关于模型优化做的工作

完成DCU移植

libtorch提供C++接口，性能比Pytho版本提升0.1（有总比没有强）

在此切换Torch_DIR

```
export LD_LIBRARY_PATH="/home/worldpeace/soft/libtorch/lib:$LD_LIBRARY_PATH"
export PATH="/home/worldpeace/soft/deepmd-c++/bin:$PATH"
export LD_LIBRARY_PATH="/home/worldpeace/soft/deepmd-c++/lib:$LD_LIBRARY_PATH"
cmake -L -C ../cmake/presets/basic.cmake  \
-C ../cmake/presets/kokkos-openmp.cmake \
-C ../cmake/presets/kokkos-cuda.cmake \
-DCMAKE_BUILD_TYPE=Release \
#   取消注释二选一    #
#原始版本# -DTorch_DIR=`python -c 'import torch;print(torch.utils.cmake_prefix_path)'`/Torch \
#libtorch版本# -DTorch_DIR=/home/worldpeace/soft/libtorch/share/cmake/Torch \
-DGFLAGS_INCLUDE_DIR=/home/worldpeace/soft/libtorch/include \
-DCUDA_ARCH=AMPERE86 \
-DMKL_INCLUDE_DIR=/opt/intel/oneapi/mkl/latest/include \
-DCMAKE_PREFIX_PATH=/home/worldpeace/soft/deepmd-kit\
-DCMAKE_INSTALL_PREFIX=/opt/LMP_dp_allegro_C -DBUILD_TOOLS=ON -DBUILD_SHARED_LIBS=ON \
-DPKG_GPU=ON  \
-DFFT=FFTW3 -DFFTW3_LIBRARY=/opt/fftw3/lib/libfftw3.so \
-DFFTW3_INCLUDE_DIR=/opt/fftw3/include \
-DLAMMPS_INSTALL_RPATH=ON  ../cmake
```

TVM调研实践，卡在Input

input代码分析，获得Input形状

/home/worldpeace/anaconda3/envs/tvm/lib/python3.11/site-packages/nequip/ase/nequip_calculator.py

```
    def calculate(self, atoms=None, properties=["energy"], system_changes=all_changes):
        """
        Calculate properties.

        :param atoms: ase.Atoms object
        :param properties: [str], properties to be computed, used by ASE internally
        :param system_changes: [str], system changes since last calculation, used by ASE internally
        :return:
        """
        # call to base-class to set atoms attribute
        Calculator.calculate(self, atoms)
        # prepare data
        data = AtomicData.from_ase(atoms=atoms, r_max=self.r_max)
        for k in AtomicDataDict.ALL_ENERGY_KEYS:
            if k in data:
                del data[k]
        data = self.transform(data)
        data = data.to(self.device)
        data = AtomicData.to_AtomicDataDict(data)

        # predict + extract data
        out = self.model(data)
```

这里data为模型需要的输入，获得并保存在txt内部

```
{'edge_index': tensor([[ 0,  0,  0,  ..., 70, 70, 70],        [ 3,  5,  6,  ...,
47, 42, 18]], device='cuda:0'), 'pos': tensor([[1.1242e-01, 1.2245e+01,
4.1290e+00],        [3.6315e-01, 4.2438e+00, 8.0531e+00],        [9.3500e-02,
4.5374e+00, 4.3283e+00],        [1.2334e-01, 8.6023e+00, 8.1177e+00],
[1.2322e+01, 8.6604e+00, 4.3235e+00],        [4.2572e+00, 1.2474e+01,
8.0820e+00],        [4.3863e+00, 1.2205e+01, 4.0424e+00],        [4.3404e+00,
3.7507e+00, 8.1078e+00],        [4.3428e+00, 3.9973e+00, 3.9187e+00],
```

添加tensordict库，修改格式为可识别

```
input=AtomicData.to_AtomicDataDict({'edge_index': tensor([[ 0,  0,  0,  0,  0,
0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,          0,  0,  0,  0,  0,
0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,          0,  0,  0,  0,  0,
0,  0,  0,  0,  0,  0,  0,  1,  1,  1,  1,  1,  1,          1,  1,  1,  1,  1,
1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,          1,  1,  1,  1,  1,
1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,          1,  1,  1,  1,  1,
1,  1,  2,  2,  2,  2,  2,  2,  2,  2,  2,  2,  2,          2,  2,  2,  2,  2,
2,  2,  2,  2,  2,  2,  2,  2,  2,  2,  2,  2,  2,          2,  2,  2,  2,  2,
2,  2,  2,  2,  2,  2,  2,  2,  2,
```

jit.load script model。然后运行

```
import torch
from torch import tensor
from nequip.data import AtomicData, AtomicDataDict
from tensordict.tensordict import TensorDict
mod=torch.jit.load('deployed.pth')
mod=mod.to('cuda')
input=AtomicData.to_AtomicDataDict({'edge_index': tensor([[
out=mod(input)
print(out)
```

正确性对比：

上面两个python形式已经对比无误

nequip_calculator.py out 和LAMMPS pair_allegro.cpp的out对比即可

ASE仿真，完成模型独立运行

Script model保存机制，C++调用单步性能插桩

pair_allegro.cpp输出

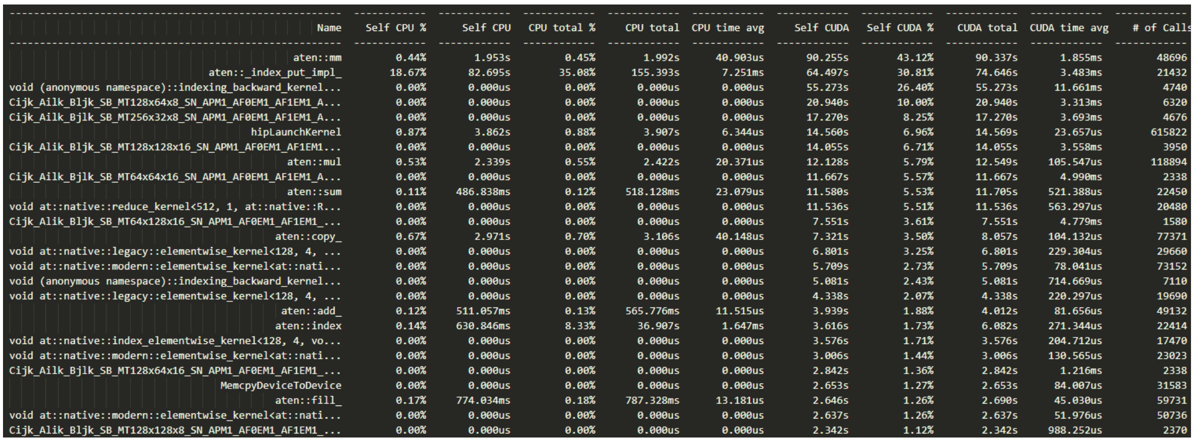
```
Per MPI rank memory allocation (min/avg/max) = 5.31 | 5.31 | 5.31 Mbytes   Step
       Time          PotEng         KinEng         TotEng          Temp
Press          Volume        Density
0   0              -5115.3514      247.66244     -4867.6889      1000
124912.37      35001.599      9.8101743
model.forward Time is : 0.445529 stoTensor().cpu Time is : 0.000252 s
Pair All Time is : 0.452624 s
model.forward Time is : 3.49675 s
toTensor().cpu Time is : 7.3e-05 s
Pair All Time is : 3.50579 s
model.forward Time is : 0.283893 s
toTensor().cpu Time is : 0.000103 s
Pair All Time is : 0.286683 s
model.forward Time is : 0.282643 s
toTensor().cpu Time is : 6e-05 s
Pair All Time is : 0.290819 s
model.forward Time is : 0.283949 s
toTensor().cpu Time is : 0.000106 s
Pair All Time is : 0.28688 s
model.forward Time is : 0.282495 s
toTensor().cpu Time is : 5.4e-05 s
Pair All Time is : 0.286974 s
model.forward Time is : 0.282829 s
toTensor().cpu Time is : 0.000113 s
```

python 单独load Model对比

```
/home/worldpeace/anaconda3/envs/tvm/lib/python3.11/site-
packages/nequip/__init__.py:20: UserWarning: !! PyTorch version 2.5.1 found.
Upstream issues in PyTorch versions 1.13.* and 2.* have been seen to cause
unusual performance degredations on some CUDA systems that become worse over
time; see https://github.com/mir-group/nequip/discussions/311. The best tested
PyTorch version to use with CUDA devices is 1.11; while using other versions if
you observe this problem, an unexpected lack of this problem, or other strange
behavior, please post in the linked GitHub issue.  warnings.warn(
Time :0.6278s
Time :0.5497s
Time :1.6636s
Time :0.3279s
Time :0.0221s
Time :0.0236s
Time :0.0239s
Time :0.0243s
Time :0.0231s
Time :0.0288s
```

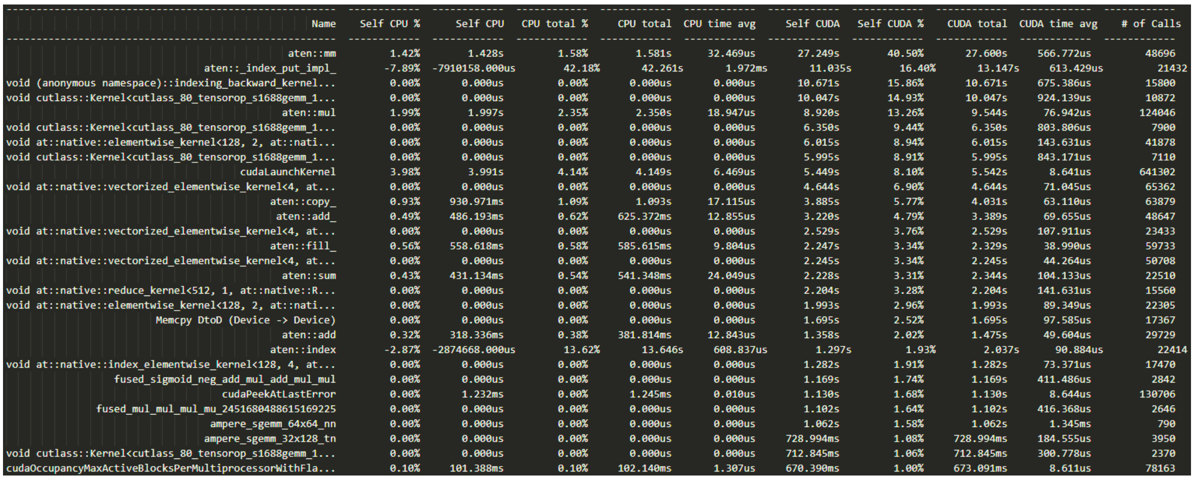
性能分析，不同硬件平台底层走不同计算库

Z100



```
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                   Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total  CUDA time avg    # of Calls
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                               aten::mm         0.44%        1.953s         0.45%        1.992s      40.903us       90.255s        43.12%       90.337s       1.855ms         48696
                                   aten::_index_put_impl_        18.67%       82.695s        35.08%      155.393s       7.251ms       64.497s        30.81%       74.646s       3.483ms         21432
void (anonymous namespace)::indexing_backward_kernel...         0.00%       0.000us         0.00%       0.000us       0.000us       55.273s        26.40%       55.273s      11.661ms          4740
Cijk_Ailk_Bljk_SB_MT128x64x8_SN_APM1_AF0EM1_AF1EM1_A...         0.00%       0.000us         0.00%       0.000us       0.000us       20.940s        10.00%       20.940s       3.313ms          6320
Cijk_Ailk_Bjlk_SB_MT256x32x8_SN_APM1_AF0EM1_AF1EM1_A...         0.00%       0.000us         0.00%       0.000us       0.000us       17.270s         8.25%       17.270s       3.693ms          4676
                                        hipLaunchKernel         0.87%        3.862s         0.88%        3.907s       6.344us       14.560s         6.96%       14.569s      23.657us        615822
Cijk_Alik_Bljk_SB_MT128x128x16_SN_APM1_AF0EM1_AF1EM1...         0.00%       0.000us         0.00%       0.000us       0.000us       14.055s         6.71%       14.055s       3.558ms          3950
                                              aten::mul         0.53%        2.339s         0.55%        2.422s      20.371us       12.128s         5.79%       12.549s     105.547us        118894
Cijk_Ailk_Bjlk_SB_MT64x64x16_SN_APM1_AF0EM1_AF1EM1_A...         0.00%       0.000us         0.00%       0.000us       0.000us       11.667s         5.57%       11.667s       4.990ms          2338
                                              aten::sum         0.11%     486.838ms         0.12%     518.128ms      23.079us       11.580s         5.53%       11.705s     521.388us         22450
void at::native::reduce_kernel<512, 1, at::native::R...         0.00%       0.000us         0.00%       0.000us       0.000us       11.536s         5.51%       11.536s     563.297us         20480
Cijk_Alik_Bljk_SB_MT64x128x16_SN_APM1_AF0EM1_AF1EM1_...         0.00%       0.000us         0.00%       0.000us       0.000us        7.551s         3.61%        7.551s       4.779ms          1580
                                            aten::copy_         0.67%        2.971s         0.70%        3.106s      40.148us        7.321s         3.50%        8.057s     104.132us         77371
void at::native::legacy::elementwise_kernel<128, 4, ...         0.00%       0.000us         0.00%       0.000us       0.000us        6.801s         3.25%        6.801s     229.304us         29660
void at::native::modern::elementwise_kernel<at::nati...         0.00%       0.000us         0.00%       0.000us       0.000us        5.709s         2.73%        5.709s      78.041us         73152
void (anonymous namespace)::indexing_backward_kernel...         0.00%       0.000us         0.00%       0.000us       0.000us        5.081s         2.43%        5.081s     714.669us          7110
void at::native::legacy::elementwise_kernel<128, 4, ...         0.00%       0.000us         0.00%       0.000us       0.000us        4.338s         2.07%        4.338s     220.297us         19690
                                             aten::add_         0.12%     511.057ms         0.13%     565.776ms      11.515us        3.939s         1.88%        4.012s      81.656us         49132
                                            aten::index         0.14%     630.846ms         8.33%       36.907s       1.647ms        3.616s         1.73%        6.082s     271.344us         22414
void at::native::index_elementwise_kernel<128, 4, vo...         0.00%       0.000us         0.00%       0.000us       0.000us        3.576s         1.71%        3.576s     204.712us         17470
void at::native::modern::elementwise_kernel<at::nati...         0.00%       0.000us         0.00%       0.000us       0.000us        3.006s         1.44%        3.006s     130.565us         23023
Cijk_Ailk_Bjlk_SB_MT128x64x16_SN_APM1_AF0EM1_AF1EM1_...         0.00%       0.000us         0.00%       0.000us       0.000us        2.842s         1.36%        2.842s       1.216ms          2338
                                   MemcpyDeviceToDevice         0.00%       0.000us         0.00%       0.000us       0.000us        2.653s         1.27%        2.653s      84.007us         31583
                                            aten::fill_         0.17%     774.034ms         0.18%     787.328ms      13.181us        2.646s         1.26%        2.690s      45.030us         59731
void at::native::modern::elementwise_kernel<at::nati...         0.00%       0.000us         0.00%       0.000us       0.000us        2.637s         1.26%        2.637s      51.976us         50736
Cijk_Alik_Bljk_SB_MT128x128x8_SN_APM1_AF0EM1_AF1EM1_...         0.00%       0.000us         0.00%       0.000us       0.000us        2.342s         1.12%        2.342s     988.252us          2370
```

RTX3090



```
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                   Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total  CUDA time avg    # of Calls
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                               aten::mm         1.42%        1.428s         1.58%        1.581s      32.469us       27.249s        40.50%       27.600s     566.772us         48696
                                   aten::_index_put_impl_        -7.89%  -7910158.000us        42.18%       42.261s       1.972ms       11.035s        16.40%       13.147s     613.429us         21432
void (anonymous namespace)::indexing_backward_kernel...         0.00%       0.000us         0.00%       0.000us       0.000us       10.671s        15.86%       10.671s     675.386us         15800
void cutlass::Kernel<cutlass_80_tensorop_s1688gemm_1...         0.00%       0.000us         0.00%       0.000us       0.000us       10.047s        14.93%       10.047s     924.139us         10872
                                              aten::mul         1.99%        1.997s         2.35%        2.350s      18.947us        8.920s        13.26%        9.544s      76.942us        124046
void cutlass::Kernel<cutlass_80_tensorop_s1688gemm_1...         0.00%       0.000us         0.00%       0.000us       0.000us        6.350s         9.44%        6.350s     803.806us          7900
void at::native::elementwise_kernel<128, 2, at::nati...         0.00%       0.000us         0.00%       0.000us       0.000us        6.015s         8.94%        6.015s     143.631us         41878
void cutlass::Kernel<cutlass_80_tensorop_s1688gemm_1...         0.00%       0.000us         0.00%       0.000us       0.000us        5.995s         8.91%        5.995s     843.171us          7110
                                       cudaLaunchKernel         3.98%        3.991s         4.14%        4.149s       6.469us        5.449s         8.10%        5.542s       8.641us        641302
void at::native::vectorized_elementwise_kernel<4, at...         0.00%       0.000us         0.00%       0.000us       0.000us        4.644s         6.90%        4.644s      71.045us         65362
                                            aten::copy_         0.93%     930.971ms         1.09%        1.093s      17.115us        3.885s         5.77%        4.031s      63.110us         63879
                                             aten::add_         0.49%     486.193ms         0.62%     625.372ms      12.855us        3.220s         4.79%        3.389s      69.655us         48647
void at::native::vectorized_elementwise_kernel<4, at...         0.00%       0.000us         0.00%       0.000us       0.000us        2.529s         3.76%        2.529s     107.911us         23433
                                            aten::fill_         0.56%     558.618ms         0.58%     585.615ms       9.804us        2.247s         3.34%        2.329s      38.990us         59733
void at::native::vectorized_elementwise_kernel<4, at...         0.00%       0.000us         0.00%       0.000us       0.000us        2.245s         3.34%        2.245s      44.264us         50708
                                              aten::sum         0.43%     431.134ms         0.54%     541.348ms      24.049us        2.228s         3.31%        2.344s     104.133us         22510
void at::native::reduce_kernel<512, 1, at::native::R...         0.00%       0.000us         0.00%       0.000us       0.000us        2.204s         3.28%        2.204s     141.631us         15560
void at::native::elementwise_kernel<128, 2, at::nati...         0.00%       0.000us         0.00%       0.000us       0.000us        1.993s         2.96%        1.993s      89.349us         22305
                          Memcpy DtoD (Device -> Device)         0.00%       0.000us         0.00%       0.000us       0.000us        1.695s         2.52%        1.695s      97.585us         17367
                                              aten::add         0.32%     318.336ms         0.38%     381.814ms      12.843us        1.358s         2.02%        1.475s      49.604us         29729
                                            aten::index        -2.87%  -2874668.000us        13.62%       13.646s     608.837us        1.297s         1.93%        2.037s      90.884us         22414
void at::native::index_elementwise_kernel<128, 4, at...         0.00%       0.000us         0.00%       0.000us       0.000us        1.282s         1.91%        1.282s      73.371us         17470
                     fused_sigmoid_neg_add_mul_add_mul_mul         0.00%       0.000us         0.00%       0.000us       0.000us        1.169s         1.74%        1.169s     411.486us          2842
                                    cudaPeekAtLastError         0.00%       1.232ms         0.00%       1.245ms       0.010us        1.130s         1.68%        1.130s       8.644us        130706
                 fused_mul_mul_mul_mu_2451680488615169225         0.00%       0.000us         0.00%       0.000us       0.000us        1.102s         1.64%        1.102s     416.368us          2646
                                   ampere_sgemm_64x64_nn         0.00%       0.000us         0.00%       0.000us       0.000us        1.062s         1.58%        1.062s       1.345ms           790
                                  ampere_sgemm_32x128_tn         0.00%       0.000us         0.00%       0.000us       0.000us     728.994ms         1.08%     728.994ms     184.555us          3950
void cutlass::Kernel<cutlass_80_tensorop_s1688gemm_1...         0.00%       0.000us         0.00%       0.000us       0.000us     712.845ms         1.06%     712.845ms     300.778us          2370
cudaOccupancyMaxActiveBlocksPerMultiprocessorWithFla...         0.10%     101.388ms         0.10%     102.140ms       1.307us     670.390ms         1.00%     673.091ms       8.611us         78163
```

性能分析脚本

找到主函数，替换

显示底层计算库算子耗时

```python
if __name__ == "__main__":
#     cProfile.run('main(running_as_script=True)')

    with torch.profiler.profile(
    activities=[
        torch.profiler.ProfilerActivity.CPU,
        torch.profiler.ProfilerActivity.CUDA,
    ]
) as p:
        main(running_as_script=True)
    print(p.key_averages().table(sort_by="self_cuda_time_total", row_limit=-1))
```

torchboard的分析

```
with torch.profiler.profile(
activities=[
torch.profiler.ProfilerActivity.CPU,
torch.profiler.ProfilerActivity.CUDA,
],
on_trace_ready=torch.profiler.tensorboard_trace_handler('./log/torchboard'),
record_shapes=True,
profile_memory=True,
with_stack=True) as p:
    main(running_as_script=True)
```

Magpy调研与应用，失败

后续方向

先重新保存模型 不用script model

尝试TVM编译model，Magpy