深蓝学院
shenlanxueyuan.com

CUDA & TensorRT
保姆级安装教程

课程目标

- Driver、CUDA、CUDNN和TensorRT的关系
- 如何选择版本
- 手动安装实操
- Docker 安装

Driver、CUDA、CUDNN和TensorRT的关系

概念	解释	组成	不同版本是否可以并存
CUDA/NVIDIA Driver	软件与硬件之间的桥梁。没有驱动，就不能识别GPU硬件，不能调用其计算资源	驱动程序	不可以
		工具：NVIDIA-Smi 等	
CUDA Toolkit	cuda 安装包，有最低驱动版本要求	编译器：nvcc	可以，修改环境变量即可
		工具：nvvp，cuda-memcheck，cuda-gdb等	
		library：cudart，cublas等	
		samples：演示如何使用各种CUDA和library API的代码示例	
		cuda driver：与当前cuda版本相兼容的driver	
CUDNN	CUDA 深度神经网络库，对cuda版本有硬要求	library	
TensorRT	对cuda和cudnn版本有硬要求	library	
		工具：trtexec	
		samples：示例代码	
		python安装包：onnx_graphy、trt python版本的安装包	

如何选择版本

两个准则:

1、自2016年以来，**每新出一代显卡，为了支持显卡的新特性，CUDA会对应更新一个大版本**。根据显卡架构选择当代或者下一代的CUDA。比如T4选择安装CUDA10.2 或者 CUDA11.x。不建议跨太大版本，有可能会存在负优化情况。

2、TensorRT近几年更新比较频繁，只建议选择**稳定版本**，推荐三个版本：TensorRT6.0GA，TensorRT7.2.3和TensorRT8.5GA。TRT对CUDA和CUDNN版本有强要求。

如何选择版本

两个准则:

1、自2016年以来，**每新出一代显卡，为了支持显卡的新特性，CUDA会对应更新一个大版本**。根据显卡架构选择当代或者下一代的CUDA。比如T4选择安装CUDA10.2 或者 CUDA11.x。不建议跨太大版本，有可能会存在负优化情况。

显卡架构	CUDA版本	稳定版本
2016年 Pascal（P4， P40	CUDA8	CUDA8.0 GA2
2017年 Volta（V100	CUDA9	cuda9
2018年 Turing（T4	CUDA10	CUDA10.2
2020年 Ampere（A100	CUDA11	不知道
2022年 Hopper（H100	CUDA12	不知道

如何选择版本

2008年 Tesla	白皮书：https://images.nvidia.com/content/pdf/tesla/whitepaper/pascal-architecture-whitepaper.pdf
2010年 Fermi	Fermi是第一个完整的GPU计算架构。 1.Fermi 首款可支持与共享存储结合纯cache层次的GPU架构，支持ECC的GPU架构。 2.512个 accelerator cores即所谓CUDA cores(包含ALU和FPU)，16个SM，每个SM包含32个CUDA core。
2012年 Kepler	Kepler相较于Fermi更快，效率更高，性能更好。 1.15个SM，192个单精度CUDAcores，64个双精度单元 2.Kepler图形架构在极大提升游戏性能的同时，又在很大程度上降低了能耗。Kepler基于28纳米制造工艺
2014年 Maxwell	紧随Kepler之后，Maxwel是NVIDIA的第四代GPU架构。这一架构是下一代游戏体验的引擎，可解决视觉计算领域中最复杂的光照和图形难题。
2016年 Pascal	适用于大数据工作负载，采用HBM2的CoWoS 技术，16 纳米 FINFET 工艺。在深度学习方面，由 Pascal 支持的系统的神经网络训练性能提高了12倍
2017年 Volta	1.640个TENSOR内核，巨大的性能飞跃;Volta配备640个Tensor内核，可提供每秒超过100 万亿 Volta 次(TFLOPS)的深度学习性能。 2.它将 CUDA内核和Tensor 内核搭配使用，在 GPU 中提供人工智能超级计算机的性能。
2018年 Turing	1.用于AI加速的TENSORCORE，用于实时光线追踪的 RT CORE 2.Turing 利用多达4608个CUDA核心及软件开发套件(SDK)创建复杂的模拟 3.Turing架构能够借助增强的图形管线和全新可编程着色技术显著提高光栅性能
2020年 Ampere	基于AmpereGPU架构，基于TSMC7nm制程;NVIDIA第一个统一了数据分析，训练和推理的弹性多实例GPU。基于TF32的第三代张量核，NVLink3，结构稀疏性等特性
2022年 Hopper	NVIDIA H100集成了800亿个晶体管，采用台积电N4工艺，是全球范围内最大的加速器，拥有Transformer引擎和高度可扩展的NVLink互连技术（最多可连接达256个H100 GPU，相较于上一代采用HDR Quantum InfiniBand网络，带宽高出9倍，带宽速度为900GB/s）等功能，可推动庞大的AI语言模型、深度推荐系统、基因组学和复杂数字孪生的发展。

Archived Releases

CUDA Toolkit 12.0.1 (January 2023), Versioned Online Documentation
CUDA Toolkit 12.0.0 (December 2022), Versioned Online Documentation
CUDA Toolkit 11.8.0 (October 2022), Versioned Online Documentation
CUDA Toolkit 11.7.1 (August 2022), Versioned Online Documentation
CUDA Toolkit 11.7.0 (May 2022), Versioned Online Documentation
CUDA Toolkit 11.6.2 (March 2022), Versioned Online Documentation
CUDA Toolkit 11.6.1 (February 2022), Versioned Online Documentation
CUDA Toolkit 11.6.0 (January 2022), Versioned Online Documentation
CUDA Toolkit 11.5.2 (February 2022), Versioned Online Documentation
CUDA Toolkit 11.5.1 (November 2021), Versioned Online Documentation
CUDA Toolkit 11.5.0 (October 2021), Versioned Online Documentation
CUDA Toolkit 11.4.4 (February 2022), Versioned Online Documentation
CUDA Toolkit 11.4.3 (November 2021), Versioned Online Documentation
CUDA Toolkit 11.4.2 (September 2021), Versioned Online Documentation
CUDA Toolkit 11.4.1 (August 2021), Versioned Online Documentation
CUDA Toolkit 11.4.0 (June 2021), Versioned Online Documentation
CUDA Toolkit 11.3.1 (May 2021), Versioned Online Documentation
CUDA Toolkit 11.3.0 (April 2021), Versioned Online Documentation
CUDA Toolkit 11.2.2 (March 2021), Versioned Online Documentation
CUDA Toolkit 11.2.1 (February 2021), Versioned Online Documentation
CUDA Toolkit 11.2.0 (December 2020), Versioned Online Documentation
CUDA Toolkit 11.1.1 (October 2020), Versioned Online Documentation
CUDA Toolkit 11.1.0 (September 2020), Versioned Online Documentation
CUDA Toolkit 11.0.3 (August 2020), Versioned Online Documentation
CUDA Toolkit 11.0.2 (July 2020), Versioned Online Documentation
CUDA Toolkit 11.0.1 (June 2020), Versioned Online Documentation
CUDA Toolkit 10.2 (Nov 2019), Versioned Online Documentation
CUDA Toolkit 10.1 update2 (Aug 2019), Versioned Online Documentation
CUDA Toolkit 10.1 update1 (May 2019), Versioned Online Documentation
CUDA Toolkit 10.1 (Feb 2019), Online Documentation
CUDA Toolkit 10.0 (Sept 2018), Online Documentation
CUDA Toolkit 9.2 (May 2018),Online Documentation
CUDA Toolkit 9.1 (Dec 2017), Online Documentation
CUDA Toolkit 9.0 (Sept 2017), Online Documentation
CUDA Toolkit 8.0 GA2 (Feb 2017), Online Documentation
CUDA Toolkit 8.0 GA1 (Sept 2016), Online Documentation
CUDA Toolkit 7.5 (Sept 2015)

如何选择版本

准则2：TRT建议选择稳定版本，TensorRT6.0GA，TensorRT7.2.3和TensorRT8.5GA。
TRT对CUDA和CUDNN版本有强要求。

TensorRT6.0GA 支持 CUDA9.0，10.0，10.1和10.2

Tar File Install Packages For Linux x86

- TensorRT 6.0.1.8 GA for Ubuntu 18.04 and CUDA 10.2 tar package
- TensorRT 6.0.1.5 GA for Ubuntu 18.04 and CUDA 10.1 tar package
- TensorRT 6.0.1.5 GA for Ubuntu 18.04 and CUDA 10.0 tar package
- TensorRT 6.0.1.8 GA for Ubuntu 16.04 and CUDA 10.2 tar package
- TensorRT 6.0.1.5 GA for Ubuntu 16.04 and CUDA 10.1 tar package
- TensorRT 6.0.1.5 GA for Ubuntu 16.04 and CUDA 10.0 tar package
- TensorRT 6.0.1.5 GA for Ubuntu 16.04 and CUDA 9.0 tar package
- TensorRT 6.0.1.5 GA for Ubuntu 14.04 and CUDA 10.1 tar package
- TensorRT 6.0.1.5 GA for Ubuntu 14.04 and CUDA 10.0 tar package
- TensorRT 6.0.1.5 GA for Ubuntu 14.04 and CUDA 9.0 tar package
- TensorRT 6.0.1.8 GA for CentOS/RedHat 7 and CUDA 10.2 tar package
- TensorRT 6.0.1.5 GA for CentOS/RedHat 7 and CUDA 10.1 tar package
- TensorRT 6.0.1.5 GA for CentOS/RedHat 7 and CUDA 10.0 tar package
- TensorRT 6.0.1.5 GA for CentOS/RedHat 7 and CUDA 9.0 tar package

如何选择版本

准则2：TRT建议选择稳定版本，TensorRT6.0GA，TensorRT7.2.3和TensorRT8.5GA。
TRT对CUDA和CUDNN版本有强要求。

TensorRT7.2.3 支持 CUDA10.2，11.0 11.1 和 11.2

TensorRT 7.2.3 for Linux and CUDA 11.1 & 11.2

Debian, RPM and TAR Install Packages for x86_64 Architecture

- TensorRT 7.2.3 for Ubuntu 18.04 and CUDA 11.1 & 11.2 DEB local repo package
- TensorRT 7.2.3 for Ubuntu 18.04 and CUDA 11.1 & 11.2 TAR package
- TensorRT 7.2.3 for Ubuntu 16.04 and CUDA 11.1 & 11.2 DEB local repo package
- TensorRT 7.2.3 for Ubuntu 16.04 and CUDA 11.1 & 11.2 TAR package
- TensorRT 7.2.3 for CentOS / RedHat 7 and CUDA 11.1 & 11.2 RPM local repo package
- TensorRT 7.2.3 for CentOS / RedHat 7 and CUDA 11.1 & 11.2 TAR package

TensorRT 7.2.3 for Linux and CUDA 11.0

Debian, RPM and TAR Install Packages for x86_64 Architecture

- TensorRT 7.2.3 for Ubuntu 18.04 and CUDA 11.0 DEB local repo package
- TensorRT 7.2.3 for Ubuntu 18.04 and CUDA 11.0 TAR package
- TensorRT 7.2.3 for Ubuntu 16.04 and CUDA 11.0 DEB local repo package
- TensorRT 7.2.3 for Ubuntu 16.04 and CUDA 11.0 TAR package
- TensorRT 7.2.3 for CentOS / RedHat 7 and CUDA 11.0 RPM local repo package
- TensorRT 7.2.3 for CentOS / RedHat 7 and CUDA 11.0 TAR package

TensorRT 7.2.3 for Linux and CUDA 10.2

Debian, RPM and TAR Install Packages for x86_64 Architecture

- TensorRT 7.2.3 for Ubuntu 18.04 and CUDA 10.2 DEB local repo package
- TensorRT 7.2.3 for Ubuntu 18.04 and CUDA 10.2 TAR package
- TensorRT 7.2.3 for Ubuntu 16.04 and CUDA 10.2 DEB local repo package
- TensorRT 7.2.3 for Ubuntu 16.04 and CUDA 10.2 TAR package

如何选择版本

准则2：TRT建议选择稳定版本，TensorRT6.0GA，TensorRT7.2.3和TensorRT8.5GA。TRT对CUDA和CUDNN版本有强要求。

TensorRT8.5GA 支持所有cuda11版本。

如何选择版本-举例

两个准则:

1、自2016年以来，**每新出一代显卡，为了支持显卡的新特性，CUDA会对应更新一个大版本**。根据显卡架构选择当代或者下一代的CUDA。比如T4选择安装CUDA10.2 或者 CUDA11.x。不建议跨太大版本，有可能会存在负优化情况。

2、TensorRT近几年更新比较频繁，只建议选择**稳定版本**，推荐三个版本：TensorRT6.0GA，TensorRT7.2.3和TensorRT8.5GA。TRT对CUDA和CUDNN版本有强要求。

比如Tesla T4显卡，推荐组合就是CUDA10.2+TensorRT7.2.3 和 CUDA11.0、1、2+TensorRT8.5GA

确定CUDA版本和TensorRT版本，CUDNN版本就确定下来了

手动安装实操

唯一建议方案，官网安装包安装。
CUDA driver：https://www.nvidia.com/download/index.aspx
CUDA版本和驱动版本要求：https://docs.nvidia.com/cuda/cuda-toolkit-release-notes/index.html
CUDA各个版本：https://developer.nvidia.com/cuda-toolkit-archive
TensorRT各个版本：https://developer.nvidia.com/nvidia-tensorrt-download
CUDNN各个版本：https://developer.nvidia.com/rdp/cudnn-archive

举例：Tesla T4显卡，CUDA10.2+TensorRT7.2.3

手动安装实操

举例：Tesla T4显卡，CUDA10.2+TensorRT7.2.3

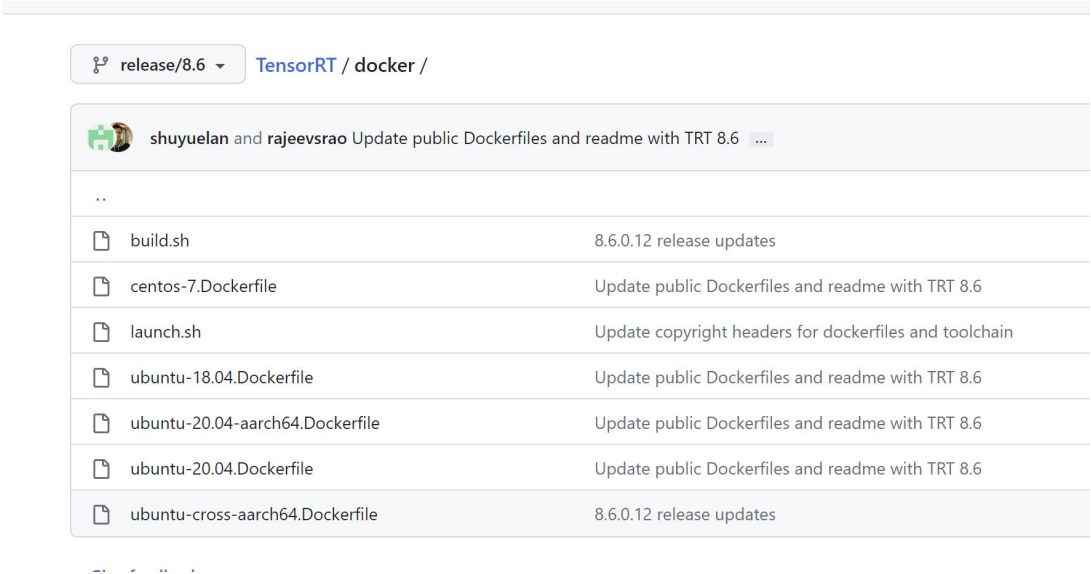
考虑问题：
1、Driver手动安装还是使用CUDA安装包内的版本？
如果后续需要升级CUDA，或者需要共存多个版本的CUDA，建议手动安装。
2、机器上是否已经安装Driver？版本是否满足要求？

唯一建议方案，官网安装包安装。
CUDA driver：https://www.nvidia.com/download/index.aspx
CUDA版本和驱动版本要求：https://docs.nvidia.com/cuda/cuda-toolkit-release-notes/index.html

CUDA各个版本：https://developer.nvidia.com/cuda-toolkit-archive
TensorRT各个版本：https://developer.nvidia.com/nvidia-tensorrt-download
CUDNN各个版本：https://developer.nvidia.com/rdp/cudnn-archive

Docker 安装-我没用过

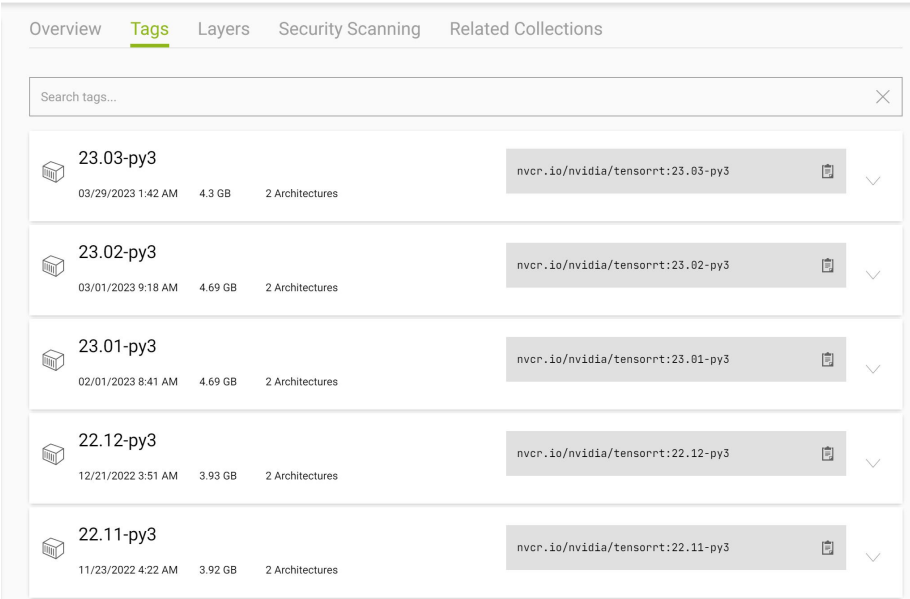
方式一： Dockerfile
https://github.com/NVIDIA/TensorRT



Docker 安装-我没用过

方式二：docker 源

https://catalog.ngc.nvidia.com/orgs/nvidia/containers/tensorrt



没找到版本信息.....

深蓝学院
shenlanxueyuan.com

感谢聆听！

Thanks for Listening