   | Connection MatriX            | Compute / mem tiles         |
|                                                                                                                                                 | 1                                                            | 3 4                           | 6                             | 7                            | 9 11                        |
| Instruction storage                                                                                                                             | IMEM + instruction cache                                     | L1 instruction cache          | -                             | Shared instruction L2 cache  | Program memory              |
| Stores or caches code for programmable compute processors. Program RAM and hardware-managed instruction caches differ.                          | VP6                                                          | Hexagon core                  |                               | ACT-SHAVE                    | Inside each compute tile    |
|                                                                                                                                                 | 1                                                            | 19                            |                               | 15                           | 20                          |
| Data movement                                                                                                                                   | EDMA                                                         | VMEM loads / stores           | Tile DMA / kernel DMA*        | DMA engines                  | Tile DMA / shim DMA         |
| Hardware paths that transfer data between memory and compute resources; transfers may overlap computation.                                      | External DMA; format conversion                              | External DMA via AXI          | Operands, results and weights | DRAM / CMX transfers         | Memory transfers            |
|                                                                                                                                                 | 1                                                            | 3                             | 6                             | 7                            | 11 12                       |
| Interconnect                                                                                                                                    | APU NoC                                                      | AXI slave port                | -                             | -                            | AXI stream switch fabric    |
| Connections carrying data between engines and memory. Memory-mapped interfaces and streaming fabrics use different protocols.                   | On-chip network                                              | External DMA interface        |                               |                              | Tile-to-tile streams        |
|                                                                                                                                                 | 1                                                            | 3                             |                               |                              | 12                          |
| Layout transformation                                                                                                                           | Transpose engine                                             | HVX shuffle / deal            | Transpose configuration*      | -                            | Tile DMA address generation |
| Changes tensor or vector element order. Dedicated transpose units, shuffle instructions and DMA addressing have different scopes.               | C-W / Depth2Space / Space2Depth                              | vshuff / vdeal                | Task-descriptor fields        |                              | to_stream / from_stream     |
|                                                                                                                                                 | 1                                                            | 23                            | 6                             |                              | 22                          |
| Weight compression                                                                                                                              | Compressed weights                                           | -                             | Palette / sparse weights*     | In-line weight decompression | -                           |
| Compact weight representations reduce storage or transfer volume. Compression formats do not determine compute precision.                       | MDLA; codec unspecified                                      |                               | M1 native streaming formats   | DMA engines                  |                             |
|                                                                                                                                                 | 1                                                            |                               | 6                             | 7                            |                             |
| Address translation                                                                                                                             | IOMMU                                                        | Hexagon MMU                   | DART*                         | MMU                          | Host IOMMU                  |
| Maps virtual or device addresses to physical memory. Processor MMUs and host IOMMUs serve different address spaces.                             | APUSYS memory path                                           | Processor address space       | ANE IOMMU                     | NPU memory path              | PASID-based isolation       |
|                                                                                                                                                 | 1                                                            | 19                            | 6                             | 16                           | 20                          |

|                                                                                                                         |                                                                |                                    |                            |                          |                          |
|-------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|------------------------------------|----------------------------|--------------------------|--------------------------|
| 02 Programming model                                                                                                    | Execution, transfer records, tensor blocks and synchronization |                                    |                            |                          |                          |
| Execution model                                                                                                         | VP6: VLIW + SIMD                                               | Hexagon / HVX: VLIW                | Hardware task descriptors* | DPU / ACT-SHAVE work     | Tile kernels + dataflow  |
| How instructions or tasks express parallel work at the hardware boundary. These are distinct programming models.        | MDLA: command-driven                                           | HMX: matrix instructions           | Register configuration     | Scheduled hardware tasks | VLIW / SIMD within tiles |
|                                                                                                                         | 1                                                              | 2 3 4                              | 6                          | 7 8                      | 9 11                     |
| Transfer descriptors                                                                                                    | -                                                              | External DMA descriptor            | Hardware task descriptor*  | vpu_dma_descriptor       | Buffer descriptor (BD)   |
| Records configuring a transfer or a combined compute/DMA task. Scope and record layout depend on the engine.            |                                                                | Polled by the scalar core          | Includes DMA configuration | NPU37XX firmware record  | Tile DMA transfer record |
|                                                                                                                         |                                                                | 3                                  | 6                          | 26                       | 12                       |
| Tensor blocking                                                                                                         | -                                                              | Activation crouton                 | Output-channel group*      | -                        | Tile / micro-tile        |
| Partitions of tensor data processed or transferred together. The blocked axes, formats and sizes differ.                |                                                                | HMX operand block                  | OCG                        |                          | Kernel data blocks       |
|                                                                                                                         |                                                                | 4                                  | 6                          |                          | 18                       |
| Vector element selection                                                                                                | -                                                              | Vector predicate                   | -                          | -                        | aie::mask                |
| Per-element comparison results used to select vector values. Architectural registers and API types are distinct levels. |                                                                | Q registers / vmux                 |                            |                          | Vector element mask type |
|                                                                                                                         |                                                                | 3                                  |                            |                          | 17 21                    |
| Synchronization                                                                                                         | -                                                              | External DMA sync                  | Task event masks*          | Hardware barriers        | Hardware locks           |
| Mechanisms that order compute and data transfers so a consumer uses data after its producer completes.                  |                                                                | Store-release / descriptor polling | Wait / signal fields       | Task dependencies        | Core / DMA coordination  |
|                                                                                                                         |                                                                | 3                                  | 6                          | 7                        | 12                       |

## Notes

**Apple: independent research**  
Marked entries summarize Bryngelson's M1 reverse engineering. Event masks describe on-die ordering; the vector-processor and vector-selection entries remain unmapped.

Column order uses smartphone SoC shipments (Q2 2026) and AI-PC data (2025) as separate proxies. Combined order is editorial; NPU usage is not measured.

Examples: MT8370; Hexagon V73 HVX / V81 HMX; Apple M1; Intel Core Ultra 200S/HX; AMD XDNA / XDNA 2.

## Keep tile meanings separate

Compute tiles are hardware. Croutons, output-channel groups and matrix tiles partition tensor data along different axes. Their sizes and scopes differ.

## Precision has several meanings

INT8 is 8-bit integer. FP16 and BF16 (bfloat16) differ; Block FP16 uses shared exponents. Operand, accumulator and stored-weight precision are separate.