Report

Environment

Since the UCI cluster is way too slow, I rent an online GPU to do the experiment.

Spec:

- Vendor
  - vast.ai
  - 0.5$ / hr

- GPU
  - RTX 4090
  - 83 TFlops
  - 22.5GB, 3572GB/s

- CPU
  - AMD Ryzen Threadripper 5975X 32-Core Processor

- Peripherals
  - PCIe 4.0 16x, 24.3GB/s
  - nvme, 3684MB/s, 16GB
  - RAM 37GB

- Env
  - Ubuntu 20.04
  - CUDA 12.2

Scheme

I let each thread to calculate each operation. Only if the index is in the range, the thread will do the calculation.

The kernel is as follows:

```
__global__ void stencil_kernel(float* temp, float* temp2, float* conduct
    int i = blockIdx.x * blockDim.x + threadIdx.x;
    int x = i % w, y = i / w;
    if (x > 0 && y > 0 && x < (w-1) && y < (h-1)) {
        float e = temp[i];
        temp2[i] = e +
            (
                (temp[i-1] - e) * conduct[i-1]
                +
                (temp[i+1] - e) * conduct[i+1]
                +
                (temp[i-w] - e) * conduct[i-w]
                +
                (temp[i+w] - e) * conduct[i+w]
            ) * 0.2;
    }
}
```

I try to write the halo version, but I had difficulty in implementing it, so I just use the naive version.

Evaluation

- CPU native time: 1.4s
- GPU native time: 9.7s
- GPU parallel time: 0.13s ~ 0.16s

In my simple setup, the GPU parallel version is 10x faster than the CPU native version and 60x fater than the GPU native version.