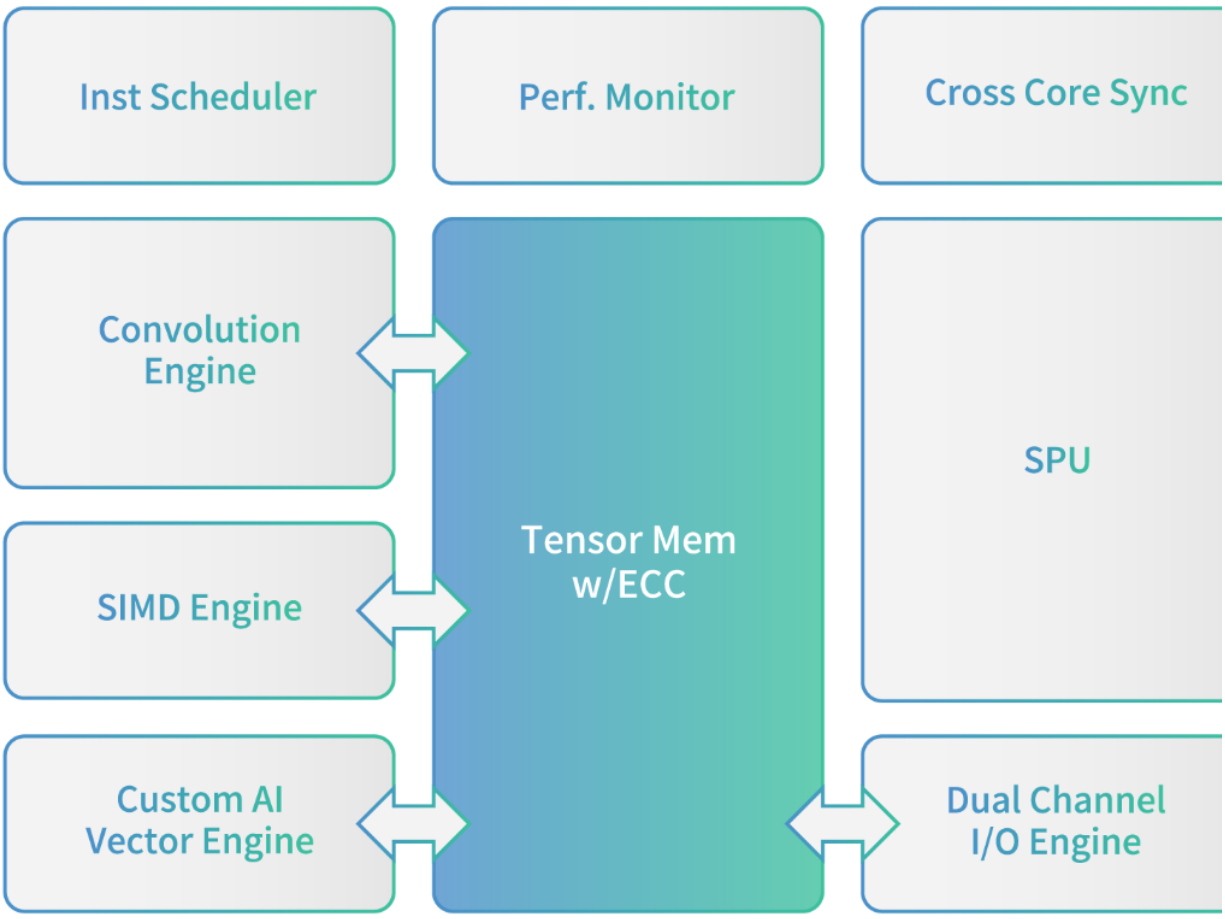
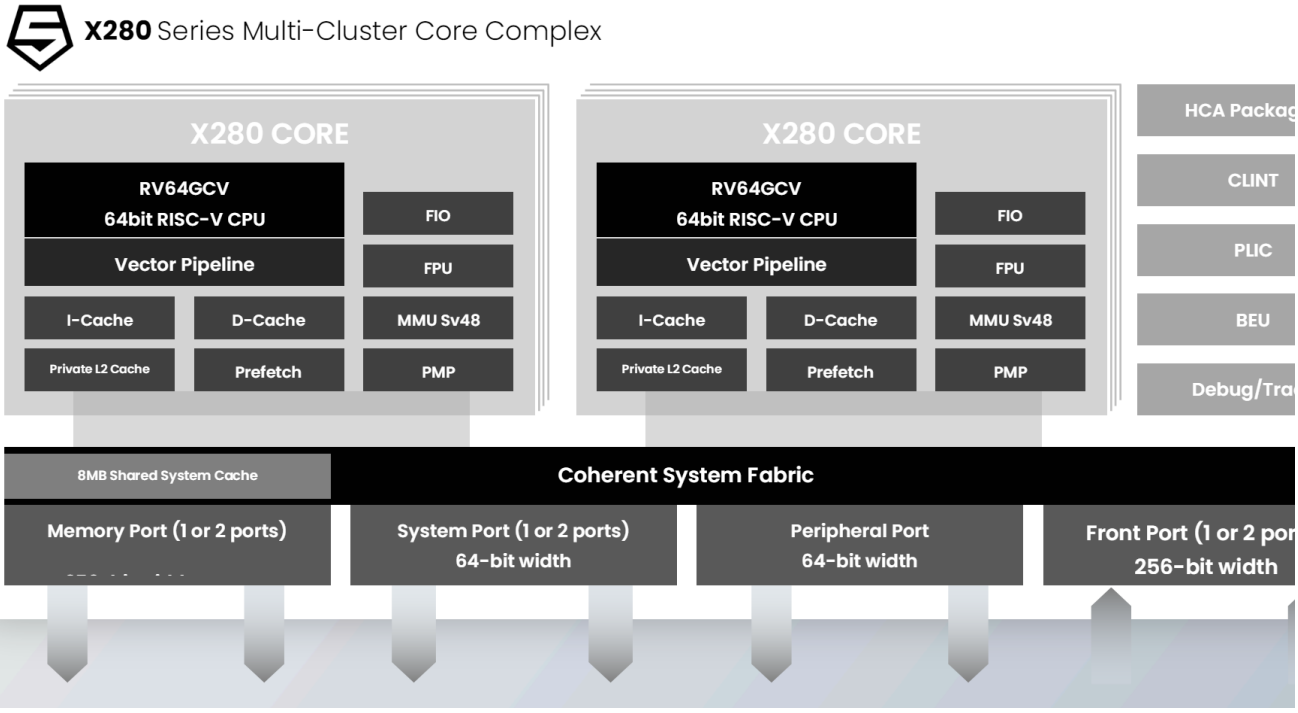
???????

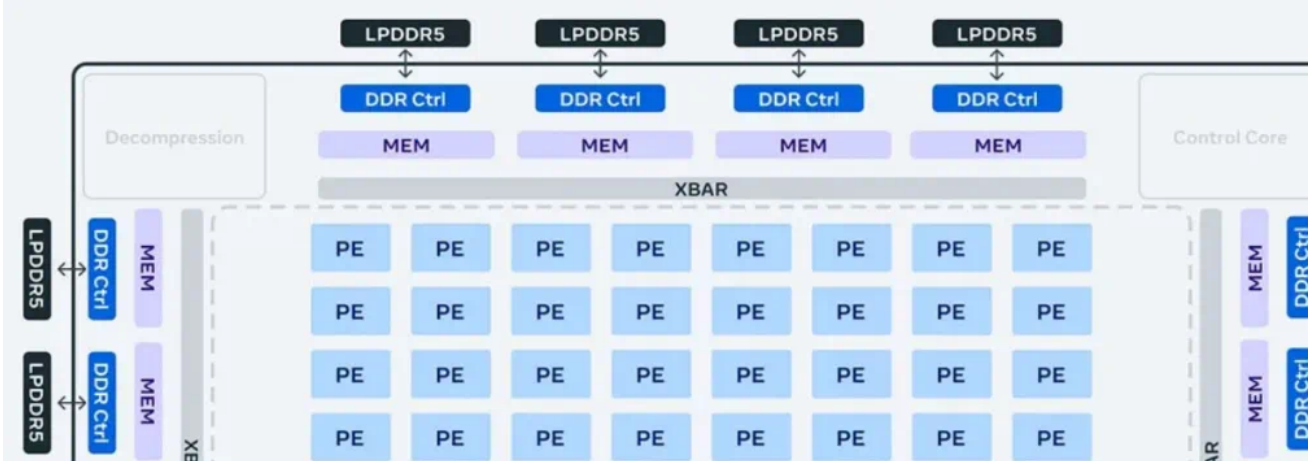
1. NVIDIA Jetson Orin
2. HAILO Hailo-8 15 等等 domain-specific-dataflow-processing
   - 据悉，它能够在功耗低于 5W的情况下，以每秒生成 10 个 token（ TPS）的速度流畅运行 Llama2-7B模型。同时，在处理 Stable Diffusion 2.1 模型（该模型可根据文本提示生成图像）时， Hailo-10 同样在超低功耗范围内实现了每张图像评级低于 5秒的优秀表现。 Hailo-10的性能高达每秒 40 万亿次（ TOPS）。根据最新公布的基准测试数据， Hailo-10 在性能上超越了集成神经处理单元（ NPU）解决方案，同时在功耗方面表现更为出色。与 Intel 的 Core Ultra NPU相比， Hailo-10在保持功耗减半的同时，实现了至少 2倍的性能提升。
3. 算能 1684X 支持高达 17.6T 的 INT8峰值算力 LPDDR4x 68.3GB/s，最大支持 16GB 17W
4. 爱芯元智
5. 地瓜机器人 https://d-robotics.cc
6. 聆思科技 https://www.listenai.com/products/chips/csk6
7. AMD旗下的 Versal自适应片上系统 (SoC)产品升级全新第二代，包括 面向 **AI** 驱动型嵌入式系统第二代的 **Versal AI Edge**系列、面向经典嵌入式系统的第二代 **Versal Prime**系列。
8. Sophgo SG2380
9. 超星未来 https://www.novauto.com.cn/



10. sifive-intelligence-x280
the Vector Coprocessor Interface Extension (VCIX)
现代工作负载和应用程序通常需要最高的性能，但需要在有限的功率环境中实现。由于标准
RISC-V Vector ISA 和 SiFive Intelligence Extensions
提供的出色矢量计算能力，设计人员能够将各种专用 DSP 加速器功能整合到单个 X280
处理器设计中，从而保持更简单的系统设计和易编程性，同时仍能实现所需的性能和效率目标
。
在某些情况下，例如需要对矢量数据进行高度密集计算的工作负载，设计人员仍然需要使用经
过高度优化的定制加速器来从主处理器卸载这些任务。这种方法的挑战在于，定制加速器需要
设计为与主处理器共存，这会带来一些微妙的设计后果，即加速器：
RVV The X280 processor implements a 512-bit vector length architecture (VLEN), fully supporting the vector extension standard, with dynamic variable vector length operations. The vector ALU and load/store architecture data width (DLEN) is 256-bits.

- 

11. □□□□□□□□ AI□□ DeepEdge10□□□□□□□□□□□□□□ NNP400T
    - 

12. Meta□ MTIA□□□□□□□ 256MB □□□□□□□□ 1.3GHz□□ v1 □□□□□□ 128MB □ 800GHz□□□□□□□□□□□□ 8x8 □□□□□ (PE) □□□□□ PE □□□□□□□□□□□□□ MTIA v1 □□□ 3.5 □□□□□□□□□□□□□ 7 □□□□□□□□□□□□□□□□□□□□□□□□□□□□ PE □□□□□□□□ Meta □□□ PE □□□□□□□□□□□□□ SRAM □□□□□□□□□□□□□ 3.5 □□□□ LPDDR5 □□□□□□□□□
    - 

First Gen MTIA

Technology
TSMC 7nm

Frequency
800MHz

Instances
1.12B gates, 65M flops

Area
19.34mm x 19.1mm, 373mm2

Package
43mm x 43mm

Voltage
0.67V logic, 0.75V memory

TDP
25W

Host Connection
8x PCIe Gen4 (16 GB/s)

GEMM TOPS
102.4 TFLOPS/s (INT8)
51.2 TFLOPS/s (FP16/BF16)

SIMD TOPS
Vector core:
3.2 TFLOPS/s (INT8),
1.6 TFLOPS/s (FP16/BF16),
0.8 TFLOPS/s (FP32)
SIMD:
3.2 TFLOPS/s (INT8/FP16/BF16),
1.6 TFLOPS/s (FP32)

Memory Capacity
Local memory: 128 KB per PE
On-chip memory: 128 MB
Off-chip LPDDR5: 64 GB

Memory Bandwidth
Local memory: 400 GB/s per PE
On-chip memory: 800 GB/s
Off-chip LPDDR5: 176 GB/s

Next Gen MTIA

Technology
TSMC 5nm

Frequency
1.35GHz

Instances
2.35B gates, 103M flops

Area
25.6mm x 16.4mm, 421mm2

Package
50mm x 40mm

Voltage
0.85V

TDP
90W

Host Connection
8x PCIe Gen5 (32 GB/s)

GEMM TOPS
708 TFLOPS/s (INT8) (sparsity)
354 TFLOPS/s (INT8)
354 TFLOPS/s (FP16/BF16) (sparsity)
177 TFLOPS/s (FP16/BF16)

SIMD TOPS
Vector core:
11.06 TFLOPS/s (INT8),
5.53 TFLOPS/s (FP16/BF16),
2.76 TFLOPS/s (FP32)
SIMD:
5.53 TFLOPS/s (INT8/FP16/BF16),
2.76 TFLOPS/s (FP32)

Memory Capacity
Local memory: 384 KB per PE
On-chip memory: 256 MB
Off-chip LPDDR5: 128 GB

Memory Bandwidth
Local memory: 1 TB/s per PE
On-chip memory: 2.7 TB/s
Off-chip LPDDR5: 204.8 GB/s

Revision #2
Created 2025-01-11 09:46:28 UTC by Colin
Updated 2026-04-29 07:33:40 UTC by Colin