

【职业发展路径图 + 项目提升建议表（融合AI分析与职业趋势）】

姓名：罗腾 目标方向：软硬协同工程师 / AI模型推理加速 / CUDA内核优化 / 芯片架构设计 阶段：研一，具备数字IC与GPU并行计算基础，寻求综合性发展路径

一、选择建议：优先深耕“软硬结合”方向

1. 为何不建议单纯深入CPU或纯软件？

* 纯CPU设计：国内高端CPU市场被ARM/x86垄断，自主架构生态突破需长期积累，岗位集中在少数头部企业，职业路径相对狭窄。
* 纯软件优化：大模型推理加速虽热门，但若缺乏硬件底层理解，容易陷入“调参工程师”瓶颈，技术壁垒不足。

1. 软硬结合的核心价值

* 契合行业需求：AI芯片公司和互联网大厂亟需既懂硬件架构、又能优化模型部署的跨领域人才，Chisel/CUDA双重经验正匹配此类岗位。
* 薪资与成长性：软硬结合岗位平均薪资比纯硬件高20-30%，技术护城河更深，职业生命周期更长。

二、职业发展路径图（20242025） ▶ 技术强化阶段

* 深化CUDA kernel优化理解，学习shared memory、coalesced memory等性能关键点
* 学习TVM、TensorRT、ONNX Runtime、MLIR、Triton等工具链与优化框架
* 参与AI推理系统开源项目，或自制一个小型深度模型推理框架（如PointNet）
* 阅读AI编译器相关论文，建立软硬协同系统思维

研二（2025） ▶ 项目沉淀+实习准备阶段

* 制作GitHub工程库（硬：RISC-V SoC项目；软：CUDA模型推理项目）
* 拓展项目方向：如RISC-V向量指令加速LLM推理，结合Gem5模拟与性能分析
* 投递AI芯片公司或云平台异构计算团队实习岗

研三（2026~2027） ▶ 定向专精阶段

* 主线方向：AI推理平台开发 / 编译器优化
* 支线方向：芯片架构 / 新型计算范式（光子计算、存算一体）
* 开始技术纵深研究或准备出国/Top岗

三、目标岗位与企业推荐

1. AI芯片公司

* 华为海思：NPU架构师、AI芯片验证工程师
* 阿里平头哥：RISC-V生态工具链开发工程师
* 寒武纪/燧原/壁仞：GPU架构、模型部署、编译器优化

1. 云计算与互联网平台

* 阿里云、昇腾、腾讯云：异构计算工程师
* 字节AML、百度飞桨：大模型推理系统开发、深度学习框架优化

1. 高性能计算与新兴领域

* NVIDIA中国：TensorRT/CUDA内核开发
* 光子算数、曦智科技：光子计算芯片系统映射

1. RISC-V生态与开源协同

* 芯来科技、赛昉科技：AI扩展指令开发、工具链构建
* 中科驭数：DPU加速引擎开发

四、技能提升路径

* 工具链：掌握MLIR、Triton、AutoScheduler等编译优化框架
* 模型优化：学习TensorRT量化、FlashAttention、KV Cache优化
* 硬件感知算法设计：用Gem5分析访存瓶颈，设计带硬件感知能力的算子

五、项目包装建议

| **项目方向** | **技术关键词** | **提升方式** | **产出** |
| --- | --- | --- | --- |
| CUDA模型推理 | kernel优化/Warp并行 | 增加FP16支持+kernel融合 | GitHub库+博客 |
| AI加速器设计 | RISC-V指令/架构分析 | 向LLM指令集扩展方向演进 | Paper/开源 |
| TVM/MLIR优化 | AutoTVM/IR Pass设计 | 算子融合+图优化案例构建 | 简历亮点 |

六、长期发展路径

* 技术纵深：AI芯片架构师、大模型推理系统负责人
* 交叉探索：光子计算系统映射、存算一体芯片设计
* 学术产业结合：RISC-V AI生态建设、Chiplet系统协同设计

七、岗位筛选建议

* 优先投递软硬结合岗位：工具链开发、推理加速、模型部署
* 避免路径收窄：慎重选择仅限底层CPU设计或纯算法岗

八、兴趣验证策略

* 尝试构建TVM+MLIR端到端优化案例（如ResNet部署）
* 参与RISC-V AI/ML工作组，设计并提交一个自定义扩展指令提案

结语：你具备成为“AI时代的芯片架构师”或“懂硬件的AI系统专家”的潜力，建议从AI推理系统与编译优化切入，构建技术闭环，同时保留硬件能力形成差异化壁垒。

【职业发展路径图 + 项目提升建议表（融合AI分析与职业趋势）】

姓名：罗腾 目标方向：软硬协同工程师 / AI模型推理加速 / CUDA内核优化 / 芯片架构设计 阶段：研一，具备数字IC与GPU并行计算基础，寻求综合性发展路径

（前文内容保留，略）

九、岗位搜索关键词清单表

| **岗位方向** | **推荐关键词（中文+英文）** | **适配平台** | **典型公司推荐** |
| --- | --- | --- | --- |
| AI推理加速 | AI推理工程师 / Inference Engineer / 模型部署优化 | Boss直聘、小红书 | 字节跳动、百度飞桨、阿里云 |
| CUDA/GPU优化 | CUDA开发工程师 / CUDA Kernel / GPU加速工程师 | Boss直聘、小红书 | NVIDIA中国、壁仞科技、燧原科技 |
| 编译器/工具链 | AI编译器工程师 / TVM工程师 / MLIR开发工程师 | Boss直聘、小红书 | 平头哥、地平线、华为CANN团队 |
| 软硬协同岗位 | 软硬件协同工程师 / 异构计算工程师 / AI芯片系统工程师 | Boss直聘、小红书 | 昇腾团队、阿里云、寒武纪 |
| 芯片设计扩展 | RISC-V工程师 / SoC设计 / 指令集扩展开发 | Boss直聘、小红书 | 赛昉科技、芯来科技、中科驭数 |
| 系统软件/平台 | 深度学习系统开发 / AI框架开发 / AI系统工程师 | Boss直聘、小红书 | 腾讯优图、华为MindSpore团队 |
| 实习/校招岗位 | CUDA实习 / AI推理实习 / 编译器实习 / 工具链实习 | Boss直聘、小红书、官网 | 阿里达摩院、字节校招、百度校招 |

十、岗位搜索组合关键词（快速复制使用）

1. 模型加速 TVM
2. CUDA内核优化 实习
3. AI芯片工具链 编译器
4. RISC-V 指令集 扩展
5. 大模型推理系统 工程师
6. MLIR 中间表示优化
7. PyTorch模型部署优化
8. 异构计算平台 工程师
9. AI框架优化 CUDA Kernel
10. AI芯片部署系统工程师

提示：搜索时搭配关键词“优化”“部署”“工具链”“系统开发”等辅助词可获得更精准匹配结果。

结语： 结合你的技术背景与行业趋势，建议在岗位搜索过程中主动定位“软硬结合+高成长性+工具链导向”的工程岗位，并以开源项目和系统工程能力作为求职亮点，在Boss直聘/小红书等平台筛选最契合的岗位信息。