AI芯片软硬件相关技术术语

## AI硬件

### 指令集架构

* 复杂指令集 (CISC, Complex Instruction Set Computing)
* 精简指令集 (RISC, Reduced Instruction Set Computing)
* x86
* ARM
* RISC-V
* CPU (中央处理器)
* GPU (图形处理器)
* NPU (神经网络处理器)
* TPU (张量处理器)
* ASIC (专用集成电路)
* FPGA (现场可编程门阵列)
* ISA (指令集架构)
* Microarchitecture (微架构)
* PPA (Performance, Power, Area)
* Load/Store
* 标量 (Scalar)
* 超标量 (Superscalar)
* 向量 (Vector)
* 张量 (Tensor)
* 矩阵 (Matrix)
* MAC (乘加单元)
* 图灵完备 (Turing Complete)

### 并行计算技术

* VLIW (超长指令字)
* SIMD (单指令多数据)
* SIMT (单指令多线程)

### AI芯片子系统

* Systolic Array (脉动阵列)
* GEMM引擎 (General Matrix Multiply Engine)
* Tensor Core (张量核心)
* Matrix Engine (矩阵引擎)
* PE Array (Processing Element Array, 处理单元阵列)
* MAC阵列 (Multiply-Accumulate Array, 乘加阵列)
* 脉动阵列优化 (Systolic Array Optimization)
* 激活函数单元 (Activation Function Unit)
* Softmax单元 (Softmax Unit)
* 归一化单元 (Normalization Unit)
* 池化单元 (Pooling Unit)
* 卷积引擎 (Convolution Engine)
* FFT单元 (Fast Fourier Transform Unit)
* 随机数生成器 (Random Number Generator)

### AI芯片

* NPU (神经网络处理器)
* TPU (张量处理器)
* ASIC (专用集成电路)
* FPGA (现场可编程门阵列)

#### 芯片制造与封装

* Chiplet (小芯片, Chiplet)
* 3D NAND
* 3D堆叠 (3D Stacking)
* WoW (Wafer-on-Wafer)
* CoW (Chip-on-Wafer)
* 硅通孔互连 (Through Silicon Via Interconnect)
* 微凸点技术 (Microbump Technology)
* Hybrid Bonding (混合键合)
* 光电集成 (Opto-electronic Integration)
* COWOS (Chip-on-Wafer-on-Substrate)
* Wafer (晶圆)
* Die (芯片裸片)
* Packaging 封装技术
* 先进制程 (Advanced Node)
* FinFET
* GAAFET (Gate-All-Around FET)
* 流片 (Tape-out)
* Yield (良率)
* 异构集成 (Heterogeneous Integration)
* 2.5D/3D 封装
* Fan-out/Fan-in

#### 芯片设计流程与测试

* RTL设计 (Register-Transfer Level Design)
* 逻辑综合 (Logic Synthesis)
* 物理设计 (Physical Design)
* DFT (Design for Test, 可测性设计)
* 静态时序分析 (Static Timing Analysis, STA)
* 功耗分析 (Power Analysis)
* Wafer (晶圆)
* Die (裸片)
* Packaging封装技术 (Packaging Technology)
* Chiplet (小芯片)
* COWOS (Chip-on-Wafer-on-Substrate)
* 先进封装 (Advanced Packaging)
* 系统级封装 (SiP, System in Package)
* DFT (Design For Testability)
* EDA (Electronic Design Automation) 工具
* RTL (Register-Transfer Level)
* Synthesis (综合)
* PnR (Place and Route)
* STA (Static Timing Analysis)
* ATPG (Automatic Test Pattern Generation)
* MBIST/LBIST (Memory/Logic Built-In Self-Test)
* 仿真 (Simulation)
* 硬件仿真器 (Emulator)
* 调试 (Debugging)
* 性能分析 (Profiling)
* 低功耗设计 (Low-power Design)
* 异构多核 (Heterogeneous Multi-core)
* 大小核架构 (Big.LITTLE Architecture)

### AI板级硬件

* PCIe计算卡
* OAM 模组 (OCP Accelerator Module)

### 互联标准与技术

* PCIe (Peripheral Component Interconnect Express)
* CXL (Compute Express Link)
* NVLink
* Infinity Fabric
* UALink
* UEC
* Scale-up Ethernet (SUE)

### AI系统级硬件

* Switch Tray
* Compute Tray
* Cable Tray
* QSFP-DD (Quad Small Form-factor Pluggable Double Density)
* CPO (Co-packaged Optics)
* 正交背板 (Orthogonal Backplane)
* SuperPod 超节点
* 高速 SerDes (Serializer/Deserializer)
* Controller/PHY
* NIC (Network Interface Card)

#### 网络拓扑结构

* Scale-up (纵向扩展)
* Scale-out (横向扩展)
* Dragonfly拓扑 (Dragonfly Topology)
* 胖树 (Fat Tree)
* 环面 Torus 拓扑 (Torus Topology)
* Mesh拓扑 (Mesh Topology)
* Ring拓扑 (Ring Topology)
* Tree拓扑 (Tree Topology)

#### 通信标准与技术

* OSI七层参考模型
* 以太网 (Ethernet)
* TCP/UDP
* InfiniBand
* RoCE (RDMA over Converged Ethernet)
* RDMA (Remote Direct Memory Access)
* QSFP-DD (Quad Small Form-factor Pluggable Double Density)
* OSFP (Octal Small Form Factor Pluggable)
* CPO (Co-Packaged Optics)
* 高速SerDes (Serializer/Deserializer)

### 存储技术

* Register
* Cache
* SRAM
* DDR/GDDR SDRAM
* HBM
* NVMe协议
* SSD

### 安全性/可靠性/性能评估技术和指标

* FLOPS (Floating-point Operations Per Second)
* TOPS (Tera Operations Per Second)
* Throughput (吞吐量)
* Latency (延迟)
* TTFT (Time To First Token)
* TPOT (Time Per Output Token)
* TPS (Tokens Per Second)
* MAC利用率 (MAC Utilization)
* 带宽利用率 (Bandwidth Utilization)
* PerfML
* PUE (Power Usage Effectiveness)
* TCO (Total Cost of Ownership)
* 能耗比 (Energy Efficiency Ratio)
* 算力密度 (Computing Power Density)

## AI软件

### 数据表示与精度

* FP32 (单精度浮点数)
* FP16 (半精度浮点数)
* BF16 (脑浮点数)
* FP8 (8位浮点数)
* INT8 (8位整数)
* INT4 (4位整数)
* FP4 (4位浮点数)
* NVFP4/MXFP4 (厂商特定浮点格式)
* E4M3 (指数4位尾数3位的FP8变体)
* E5M2 (指数5位尾数2位的FP8变体)
* 定点数 (Fixed Point Number)
* 浮点数 (Floating Point Number)
* 混合精度训练 (Mixed Precision Training)

### 模型并行计算技术

* 数据并行 (Data Parallelism, DP)
* 流水并行 (Pipeline Parallelism, PP)
* 张量并行 (Tensor Parallelism, TP)
* 专家并行 (Expert Parallelism, EP)
* 模型并行 (Model Parallelism)
* 序列并行 (Sequence Parallelism)

### 集合通信操作

* AllReduce
* AllGather
* ReduceScatter
* Broadcast
* Reduce
* All-to-All
* Send/Recv
* Scatter
* Gather
* Neighbor exchange

### AI模型基础

#### 基础模型

* CNN卷积神经网络 (Convolutional Neural Network)
* RNN循环神经网络 (Recurrent Neural Network)
* 大语言模型 (Large Language Model, LLM)

#### 大语言模型

* BERT (Bidirectional Encoder Representations from Transformers)
* GPT (Generative Pre-trained Transformer)
* Dense 模型 (稠密模型)
* MoE 模型 (Mixture-of-Experts)

#### 视觉模型

* ResNet (Residual Neural Network)：图像分类
* MobileNet：图像分类
* EfficientNet：图像分类
* YOLO (You Only Look Once)：目标检测
* UNet：语义分割
* DeepLab：语义分割
* PSPNet (Pyramid Scene Parsing Network)：语义分割

#### 生成模型

* 生成对抗网络 (GAN, Generative Adversarial Network)
* 变分自编码器 (VAE, Variational Autoencoder)

#### CNN卷积神经网络 (Convolutional Neural Network)

* 批归一化 (Batch Normalization)
* 激活函数 (Activation Function)
* 反向传播 (Backpropagation)
* 梯度下降 (Gradient Descent)
* Adam优化器
* 学习率调度 (Learning Rate Scheduling)

#### Transformer网络

* Transformer算法 (Transformer Algorithm)
* 注意力机制 (Attention Mechanism)
* 多头注意力 (Multi-Head Attention)
* 位置编码 (Positional Encoding)
* 残差连接 (Residual Connection)
* 层归一化 (Layer Normalization)
* Dropout

### 模型处理技术

* 前向传播
* 反向传播
* 权重更新
* 梯度下降
* 模型训练 (Model Training)
* 模型推理 (Model Inference)
* 推理Prefill阶段
* 推理Decode阶段
* PD分离推理技术
* 微调 (Fine-tuning)
* 监督微调 (SFT, Supervised Fine-Tuning)
* 强化学习 (RLHF, Reinforcement Learning with Human Feedback)
* 后训练 (Post-training)
* 蒸馏 (Distillation)
* 量化 (Quantization)
* 剪枝 (Pruning)
* 模型压缩 (Model Compression)
* 知识蒸馏 (Knowledge Distillation)
* 多模态模型 (Multimodal Model)
* 稀疏化 (Sparsity)
* 模型上下文 (Context Length)
* 激活函数 (Activation Function)
* 损失函数 (Loss Function)
* 注意力机制 (Attention Mechanism)
* 自监督学习 (Self-Supervised Learning)

### AI编译器

* LLVM
* LLDB
* GCC
* GDB
* MLIR
* TVM (Tensor Virtual Machine)
* OpenXLA
* Runtime运行时 (Runtime Environment)
* JIT编译 (Just-In-Time Compilation)
* AOT编译 (Ahead-Of-Time Compilation)

### AI软件栈

* PyTorch
* TensorFlow
* ONNX
* Triton
* Torch-XLA
* JAX
* PaddlePaddle
* vLLM
* TensorRT
* ONNX Runtime
* OpenVINO
* TVM