先进编译实验室
Advanced Compiler

Triton的安装方式

嘉宾：孙春豪

先进编译实验室
Advanced Compiler

Triton安装方式简述(一)

环境配置：

Ubuntu 22.04
Python 3.10
Cuda 12.1

以安装目前最新版本triton3.0.0为例

一、 pip 安装

Quick Installation

You can install the latest stable release of Triton from pip:

```
pip install triton
```

Triton安装方式简述(二)

二、 源码安装

```
git clone https://github.com/triton-lang/triton.git;
cd triton;

pip install ninja cmake wheel; # build-time dependencies
pip install -e python
```

(1) 下载triton源码
(2) 创建虚拟环境
(3) 下载llvm源码
(4) 源码编译llvm
(5) 构建triton
(6) 结果验证

Building with a custom LLVM

Triton uses LLVM to generate code for GPUs and CPUs. Normally, the Triton build downloads a prebuilt LLVM, but you can also build LLVM from source and use that.

LLVM does not have a stable API, so the Triton build will not work at an arbitrary LLVM version.

1. Find the version of LLVM that Triton builds against. Check `cmake/llvm-hash.txt` to see the current version. For example, if it says: 49af6502c6dcb4a7f7520178bd14df396f78240c

   This means that the version of Triton you have builds against LLVM 49af6502.

2. `git checkout` LLVM at this revision. Optionally, make additional modifications to LLVM.

3. Build LLVM. For example, you might run

```
$ cd $HOME/llvm-project  # your clone of LLVM.
$ mkdir build
$ cd build
$ cmake -G Ninja -DCMAKE_BUILD_TYPE=Release -DLLVM_ENABLE_ASSERTIONS=ON ../llvm -DLLVM_ENABLE_PROJE
$ ninja
```

4. Grab a snack, this will take a while.

5. Build Triton as above, but set the following environment variables.

```
# Modify as appropriate to point to your LLVM build.
$ export LLVM_BUILD_DIR=$HOME/llvm-project/build

$ cd <triton install>
$ LLVM_INCLUDE_DIRS=$LLVM_BUILD_DIR/include \
  LLVM_LIBRARY_DIR=$LLVM_BUILD_DIR/lib \
  LLVM_SYSPATH=$LLVM_BUILD_DIR \
  pip install -e python
```

Triton安装过程(一)

一、 pip 安装

pip install triton==3.0.0

```
root@autodl-container-8c094bba4f-27ffa067:~# pip install triton==3.0.0
Looking in indexes: http://mirrors.aliyun.com/pypi/simple
Collecting triton==3.0.0
  Downloading http://mirrors.aliyun.com/pypi/packages/45/27/14cc3101409b9b4b92
```

```
root@autodl-container-8c094bba4f-27ffa067:~# python
Python 3.10.8 (main, Nov 24 2022, 14:13:03) [GCC 11.2.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import triton
>>>
```

Triton安装过程(二)

二、 源码安装 (源码编译llvm)

1. 下载triton源码

```
root@autodl-container-00274fbfbe-a9383609:~# git clone https://github.com/openai/triton.git
Cloning into 'triton'...
remote: Enumerating objects: 134712, done.
remote: Counting objects: 100% (5945/5945), done.
remote: Compressing objects: 100% (1051/1051), done.
remote: Total 134712 (delta 5183), reused 5407 (delta 4850), pack-reused 128767
Receiving objects: 100% (134712/134712), 319.07 MiB | 6.36 MiB/s, done.
Resolving deltas: 100% (93808/93808), done.
```

1

补充：需要安装指定版本可去官方仓库查看commit id

git checkout {commit id}

Triton安装过程(二)

二、 源码安装 (源码编译llvm)

2. 创建虚拟环境 && 安装依赖

```
root@autodl-container-00274fbfbe-a9383609:~#
root@autodl-container-00274fbfbe-a9383609:~# cd triton
root@autodl-container-00274fbfbe-a9383609:~/triton# python -m venv .venv --prompt triton
root@autodl-container-00274fbfbe-a9383609:~/triton# source /root/triton/.venv/bin/activate
(triton) root@autodl-container-00274fbfbe-a9383609:~/triton#
```

2

```
(triton) root@autodl-container-00274fbfbe-a9383609:~/triton# pip install ninja cmake wheel
Looking in indexes: http://mirrors.aliyun.com/pypi/simple
Collecting ninja
```

3

```
00274fbfbe-a9383609:~/triton# pip install scipy numpy  pytest lit pandas matplotlib
ors.aliyun.com/pypi/simple
```

4

Triton安装过程(二)

二、 源码安装 (源码编译llvm)

3. 下载llvm源码

(获取与当前triton版本适配的llvm的分支)

```
(triton) root@autodl-container-00274fbfbe-a9383609:~/triton# cat ~/triton/cmake/llvm-hash.txt
4713bd4ccc0c0d568f92916e7851d993291742c0
```

5

```
(triton) root@autodl-container-00274fbfbe-a9383609:~/triton# cd
(triton) root@autodl-container-00274fbfbe-a9383609:~# git clone https://github.com/llvm/llvm-project
Cloning into 'llvm-project'...
remote: Enumerating objects: 6041634, done.
remote: Counting objects: 100% (4600/4600), done.
remote: Compressing objects: 100% (1522/1522), done.
remote: Total 6041634 (delta 3942), reused 3262 (delta 3073), pack-reused 6037034
Receiving objects: 100% (6041634/6041634), 1.31 GiB | 8.38 MiB/s, done.
Resolving deltas: 100% (4999151/4999151), done.
Updating files: 100% (146235/146235), done.
```

6

Triton安装过程(二)

二、 源码安装 (源码编译llvm)

3. 下载llvm源码

(切换到适配当前triton的llvm版本)

```
(triton) root@autodl-container-00274fbfbe-a9383609:~# cd llvm-project/
(triton) root@autodl-container-00274fbfbe-a9383609:~/llvm-project#
(triton) root@autodl-container-00274fbfbe-a9383609:~/llvm-project# git checkout 4713bd4ccc0c0d568f92916e7851d993291742c0
Updating files: 100% (9830/9830), done.
Note: switching to '4713bd4ccc0c0d568f92916e7851d993291742c0'.
```

Triton安装过程(二)

二、 源码安装 (源码编译llvm)

4. 源码编译llvm

```
(triton) root@autodl-container-00274fbfbe-a9383609:~/llvm-project/build# cmake -G Ninja -DCMAKE_BUILD_TYPE=Release  \
-DLLVM_ENABLE_ASSERTIONS=ON ../llvm    \
-DLLVM_ENABLE_PROJECTS="mlir;llvm"    \
-DLLVM_TARGETS_TO_BUILD="host;NVPTX;AMDGPU"
```

8

```
container-00274fbfbe-a9383609:~/llvm-project/build# ninja -j24
(X object lib/Support/CMakeFiles/LLVMSupport.dir/MemoryBuffer.cpp.o
```

9

```
[4898/4898] Generating ../../bin/llvm-readelf
(triton) root@autodl-container-00274fbfbe-a9383609:~/llvm-p
```

Triton安装过程(二)

二、 源码安装 (源码编译llvm)

5. 构建triton

```
(triton) root@autodl-container-00274fbfbe-a9383609:~/llvm-project/build# export LLVM_BUILD_DIR=/root/llvm-project/build
(triton) root@autodl-container-00274fbfbe-a9383609:~/llvm-project/build# cd ~/triton
(triton) root@autodl-container-00274fbfbe-a9383609:~/triton# LLVM_INCLUDE_DIRS=$LLVM_BUILD_DIR/include \
LLVM_LIBRARY_DIR=$LLVM_BUILD_DIR/lib \
LLVM_SYSPATH=$LLVM_BUILD_DIR \
pip install -e python -i  https://pypi.tuna.tsinghua.edu.cn/simple
```

10

```
  Stored in directory: /tmp/pip-ephem-wheel-cache-ozpllbaz/
Successfully built triton
Installing collected packages: triton
Successfully installed triton-3.0.0
```

Triton安装过程(二)

二、 源码安装 (源码编译llvm)

6. 结果验证

```
(triton) root@autodl-container-00274fbfbe-a9383609:~/triton/python/tutorials# python 01-vector-add.py
tensor([1.3713, 1.3076, 0.4940,  ..., 0.6724, 1.2141, 0.9733], device='cuda:0')
tensor([1.3713, 1.3076, 0.4940,  ..., 0.6724, 1.2141, 0.9733], device='cuda:0')
The maximum difference between torch and triton is 0.0
vector-add-performance:
           size      Triton       Torch
0        4096.0   12.000000   12.387097
1        8192.0   24.000000   24.000000
2       16384.0   48.000000   48.000000
3       32768.0   96.000000   96.000000
4       65536.0  153.600004  153.600004
5      131072.0  255.999991  255.999991
6      262144.0  388.553351  384.000001
7      524288.0  511.999982  558.545450
8     1048576.0  722.823517  714.938186
9     2097152.0  722.823517  768.000002
```

11

AdvancedCompiler

Tel: 13839830713