宋柳斌

求职意向：架构师 / model leader

生　日：1994.7
电　话：18751963598
住　址：上海浦东
邮　箱：1037209590@qq.com

教育背景

● 南京理工大学	电子科学与光电技术学院	电子科学与技术
● 南京大学（免试保送）	电子科学与工程学院	集成电路工程

工作经历

● 大模型芯片公司	资深经理（performance leader）	2024/9-----至今
● 蔚来汽车	主任工程师（model leader）	2022/4----2024/9
● 华为海思	高级工程师 B（model）	2021/4----2022/4
● 阿里巴巴	开发工程师(验证)	2019/7----2021/4

项目经历

◆大模型芯片公司 **2024.9-至今**

项目简介：大模型推理芯片 G100

职级：架构师 / performance leader

下属人数：直线 5，虚线 14

职责描述：整体 G100 软硬件系统架构性能，G100 pim die 计算 core 架构师，负责 G100 前端系统交付

◆ 硬件架构：

1) G100 项目计算 core 架构师，负责 pim die 内部计算 core 的高性能架构设计
2) CCL 数据传输高性能架构设计，分析数据上行和下行性能瓶颈，端到端 H2D/D2H 整体带宽提升 4 倍，die 间/chip 间 CCL 性能提升 6 倍
3) 负责硬件 MMA/TMA 等 macro 指令 issue 高性能设计，指令 issue 性能提升 90%
4) 优化 tensor core 地址访存设计，端到端有效算力提升至 80%
5) 优化 TMA 访存子系统设计，端到端有效带宽提升至 95%
6) 优化 G100 kernel 间同步设计，由统一单 kernel 串行同步改为三层 task graph 同步，kernel 同步性能提升 70%
7) G100 大模型 token 性能整体提升 5 倍

◆ 软件优化：

1） llama2 7B/70B, llama3 8B,qiwen2.5 7B 单卡、多卡性能评估，分析 TP/PP/SP/DP 等不同部署模式下在 G100 芯片下的性能表现，指导上层软件部署大模型方案
2） 分析手写算子和 triton 算子，结合 G100 硬件特性针对性优化算子逻辑实现和访存特性，提升算子性能 40%

◆ 系统交付：

1） 负责 G100 整 chip 功能、性能系统联调
2） 搭建 replay 软硬件系统联调平台，在 FPGA 平台完整跑通 llama2 7B FP16, qiwen3 8B FP8 decoder
3） 负责 TO 前的功能、性能系统压力测试
4） 搭建顶层性能评估平台 perf model

◆蔚来汽车 **2022.4-2024.9**

项目简介：蔚来汽车 NX9031

职级：架构师 / model leader

下属人数：8

职责描述：架构设计，model 团队交付，系统验证

◆ 架构设计：

1) NX9031 NPU SIMT 指令集的维护与演进，搭建 AI 芯片 fast model 架构演进平台

2) 梳理蔚来汽车 BEV/LIDAR/STAMP 智驾模型，llama2 7b/70b, qiwen2.5 7b 等大语言模型，分析模型所需算力、带宽需求，以及在不同带宽算力配比下各模型的性能表现，指导芯片设计规格

3) 负责 NX9031 NPU 任务调度控制器 TD 架构，设计 SWSQ 到 HWSQ 的优先级选择机制，task graph/grid/block/dispatch，queue 间/task 间/grid/block 间的同步和依赖

◆ **model 开发交付：**

1） 负责 NX9031 NPU 芯片 function model 的搭建与交付，该 model 基于 NPU 指令集实现，承接 LLVM 团队、算子团队、图编译团队的初期开发，验证 NPU SIMT 指令集的功能正确性和完备性

2） 负责 NX9031 NPU 芯片 cycle accurate model 的开发和交付，该 model 基于指令集和硬件微架构实现，为软件图编译、算子团队提供 cycle 级的功能仿真和性能仿真，perf 准确度与硬件 align 达到 97%；该模型还作为 DV 团队的 reference model，负责硬件的模块级、子系统级、chip 级的 CO_SIM 验证

3） 负责整个模型组对上层软件和 DV 团队的交付与版本发布

◆ **系统验证：**

1） 规划 NX9031 NPU 芯片的系统验证用例，从系统层面测试 NPU 各模块、各子系统、整芯片的功能和性能是否满足架构设计需求

◆华为海思 **2021.4-2022.4**

项目简介：笛卡尔 GPU V300

职级： model 高级工程师

职责描述：V300 GPU 调度子系统设计

1) 负责 V300 GPU command processor 架构设计

2) 负责硬件 context 任务的调度 schedule 设计，binning/rendering/compute/bvhg/rttc 五种任务的多核调度分发

3) 负责 V300 state base descriptor per_frame/per_draw/per_bin 调度的预研

◆平头哥 **2019.07-2021.4**

项目简介：平头哥 PPU1.0

职级：开发工程师

职责描述：warp scheduler 模块验证

1) 搭建 warp scheduler block 验证平台和 reference model 的搭建

2) 负责 PPU1.0 指令集 single isa 的 function 验证

3) PPU1.0 指令集 RAW/WAR/WAW 的 reference model 实现

4) 开发验证 regression 返标工具 VBA