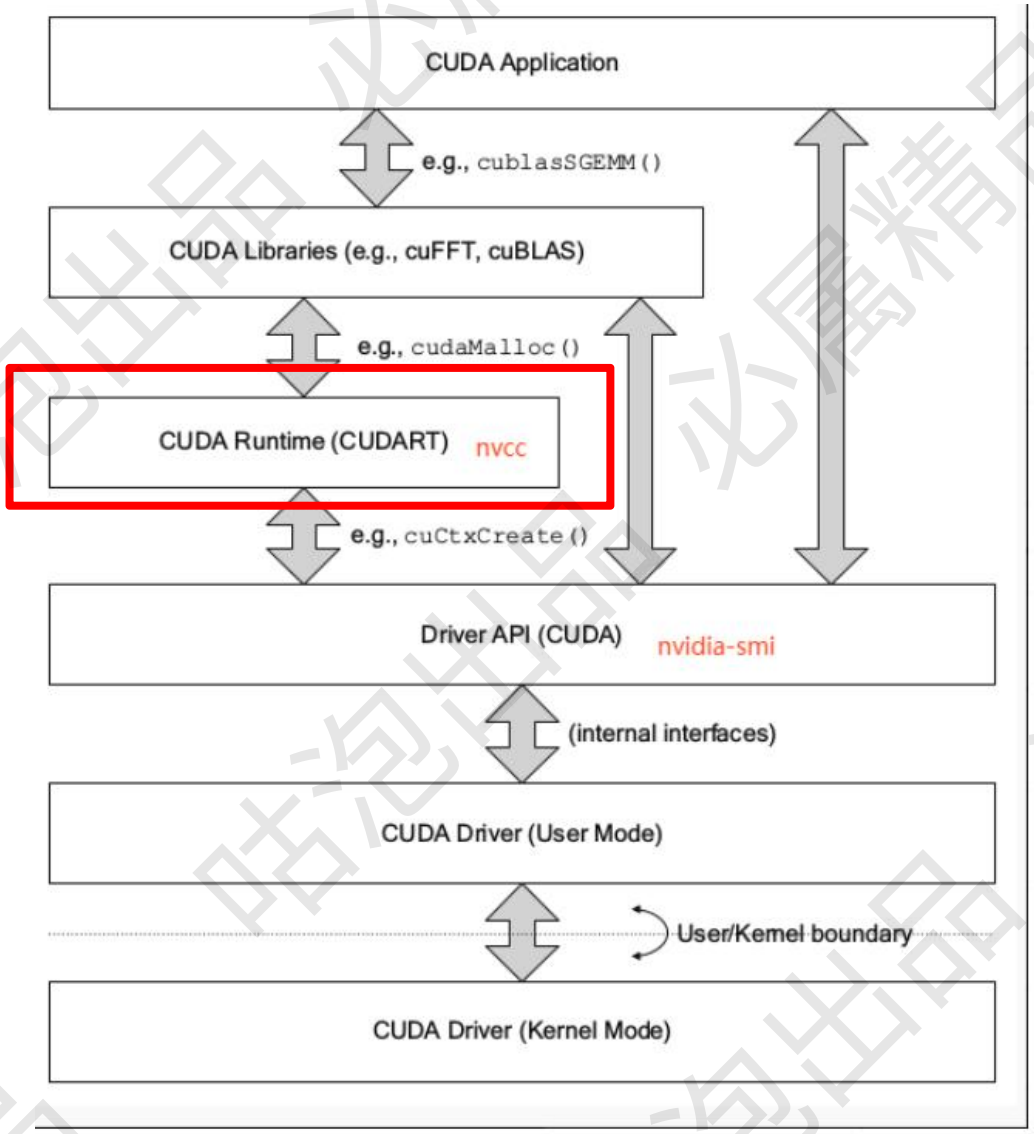
精简CUDA教程-RuntimeAPI概述

精简CUDA - RuntimeAPI



精简CUDA - RuntimeAPI

1. 对于runtimeAPI，与driver最大区别是**懒加载**
2. 即，第一个runtime API调用时，会进行**cuInit初始化**，避免驱动api的初始化窘境
3. 即，第一个需要context的API调用时，会进行context关联并创建context和设置当前context，**调用cuDevicePrimaryCtxRetain**实现
4. 绝大部分api需要context，例如查询当前显卡名称、参数、内存分配、释放等

精简CUDA - RuntimeAPI

1. CUDA Runtime是封装了CUDA Driver的高级别更友好的API
2. 使用cuDevicePrimaryCtxRetain为每个设备设置context，不再手工管理context，并且不提供直接管理context的API（可Driver API管理，通常不需要）
3. 可以更友好的执行核函数，.cpp可以与.cu文件无缝对接
4. 对应cuda_runtime.h和libcudart.so
5. runtime api随cuda toolkit发布
6. 主要知识点是**核函数的使用、线程束布局、内存模型、流的使用**
7. 主要实现**归约求和、仿射变换、矩阵乘法、模型后处理**，就可以解决绝大部分问题

精简CUDA - RuntimeAPI

对应于系列名称：cuda-runtime-api

获取代码：trtpy get-series cuda-runtime-api

查询系列清单：trtpy series-detail cuda-runtime-api

精简CUDA - RuntimeAPI

```
C:\Users\Administrator\cuda-driver-api>trtpy series-detail cuda-runtime-api
Download cuda-runtime-api.series.json: 100%|                                  | 2 KB/2 KB 00:00<00:00
List templ:
chapter: 1.1, caption: hello-runtime, description: CUDA运行时API开始，以及与CUDA驱动API的context关系解释
chapter: 1.2, caption: memory, description: 进行GPU的内存分配
chapter: 1.3, caption: stream, description: cudaStream流管理，函数的异步控制
chapter: 1.4, caption: kernel-function, description: 第一个核函数
chapter: 1.5, caption: thread-layout, description: 线程的布局设计
chapter: 1.6, caption: vector-add, description: 使用cuda核函数实现向量加法
chapter: 1.7, caption: shared-memory, description: 共享内存的操作
chapter: 1.8, caption: reduce-sum, description: 规约求和的实现，利用共享内存，高性能
chapter: 1.9, caption: atomic, description: 原子操作，实现动态数组的操作
chapter: 1.10, caption: warpaffine, description: 仿射变换双线性插值的实现，yolov5的预处理
chapter: 1.11, caption: cublas-gemm, description: 通用矩阵乘法的cuda核函数实现，以及cublasSgemm的调用
chapter: 1.12, caption: yolov5-postprocess, description: 使用cuda核函数实现yolov5的后处理案例
chapter: 1.13, caption: thrust, description: 简单用一下cuda的stl库 thrust
chapter: 1.14, caption: error, description: 关于cuda发生错误的一些理解
```

谢谢!