MadGraph5 GPU Development

Kernel Profiling

A. Thete & C. Vuosalo

University of Wisconsin-Madison

25.06.2024

ALOHA Subroutine Variations

- Recall the memory bottleneck was line 1107 in `FFV1_0` of HelAmps_sm.h to compute output amplitude vertex from input 3 wavefunctions

```
  const cxtype_sv TMP9 = ( F1[2] * ( F2[4] * ( V3[2] + V3[5] ) + F2[5] * ( V3[3] + cI * V3[4] ) )
+ . . .
```

- Tested 5 variations on the generated subroutine (thanks Olivier!) with different levels of modification on that line:
    - NoFact
    - FactByWFact
    - VectorFirst
    - VectorByHand
    - VectorByHand2

ALOHA Subroutine Variations

Routine	Wall Time [ms]	Memory Throughput [%]	Warp Stall Statistics	Comments
NoFact	67.45 (+0.02%)	62.32 (+0.05%)	13.21% (-0.92 pp)	45.11% LS, 19.31% W, 11.12% NI
FactByWFact	67.41 (-0.04%)	62.31 (+0.03%)	17.04% (+2.91 pp)	47.00% LS, 17.55% W, 12.39% NI
VectorFirst	67.34 (-0.14%)	62.28 (-0.02%)	14.06% (-0.06 pp)	43.42% LS, 19.06% W, 13.42% NI
VectorByHand	67.41 (-0.03%)	62.34 (+0.08%)	16.94% (+2.81 pp)	45.52% LS, 18.44% W, 11.55% NI
VectorByHand2	**67.33 (-0.15%)**	62.30 (+0.02%)	**12.8% (-1.33 pp)**	39.88% LS, 19.43% W, 14.10% NI
Baseline	67.44	62.29	14.13%	41.79% LS, 19.77% W, 12.69% NI

Long Scoreboard (LS): *Warp stalled waiting for L1TEX data*

Wait (W): *Warp stalled waiting on a fixed latency execution; typically shows up as top contributor in highly-optimized kernels*

No Instruction (NI): *Warp stalled waiting to be selected to fetch an instruction or waiting on an instruction cache miss.*