**第一章：**

介绍大背景、一些常识

RISC架构特性：好用！

硬件的Dennard Moore定律 终结

带宽的增长速度比延迟的降低速度更快

晶体管尺寸缩小 密度增加 管性增加，连线延迟不会降低

系统设计者关心：处理器最高功耗 正常功耗 能效比

多个层级的并行性

**第二章：Cache**

介绍：Cache造成功耗变大

优化Cache的方法：降低缺失率、降低缺失惩罚

SRAM DRAM

Flash memory

保护和隐私

TLB 用户态和内核态 读写执行权限

虚拟机：完整的软件栈，共享一套硬件

乱序处理器：取值操作解耦合

猜测执行

多核和IO都会造成一致性问题

**第三章：ILP**

介绍：数据冲突和控制冲突，指令级并行度ILP：硬件动态软件编译静态

数据冲突： 数据相关的指令不能交叠运行 限制了ILP

名字相关使用了相同寄存器或内存地址

RAW WAW WAR

控制相关：if while之前之后顺序问题

编译优化对ILP的静态发射和调度

循环展开

分支预测：考察本地、全局历史

动态调度：Tomasulo算法 多发射

硬件的猜测执行

多发射处理器：静态超标量 VLIW 动态超标量

其他方法提高ILP：超线程 同步多线程SMT

**第四章：GPU**

介绍：DLP SIMD：向量体系结构、多媒体指令扩展和GPU

RV64V在RISC-V向量扩展RVV

循环中没有循环体间相关，是可向量化的 窄向量自动向宽向量转型

向量操作执行时间：向量长度 结构相关 数据相关

SIMD 算术密度 Roolfine评估模型

CUDA编程

GPU 由一个或多个多线程SIMD处理器组成

PTX指令集

GPU对比向量处理器对分支指令有更多硬件支持

每个SIMD通道有私有内存

SIMD处理器支持多线程

挖掘循环体的并行性（编译时）

降低频率和电压、增加运算资源能降低GPU功耗

GPU使用堆叠式内存HBM提高访存带宽

GPU并不总是优于CPU

SIMD已加入gather-scatter访存

**第五章：多处理器 一致性**

介绍：TLP线程级并行性 多处理器架构 多个线程同一个任务 多个不相关进程同时工作

共享内存地址

第一种多处理器系统 SMP对称多处理器

第二种 分布式共享内存多处理器DSM

缓存一致性与内存一致性是互补的

两类一致性协议：基于目录 基于侦听

确保一致性方式：写无效协议 写更新协议

锁的同步机制

内存一致性：放松一致性

**第六章：WSC**

介绍：WSC机群

不同的带宽和延迟

不会同时最高负载 削减40%的电能

开销 导致云计算新趋势

微软、亚马逊和谷歌在各个大洲都有大量WSC

**第七章：DSA**

介绍：摩尔定律 登纳德缩放定律终结 DSA领域特定加速器

DNN：MLP CNN RNN LSTM

TPU：Pcle通道

微软将FPGA连接至Pcle通道

Pixel图象处理和计算视觉加速器：将ISP的核心流水线转为有向无环图