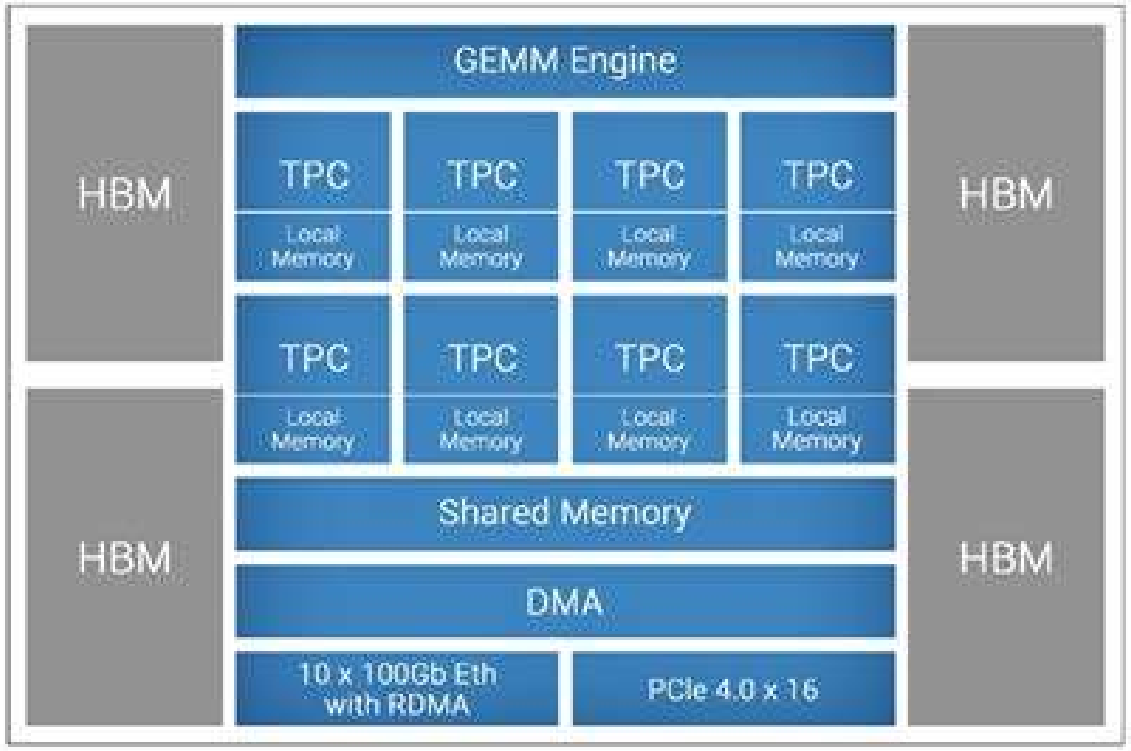
Gaudi Processor Architecture

- Heterogenous compute architecture
  - TPC, GEMM & DMA using a shared SRAM
- VLIW SIMD TPC 2.0 Core (C-programmable)
- GEMM operations engine
- Tensor addressing
- Robust to any address stride
- Latency hiding capabilities
- PCIe Gen4.0 x16
- 4 HBMs: 2GT/s, 32 GB capacity, BW 1 TB/sec
- 10 ports of 100Gb Ethernet, or 20x50 GbE
  - With integrated RDMA over Converged Ethernet (RoCE v2)
- Dedicated HW and TPC ISA for special functions acceleration (e.g. Sigmoid, GeLU, Tanh)
- Mixed-precision data types: FP32, BF16, INT32, INT16, INT8, UINT32, UINT16 and UINT8



TSMC – 16nm