

第一次讲座作业

无04 2019012137 张鸿琳

1. 和Nvidia, AMD等公司相比, 沐曦等国产GPU公司的优势和不足有哪些, 相比于CPU, memory等领域, 国产GPU产业又有哪些特点?

- 和Nvidia, AMD等公司相比, 国产GPU公司的优势: 国内市场特点是人工智能市场应用广阔, 国家高度重视芯片产业发展, 而当前国外又封锁芯片行业, 使得国内需求巨大的应用商转向和国内GPU公司加强合作, 据东吴证券测算, GPU领域国产替代的市场空间超过50亿美元。同时在各国家政策支持下, GPU行业也迎来了大量国内融资, 保证其资金充足。国内GPU公司也积极开拓GPU的各种发展方向, 不局限于图像处理, 转而投向AI+Science等领域, 从而避免了国外老牌企业的正面交锋, 同时拓展GPU的积分指标, 而不是局限于单点指标
- 和Nvidia, AMD等公司相比, 国产GPU公司的不足: 起步较晚, 英伟达、AMD等老牌企业已经占据国际市场的大部分份额, 和很多应用商相互绑定, 国内GPU难以切入。在GPU领域, 英伟达耕耘多年的CUDA开发者生态, 也是英伟达生态的核心, 很多深度学习的加速库都是基于CUDA, 并且是和英伟达的硬件深度绑定的, 这是创业公司难以撼动的壁垒
- 相比于CPU和memory领域, 国产GPU产业的特点: 国内的CPU和memory领域的市场已经较为饱和, 且其各自的架构也较为完善, 接近其发展极限, 长期被国外企业垄断, 难有突破点, 但是GPU行业还有不少新的发展机遇, 如深度学习、科学计算等方向, GPU能做到的性能相较于CPU更好, 除了这样新的应用推动, 更有如先进封装和芯粒这样的新工艺驱动, 通过解决稀疏化、GPGPU领域定制架构平台、芯粒间通信的互联接口和协议、多GPU芯粒间缓存/通信原语设计、Warp级宏指令、内存模型与片上缓存设计等问题, 进一步完善GPGPU的内存模型, 还可以进一步提升GPU的在处理各种问题上的性能, 由此可知国内GPU产业还有很大的发展空间, 同时国内应用商需求量也十分巨大, 国内GPU企业积极谋求创新仍然有和国外老牌企业正面竞争的机会

2. 为什么要面向AI应用设计领域专用架构, 领域专用架构的好处是什么? 针对AI算法, 你认为华为昇腾AI处理器的硬件设计最主要的创新点是什么? 面向AI应用, 除了处理器硬件设计外, 华为在软件平台方面做了哪些努力与创新, 软件平台和处理器硬件如何实现相互配合?

- 为什么面向AI应用设计领域专用架构: AI计算属于全新计算范式, 深度学习的核心运算逻辑为张量运算, 而传统CPU架构的张量运算比较低效, 同时摩尔定律失效, Dennard Scaling定律失效, 这都导致传统CPU架构无法支撑现有AI运算
- 领域专用架构(DSA)的好处: DSA能够在特定域利用更有效的并行形式, 更有效地利用内存层次结构, 足够的情况下可以使用较低的精度, 从而使其能实现更高的能效比。同时还有专用领域语言(DSL)专为DSA设计, DSL可以使矢量、密集矩阵和稀疏矩阵操作量式, 从而使DSL编译器能够有效地将操作映射到处理, 实现最佳的开发效率和运行开销
- 华为昇腾AI处理器的硬件设计最主要的创新点: 昇腾NPU采用达芬奇架构, 相比于需要兼顾图像渲染和AI计算的GPU, 昇腾NPU专为AI计算设计, Cube占比高, 面效比高, 算力高, 可以一个时钟周期中完成 $16 \times 16 \times 16$ 个FP16 MAC运算, 同时兼顾可扩展, 支持多种模型, 场景覆盖全面
- 华为在软件平台方面的努力与创新: 华为自研了编译器、编程语言和加速库, 构筑了计算产业的基础软件平台。编译器方面突破了异构编程技术, 构建了异构加速软件栈, 多芯片异构算力最佳; 而加速库方面也通过算法技术的持续投入和创新, 构筑了绝对领先的产品竞争力
- 软件平台和处理器硬件的相互配合: 华为利用CANN异构计算架构, 使得软件平台和处理器硬件能够协同释放算力

