GPU BOARD EXECUTION TIME COMPARISON

A30 VS A40 BOARDS

DATA SAMPLE USED

Data sample : AMS-02_PRL2019

- Ion : He3
- Starting date : 19/05/2011
- End date : 15/11/2017
- (r, θ, φ) : (1, 0, 0)
- DOI: https://dx.doi.org/10.1103/PhysRevLett.123.181102

Simulated energy (only 1 energy bin to speedup the test and remove all incidental operations)

- Rigidity : 12.5000
- Flux : 0.244600
- Initial Energy : 7.45 GeV

GPU HARWARE AND CAPABILITIES DESCRIPTION

NVIDIA A30

- Compute capability: 8.0
- Clock rate: 1 440 000
- Total global mem: 25 229 983 744
- Total constant Mem: 65 536
- Texture Alignment: 512
- Multiprocessor count: 56
- Shared mem per mp: 49 152
- Registers per mp: 65 536
- Threads in warp: 32
- Warps per Multiprocessor: 64
- Threads per Multiprocessor: 2048
- Thread Blocks per Multiprocessor: 32
- Max threads per block: 1024

NVIDIA A40

- Compute capability: 8.6
- Clock rate: 1 740 000
- Total global mem: 47 619 112 960
- Total constant Mem: 65 536
- Texture Alignment: 512
- Multiprocessor count: 84
- Shared mem per mp: 49 152
- Registers per mp: 65 536
- Threads in warp: 32
- Warps per Multiprocessor: 48
- Threads per Multiprocessor: 1536
- Thread Blocks per Multiprocessor: 16
- Max threads per block: 1024

SIMULATION LAUNCH CONFIGURATION

- The number of 5000 particles is rounded to 5632 to fit the warpsize
- The number of blocks is computed by rounding the Npart / WpB*WarpSize ratio
- The number of threads per block is computed by rounding the Npart / Nblocks ratio

Execution example

Propagation Kernel :

- Max Number ofWarp in a Block : 2
- Number of blocks : 88
- Number of threadsPerBlock : 6

Histogram Kernel :

- Number of Warp in a Block : 1024
- Number of blocks : 6
- Number of threadsPerBlock : 939

SIMULATION LAUNCH CONFIGURATION

- The number of warps per block was varied from 2 to 32 for each GPU board (only the avilable values of warps per block are taken into account)

Register allocation unit size	256
Register allocation granularity	warp
Max registers per Block	65536
Warp allocation granularity (for register allocation)	4
Registers used by the kernel	106

⬇

With more that 16 warps we exceed the GPU resources

⬇

Nregisters = 73 728 for 18 Warps

Registrs allocated = int_round(Nregisters_per_kernel*Warp_size / Register_alloc_unit*Warp_granuularity) *Register_alloc_unit*Warp_granuularity * Nwarps

We use 106 registers for the heliospheric propagation kernel function (maximum registers per thread = 32)

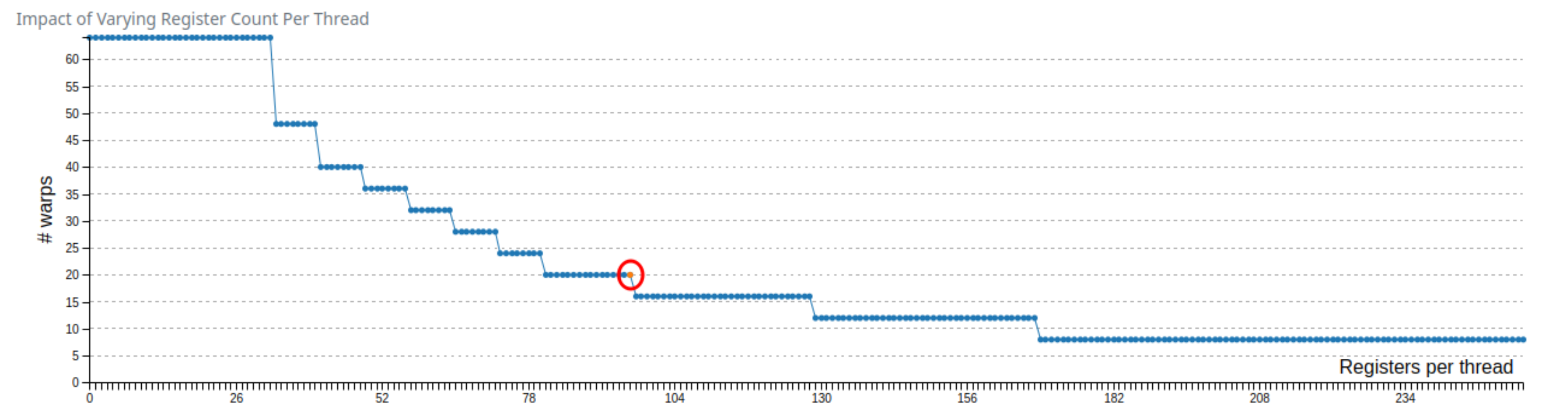
⬇

Registers from other warps are allocated and the number of active warps is sub-optimal

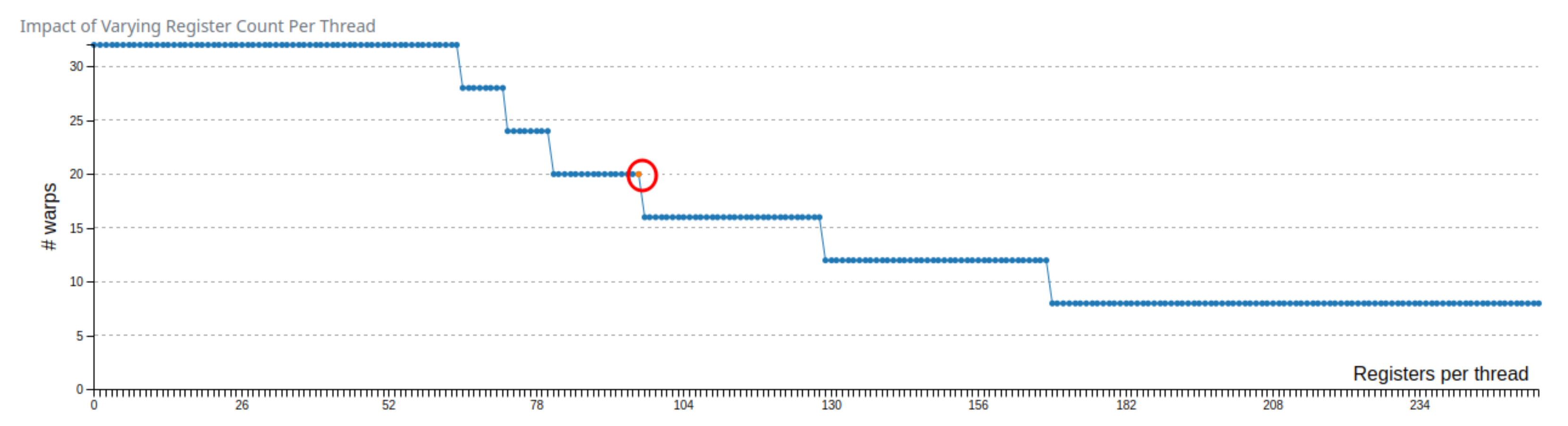
NVIDIA OCCUPANCY CALCULATOR

GPU MEMORY OCCUPANCY

A30
(compute capability 8.0)



A40
(compute capability 8.6)

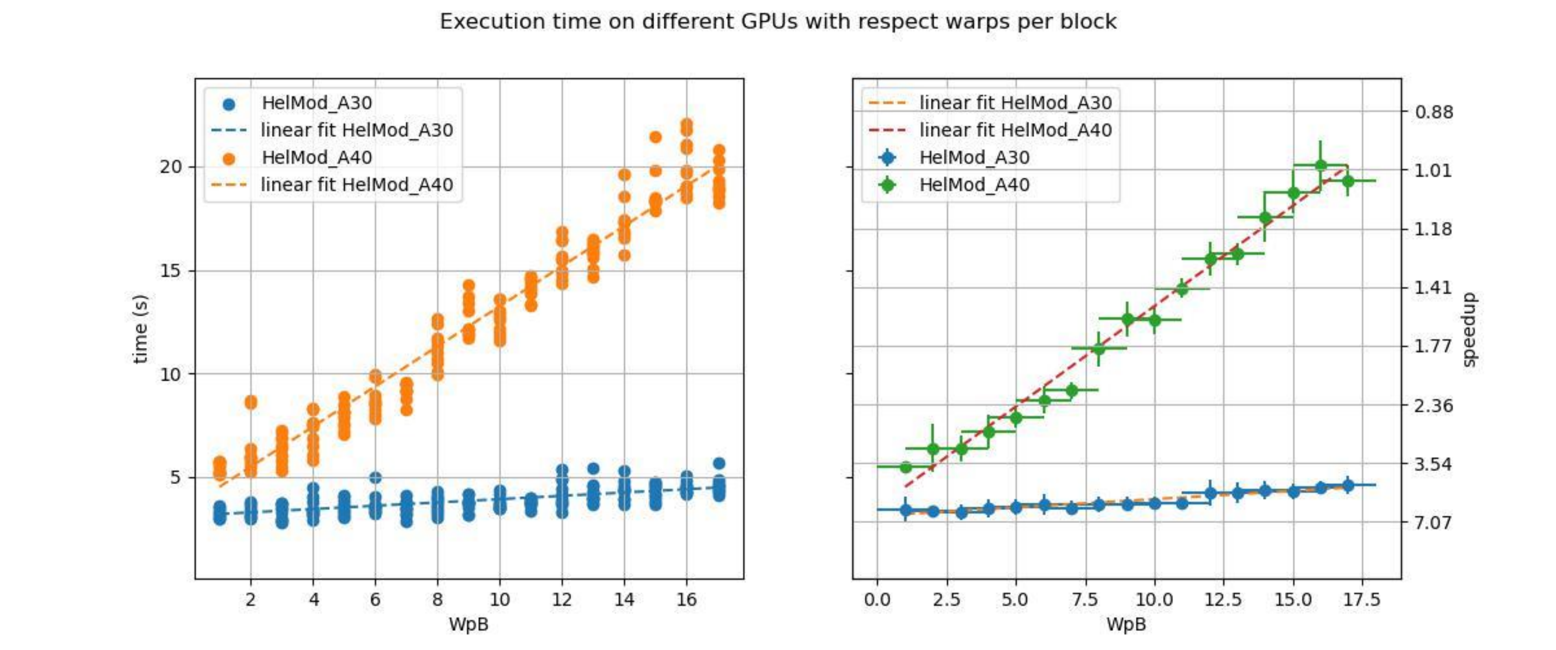


RUN OUTPUT AND CODE SECTION EXECUTION TIME

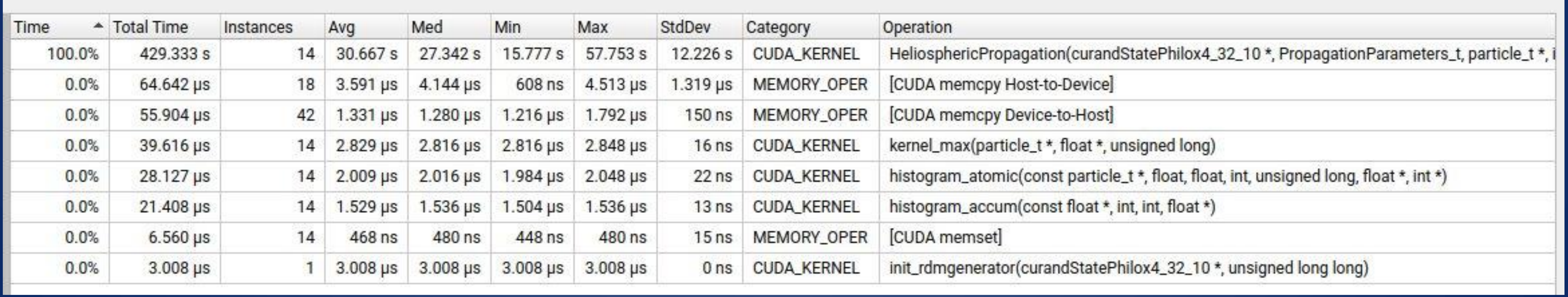
- EMin = 7.454
- Emax = 9.814
- N Output binning: 33

- Time to back-propagate particles:
  - Init : 0.06 ms
  - **propagation phase : 3911.77 ms**
  - Find Max : 0.33 ms
  - Binning : 0.15 ms
- Time to Set Memory: 12.0 ms
- Time to create Rnd: 0.1 ms
- **Time to execute : 3924.4 ms**

**The code was executed
10 times for each WpB value**



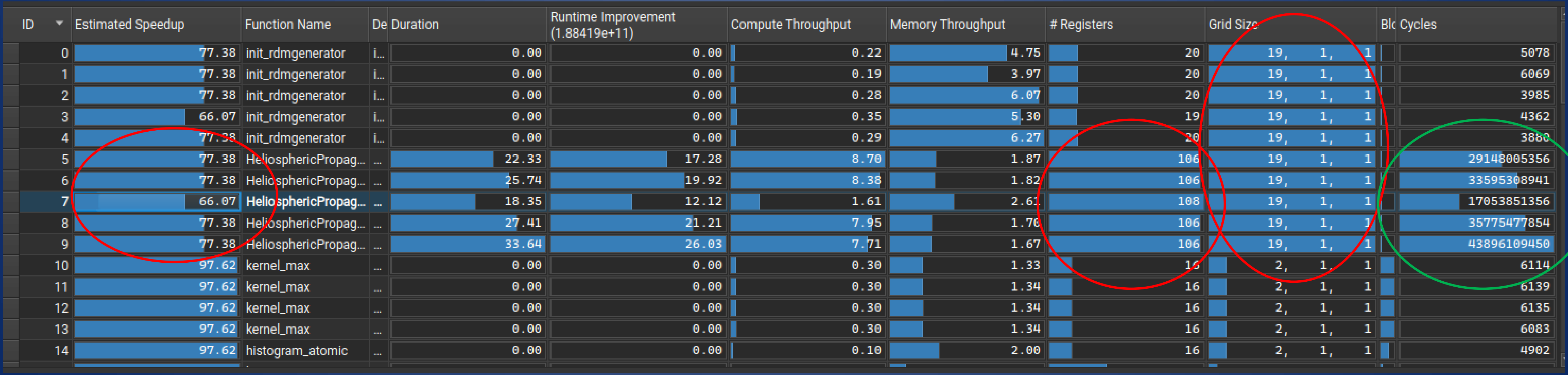
NSIGHT ANALYSIS (MULTI ENERGY A40)
GPU TIME USAGE



Time	Total Time	Instances	Avg	Med	Min	Max	StdDev	Category	Operation
100.0%	429.333 s	14	30.667 s	27.342 s	15.777 s	57.753 s	12.226 s	CUDA_KERNEL	HeliosphericPropagation(curandStatePhilox4_32_10 *, PropagationParameters_t, particle_t *, i
0.0%	64.642 μs	18	3.591 μs	4.144 μs	608 ns	4.513 μs	1.319 μs	MEMORY_OPER	[CUDA memcpy Host-to-Device]
0.0%	55.904 μs	42	1.331 μs	1.280 μs	1.216 μs	1.792 μs	150 ns	MEMORY_OPER	[CUDA memcpy Device-to-Host]
0.0%	39.616 μs	14	2.829 μs	2.816 μs	2.816 μs	2.848 μs	16 ns	CUDA_KERNEL	kernel_max(particle_t *, float *, unsigned long)
0.0%	28.127 μs	14	2.009 μs	2.016 μs	1.984 μs	2.048 μs	22 ns	CUDA_KERNEL	histogram_atomic(const particle_t *, float, float, int, unsigned long, float *, int *)
0.0%	21.408 μs	14	1.529 μs	1.536 μs	1.504 μs	1.536 μs	13 ns	CUDA_KERNEL	histogram_accum(const float *, int, int, float *)
0.0%	6.560 μs	14	468 ns	480 ns	448 ns	480 ns	15 ns	MEMORY_OPER	[CUDA memset]
0.0%	3.008 μs	1	3.008 μs	3.008 μs	3.008 μs	3.008 μs	0 ns	CUDA_KERNEL	init_rdmgenerator(curandStatePhilox4_32_10 *, unsigned long long)

- Execution time strongly dominated by the heliospheric propagation computation
- Max exit energy search and histogram building are negligible in the execution time
- Even memory set and transfer between host and device occupy less than 0.1%

NSIGHT COMPUTE ANALYSIS
A30 MULTI ENERGY

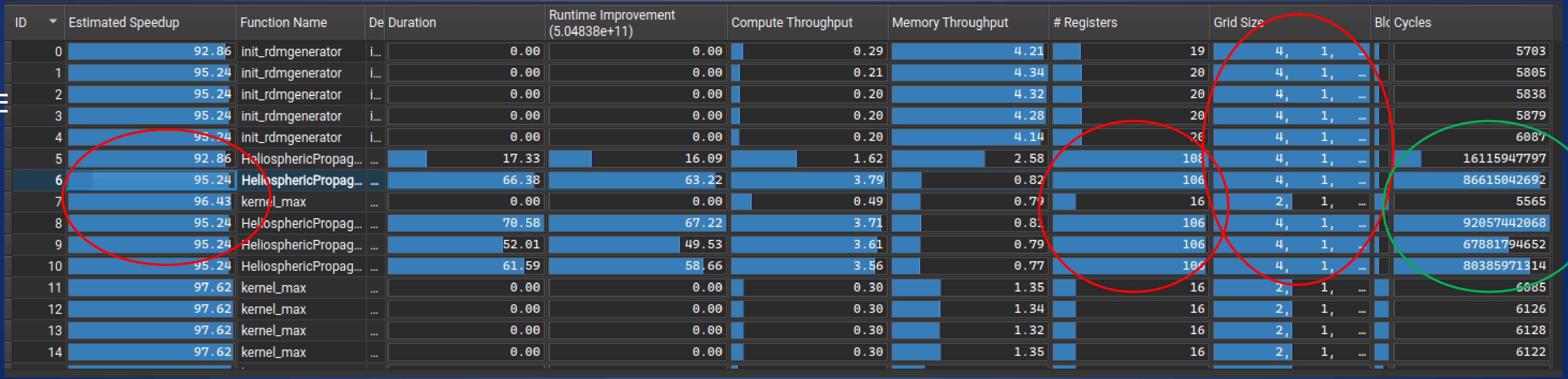


ID	Estimated Speedup	Function Name	De	Duration	Runtime Improvement (1.88419e+11)	Compute Throughput	Memory Throughput	# Registers	Grid Size	Blc	Cycles
0	77.38	init_rdmgenerator	i...	0.00	0.00	0.22	4.75	20	19, 1, 1		5078
1	77.38	init_rdmgenerator	i...	0.00	0.00	0.19	3.97	20	19, 1, 1		6069
2	77.38	init_rdmgenerator	i...	0.00	0.00	0.28	6.07	20	19, 1, 1		3985
3	66.07	init_rdmgenerator	i...	0.00	0.00	0.35	5.30	19	19, 1, 1		4362
4	77.38	init_rdmgenerator	i...	0.00	0.00	0.29	6.27	20	19, 1, 1		3880
5	77.38	HeliosphericPropag...	...	22.33	17.28	8.70	1.87	106	19, 1, 1		29148005356
6	77.38	HeliosphericPropag...	...	25.74	19.92	8.38	1.82	106	19, 1, 1		33595308941
7	66.07	HeliosphericPropag...	...	18.35	12.12	1.61	2.61	108	19, 1, 1		17053851356
8	77.38	HeliosphericPropag...	...	27.41	21.21	7.95	1.76	106	19, 1, 1		35775477854
9	77.38	HeliosphericPropag...	...	33.64	26.03	7.71	1.67	106	19, 1, 1		43896109450
10	97.62	kernel_max	...	0.00	0.00	0.30	1.33	16	2, 1, 1		6114
11	97.62	kernel_max	...	0.00	0.00	0.30	1.34	16	2, 1, 1		6139
12	97.62	kernel_max	...	0.00	0.00	0.30	1.34	16	2, 1, 1		6135
13	97.62	kernel_max	...	0.00	0.00	0.30	1.34	16	2, 1, 1		6083
14	97.62	histogram_atomic	...	0.00	0.00	0.10	2.00	16	2, 1, 1		4902

- The grid for this launch is configured to execute only 19 blocks, which is less than the GPU's 56 multiprocessors, underutilizing some multiprocessors
- Number of threads per block not a multiple of the warp dimension (rounded in the next version)
- Between 66 - 77% improvement of the most time comsuming function

NSIGHT COMPUTE ANALYSIS
A40 MULTI ENERGY

- E



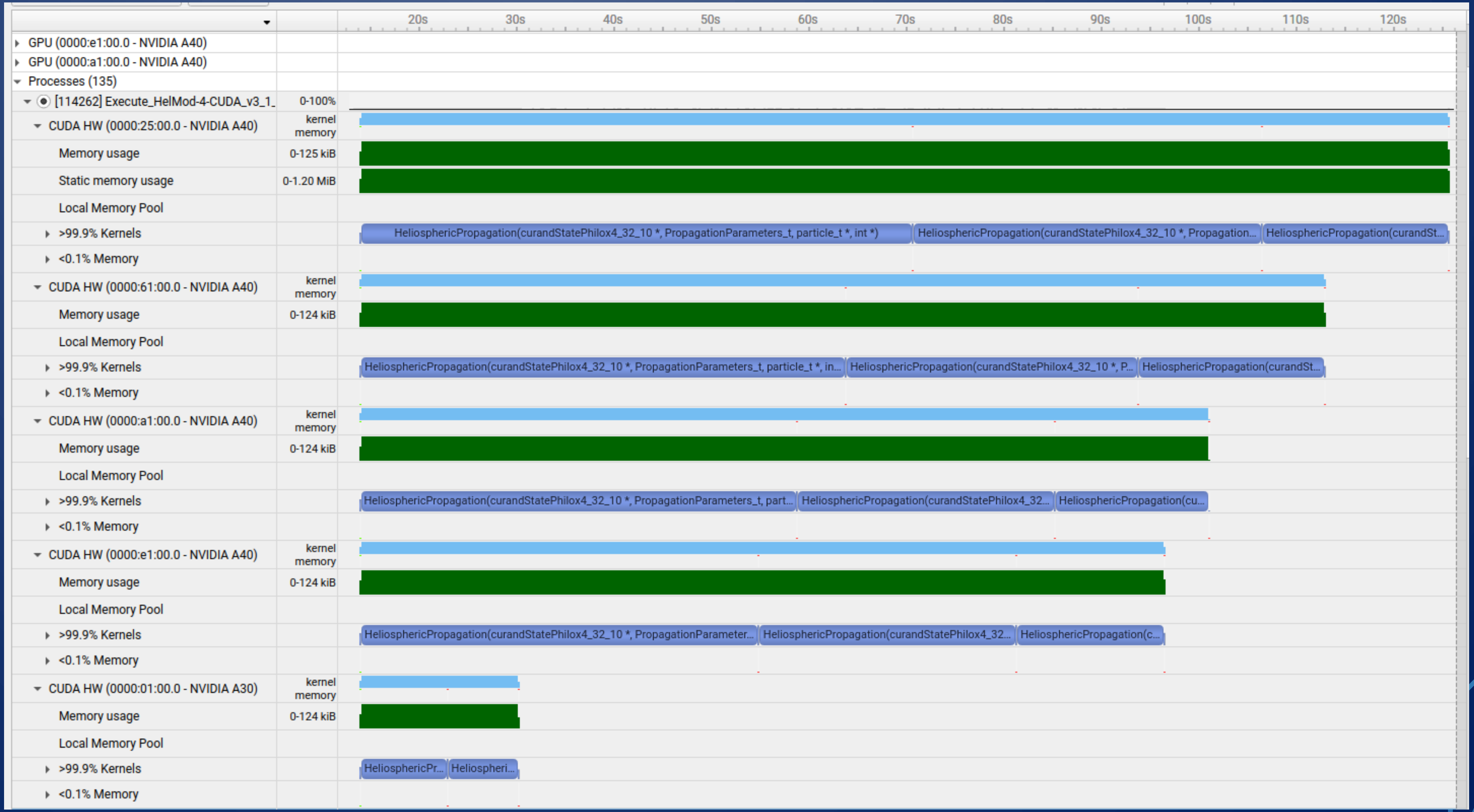
ID	Estimated Speedup	Function Name	De	Duration	Runtime Improvement (5.04838e+11)	Compute Throughput	Memory Throughput	# Registers	Grid Size	Blc	Cycles
0	92.86	init_rdmgenerator	i...	0.00	0.00	0.29	4.21	19	4, 1, ...		5703
1	95.24	init_rdmgenerator	i...	0.00	0.00	0.21	4.34	20	4, 1, ...		5805
2	95.24	init_rdmgenerator	i...	0.00	0.00	0.20	4.32	20	4, 1, ...		5838
3	95.24	init_rdmgenerator	i...	0.00	0.00	0.20	4.28	20	4, 1, ...		5879
4	95.24	init_rdmgenerator	i...	0.00	0.00	0.20	4.14	20	4, 1, ...		6087
5	92.86	HeliosphericPropag...	...	17.33	16.09	1.62	2.58	108	4, 1, ...		16115947797
6	95.24	HeliosphericPropag...	...	66.38	63.22	3.79	0.82	106	4, 1, ...		86615042692
7	96.43	kernel_max	...	0.00	0.00	0.49	0.79	16	2, 1, ...		5565
8	95.24	HeliosphericPropag...	...	70.58	67.22	3.71	0.81	106	4, 1, ...		92057442068
9	95.24	HeliosphericPropag...	...	52.01	49.53	3.61	0.79	106	4, 1, ...		67881794652
10	95.24	HeliosphericPropag...	...	61.59	58.66	3.56	0.77	106	4, 1, ...		80385971314
11	97.62	kernel_max	...	0.00	0.00	0.30	1.35	16	2, 1, ...		6085
12	97.62	kernel_max	...	0.00	0.00	0.30	1.34	16	2, 1, ...		6126
13	97.62	kernel_max	...	0.00	0.00	0.30	1.32	16	2, 1, ...		6128
14	97.62	kernel_max	...	0.00	0.00	0.30	1.35	16	2, 1, ...		6122

- The grid for this launch is configured to execute only 4 blocks, which is less than the GPU's 84 multiprocessors, underutilizing some multiprocessors
- Number of threads per block not a multiple of the warp dimension (rounded in the next version)
- Between 92 – 95% improvement of the most time comsuming function

Stronger effect of the warp per block on the A40 boards

NSIGHT ANALYSIS (MULTI ENERGY)
MULTI GPU

- Energy bins with different avarage propagation time are not equally distributed between the GPU in the cluster
- Evident faster execution of the propagation computation in the A30 boards


20s
30s
40s
50s
60s
70s
80s
90s
100s
110s
120s
GPU (0000:e1:00.0 - NVIDIA A40)
GPU (0000:a1:00.0 - NVIDIA A40)
Processes (135)
[114262] Execute_HelMod-4-CUDA_v3_1_
0-100%
CUDA HW (0000:25:00.0 - NVIDIA A40)
kernel memory
Memory usage
0-125 kiB
Static memory usage
0-1.20 MiB
Local Memory Pool
>99.9% Kernels
HeliosphericPropagation(curandStatePhilox4_32_10 *, PropagationParameters_t, particle_t *, int *)
HeliosphericPropagation(curandStatePhilox4_32_10 *, Propagation...
HeliosphericPropagation(curandSt...
<0.1% Memory
CUDA HW (0000:61:00.0 - NVIDIA A40)
kernel memory
Memory usage
0-124 kiB
Local Memory Pool
>99.9% Kernels
HeliosphericPropagation(curandStatePhilox4_32_10 *, PropagationParameters_t, particle_t *, in...
HeliosphericPropagation(curandStatePhilox4_32_10 *, P...
HeliosphericPropagation(curandSt...
<0.1% Memory
CUDA HW (0000:a1:00.0 - NVIDIA A40)
kernel memory
Memory usage
0-124 kiB
Local Memory Pool
>99.9% Kernels
HeliosphericPropagation(curandStatePhilox4_32_10 *, PropagationParameters_t, part...
HeliosphericPropagation(curandStatePhilox4_32...
HeliosphericPropagation(cu...
<0.1% Memory
CUDA HW (0000:e1:00.0 - NVIDIA A40)
kernel memory
Memory usage
0-124 kiB
Local Memory Pool
>99.9% Kernels
HeliosphericPropagation(curandStatePhilox4_32_10 *, PropagationParameter...
HeliosphericPropagation(curandStatePhilox4_32...
HeliosphericPropagation(c...
<0.1% Memory
CUDA HW (0000:01:00.0 - NVIDIA A30)
kernel memory
Memory usage
0-124 kiB
Local Memory Pool
>99.9% Kernels
HeliosphericPr...
Heliospheri...
<0.1% Memory

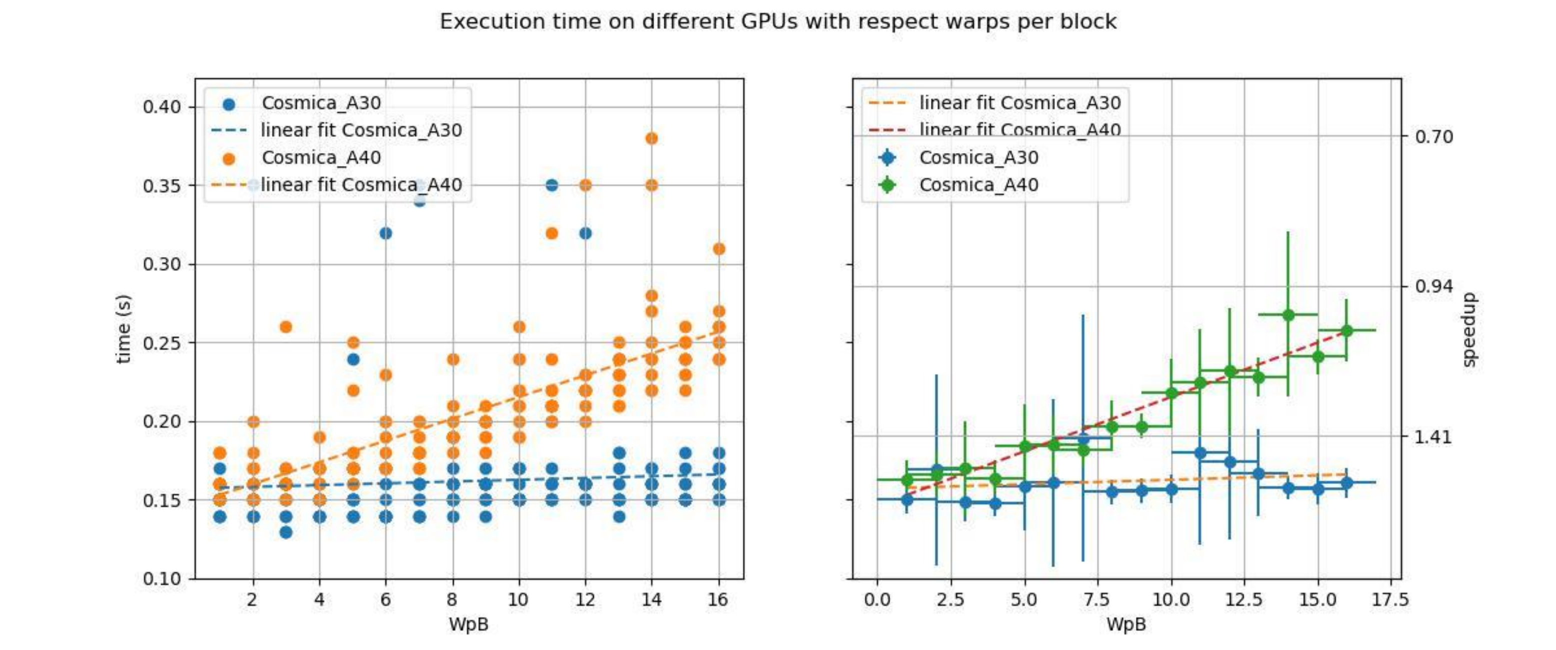
1ENERGY BIN SINGLE GPU COSMICA V1

(SAME SIMULATION CONFIGURATION)

- EMin = 7.454
- Emax = 9.814
- N Output binning: 33

- Time to back-propagate particles:
  - Init : 0.08 ms
  - **propagation phase : 93.66 ms**
  - Find Max : 47.48 ms
  - Binning : 0.17 ms
- Time to Set Memory: 30.0 ms
- Time to create Rnd: 0.1 ms
- **Time to execute : 171.7 ms**

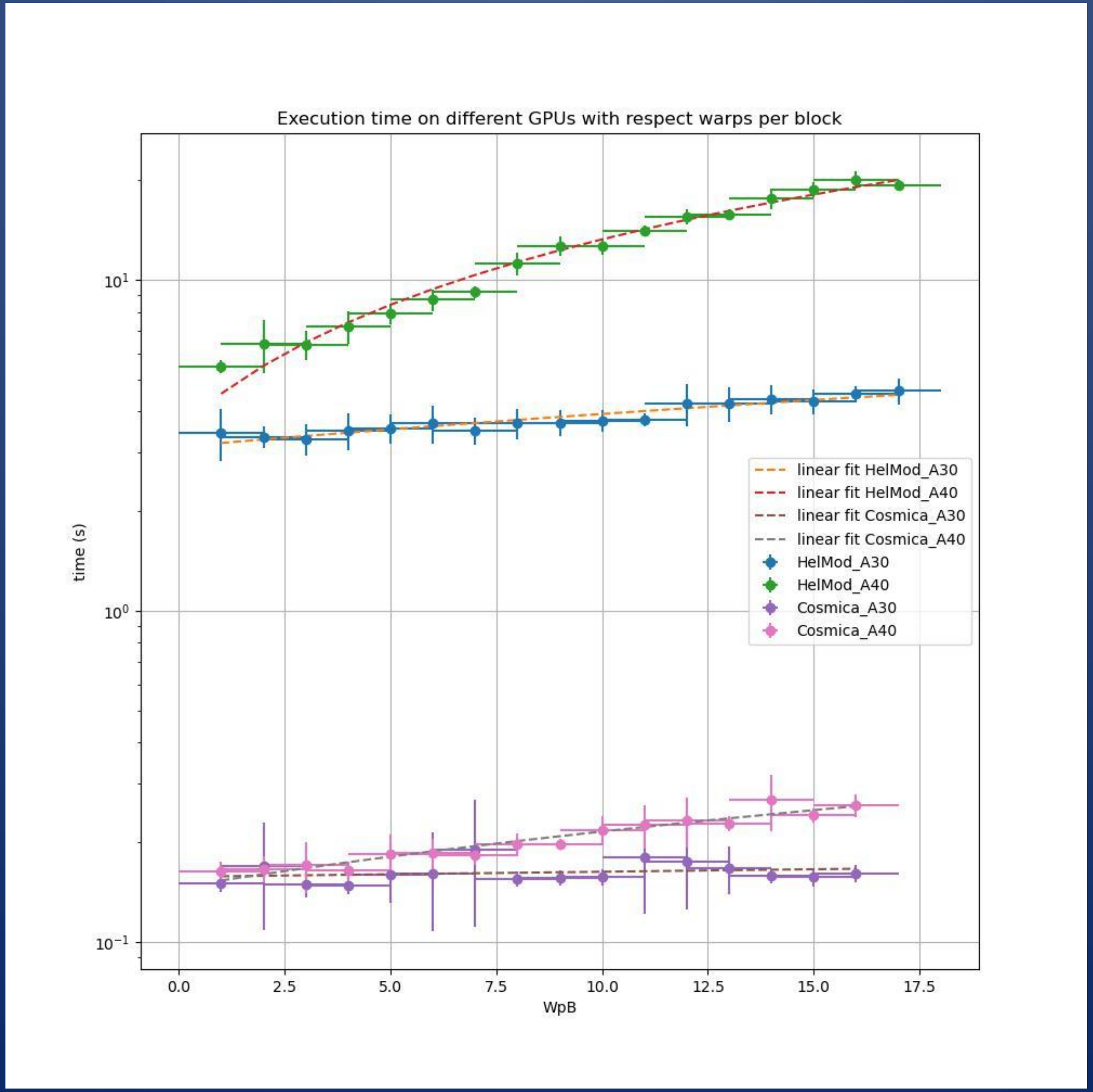
The code was executed
10 times for each WpB value



Execution time on different GPUs with respect warps per block

SPEED COMPARISON WITH COSMICA V1

(SAME SIMULATION CONFIGURATION)



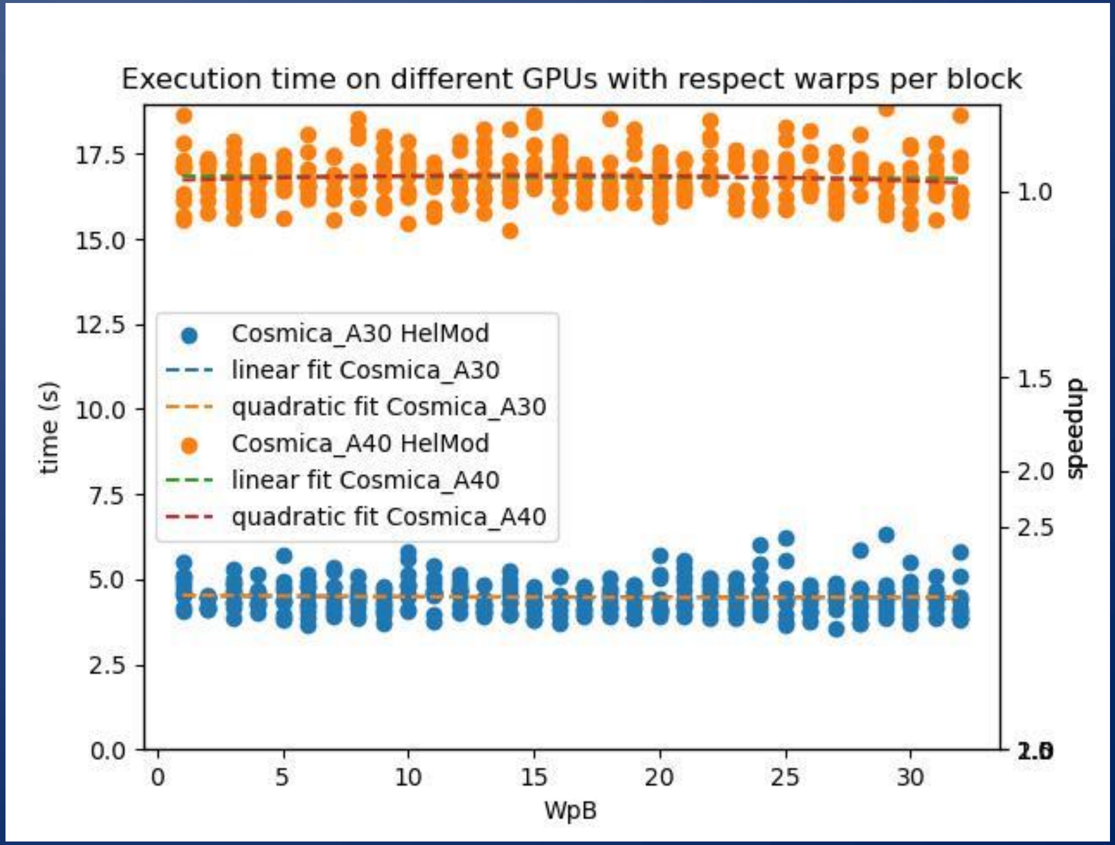
1ENERGY BIN SINGLE GPU COSMICA STABLE

(SAME SIMULATION CONFIGURATION)

- EMin = 7.454
- Emax = 9.814
- N Output binning: 33

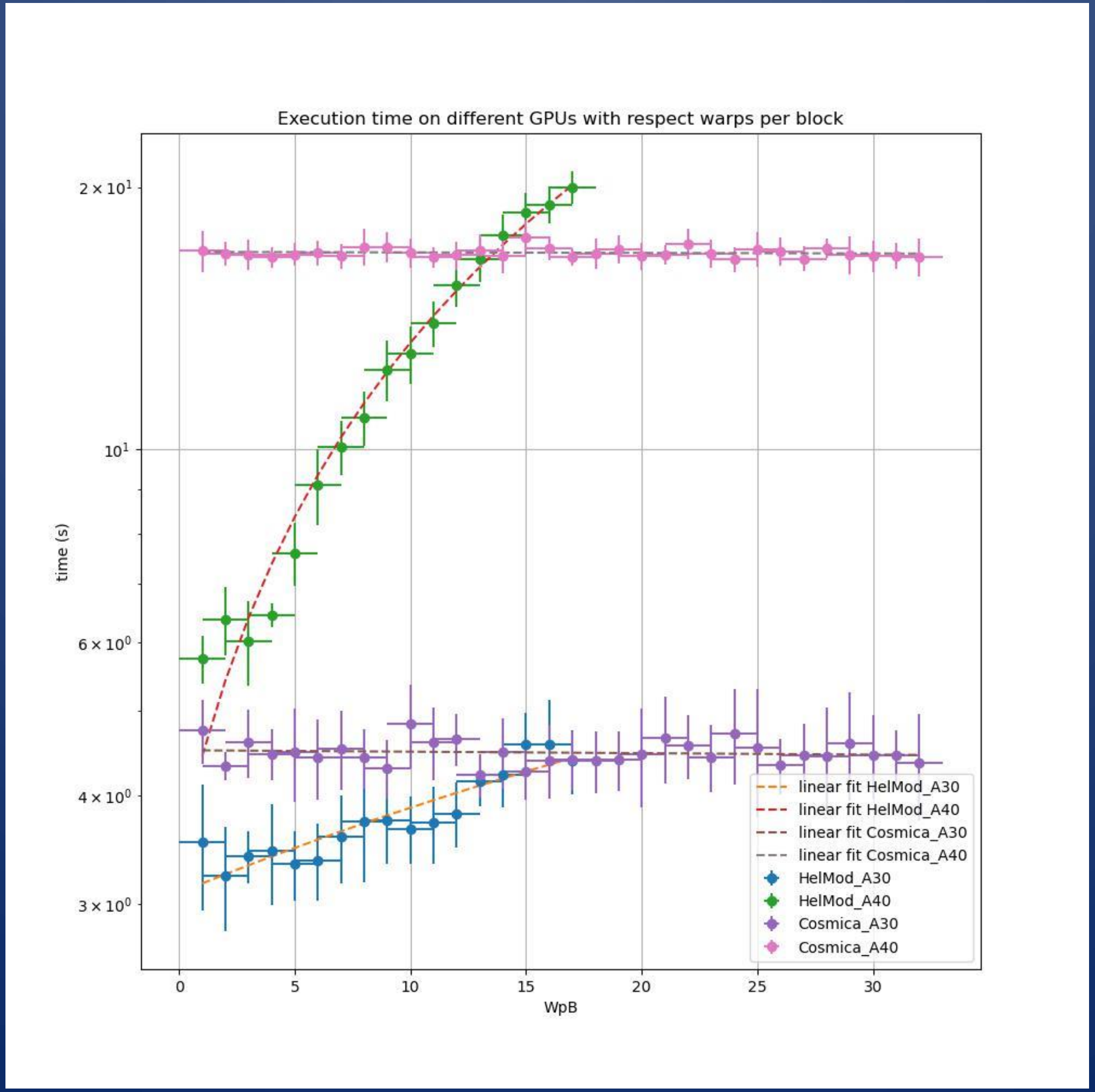
- Time to back-propagate particles:
  - Init : 0.09 ms
  - **propagation phase : 5369.13 ms**
  - Find Max : 0.15 ms
  - Binning : 0.47 ms
- Time to Set Memory: 13.6 ms
- Time to create Rnd: 0.1 ms
- **Time to execute : 5479.8 ms**

The code was executed
10 times for each WpB value



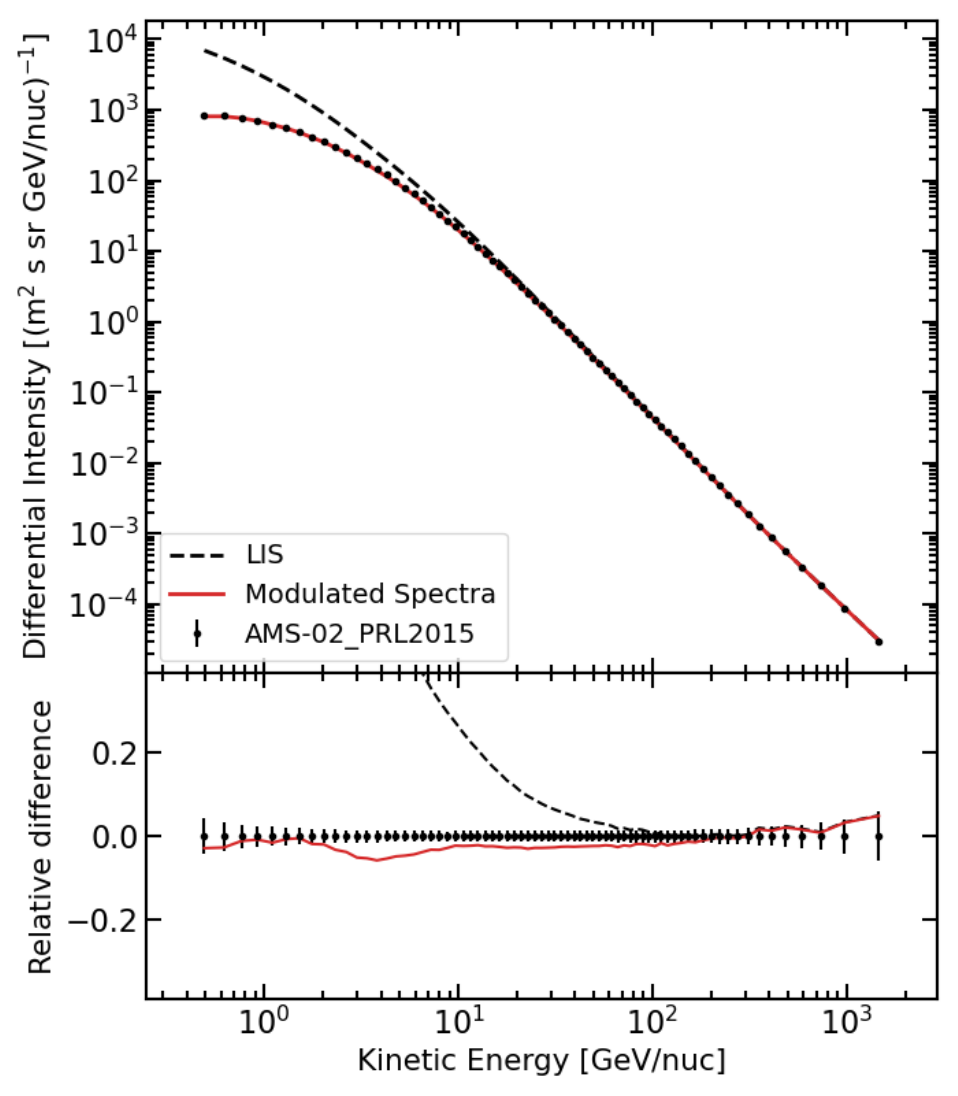
SPEED COMPARISON WITH COSMICA STABLE

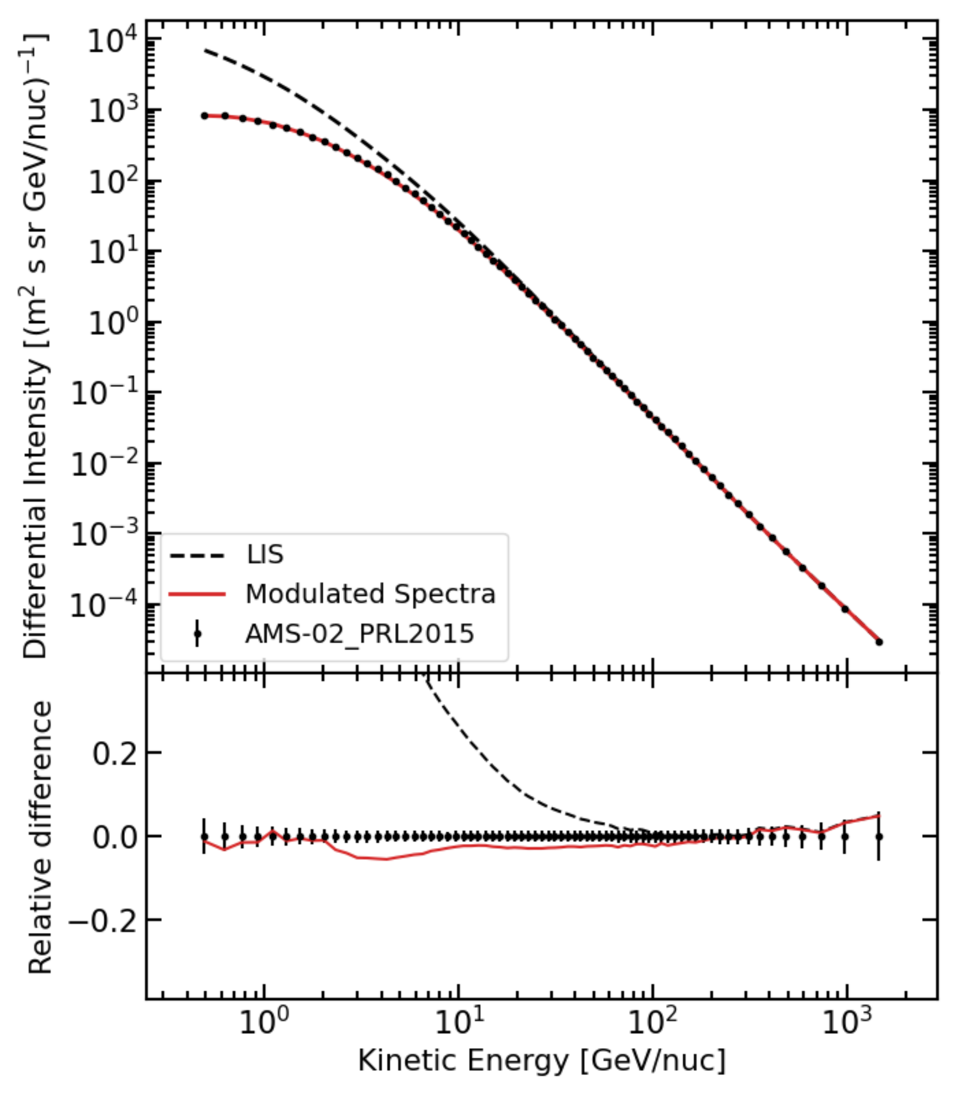
(SAME SIMULATION CONFIGURATION)



PROPAGATION OUTPUT TEST COSMICA SYABLE

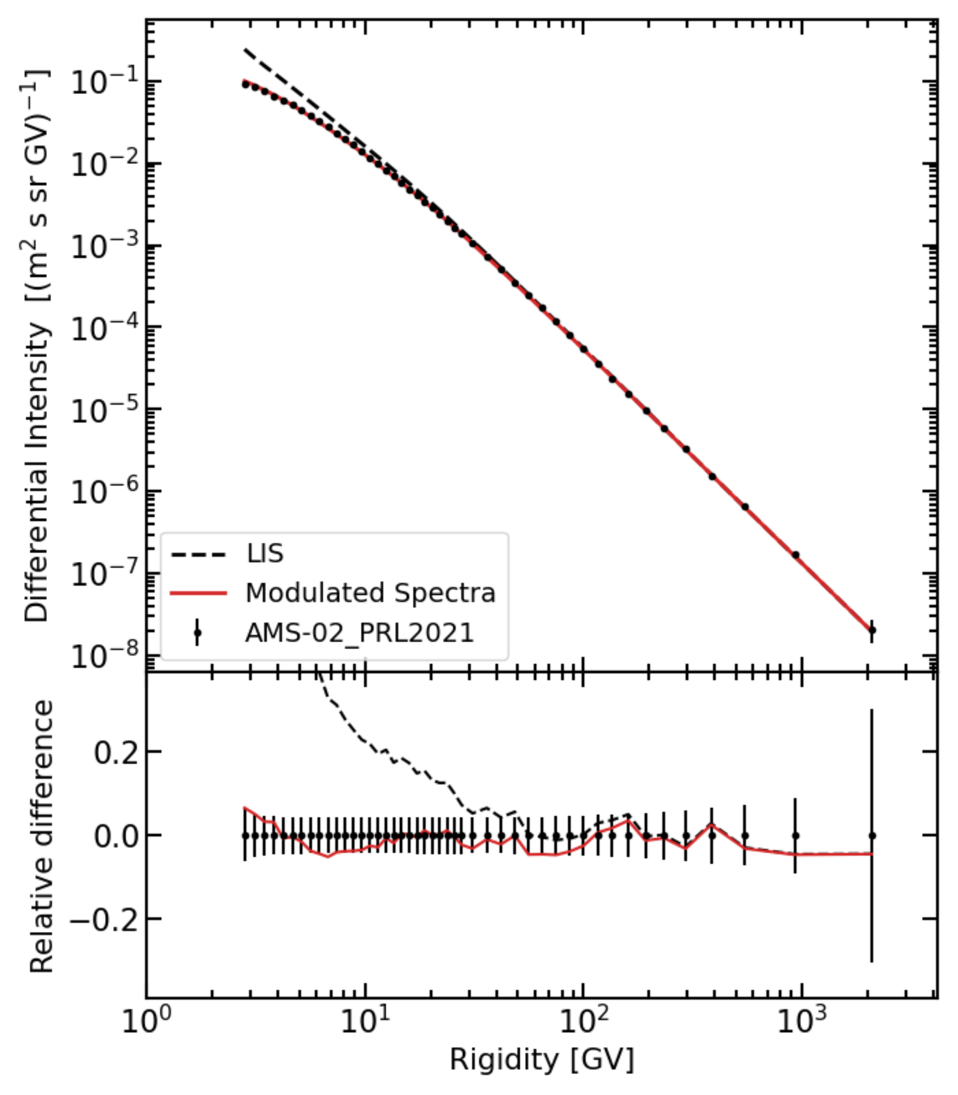
PROTON FLUX (LEFT : COSMICA, RIGHT : OLD CODE)

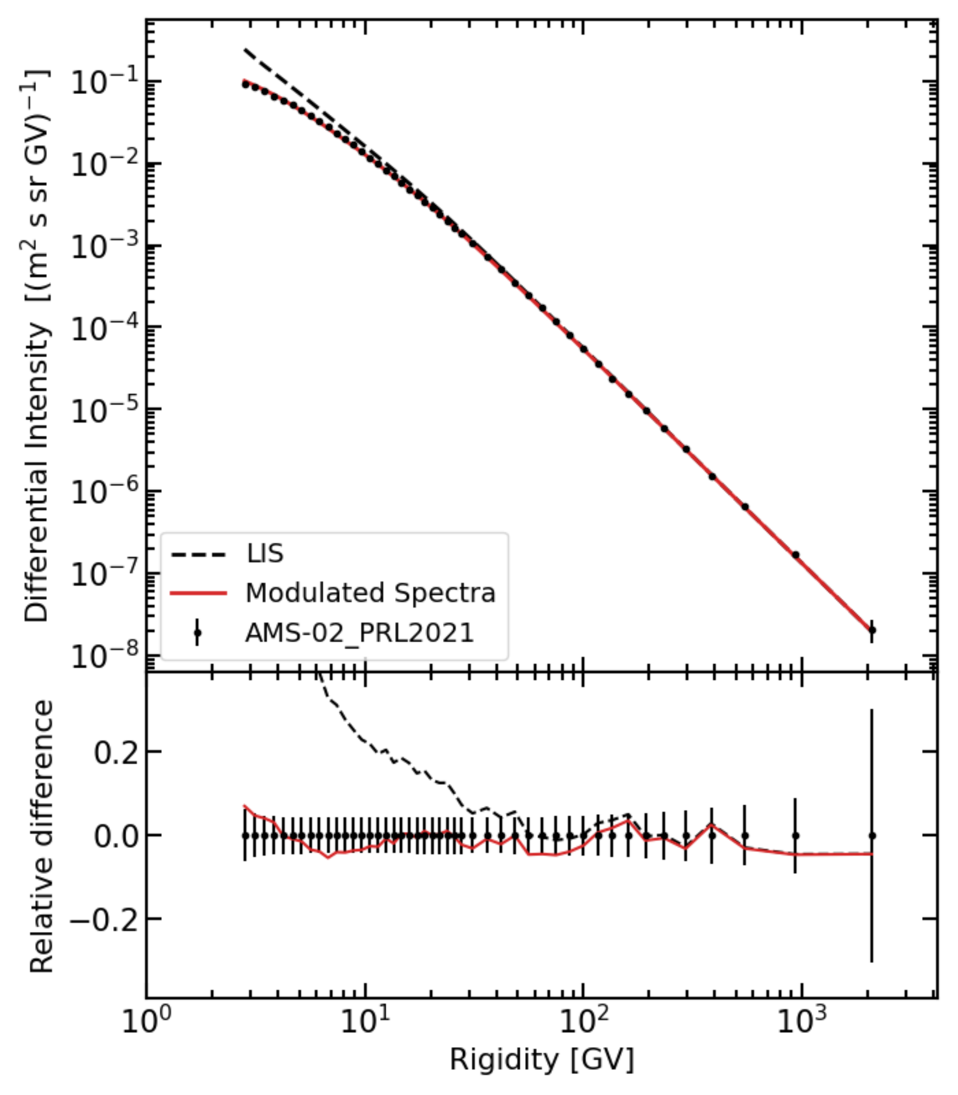




PROPAGATION OUTPUT TEST COSMICA SYABLE

IRON FLUX (LEFT : COSMICA, RIGHT : OLD CODE)


Differential Intensity [(m^2 s sr GV)$^{-1}$]
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
LIS
Modulated Spectra
AMS-02_PRL2021
Relative difference
10^0
10^1
10^2
10^3
Rigidity [GV]


Differential Intensity [(m^2 s sr GV)$^{-1}$]
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
LIS
Modulated Spectra
AMS-02_PRL2021
Relative difference
10^0
10^1
10^2
10^3
Rigidity [GV]