GPU-CPU Memory Transfer Data - Group 2

- Steven John A. Pascaran
- Charlyne Arajoy Carabeo

A. Unified memory introduced in CUDA 6

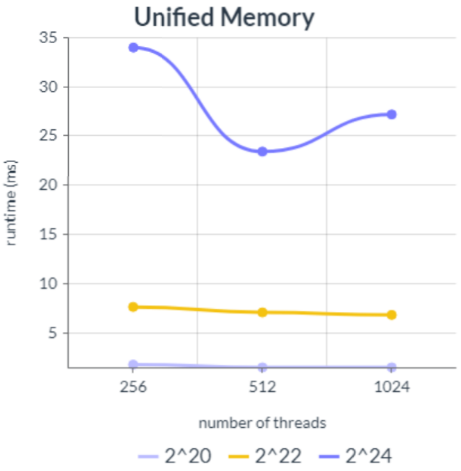
Compute method execution time	2^{20}	2^{22}	2^{24}
256	1.7767ms	7.6427ms	33.965ms
512	1.4986ms	7.0767ms	23.415ms
1024	1.4987ms	6.8222ms	27.188ms

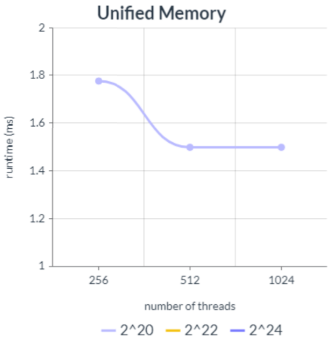
Host to Device execution time (Total Size: 64,000KB)	2^{20}	2^{22}	2^{24}
256	0.008416ms	0.008351ms	0.00816ms
512	0.007936ms	0.01264ms	0.007968ms
1024	0.007872ms	0.008128ms	0.008032ms

Device to Host execution time (Total Size: 128,000KB)	2^{20}	2^{22}	2^{24}
256	0.017088ms	0.01756ms	0.01716
512	0.01712ms	0.017088ms	0.017024ms
1024	0.017312ms	0.017248ms	0.017088ms

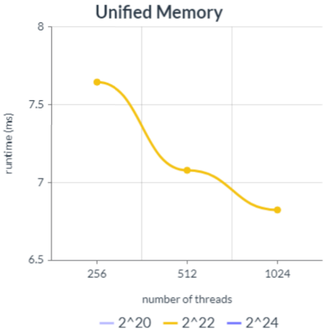
GPU Page faults execution time	2^{20}	2^{22}	2^{24}
256	1.9634ms	8.1308ms	37.871ms
512	1.573657ms	9.788310ms	21.02570ms
1024	1.683414ms	6.223072ms	24.30412ms

Compute Method Graphs:

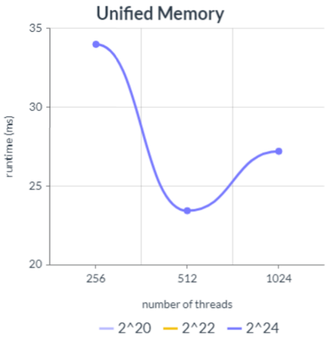




Compute Method 2^{20}



Compute Method 2^{22}



Compute Method 2^{24}

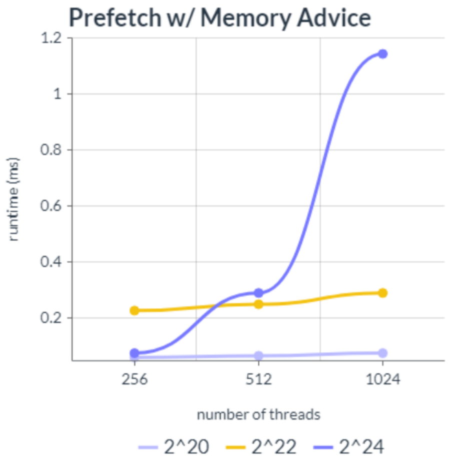
B. Prefetching of data with memory advice

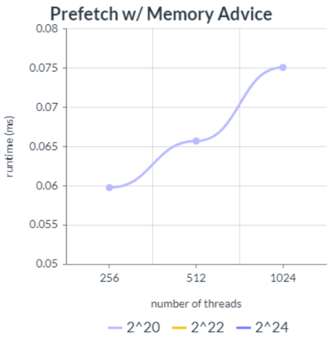
Compute method execution time	2^{20}	2^{22}	2^{24}
256	0.059743ms	0.22665ms	0.8939ms
512	0.065663ms	0.24931ms	0.98929ms
1024	0.075038ms	0.28963ms	1.1418ms

Host to Device execution time	2^{20} (Total Size: 4MB)	2^{22} (Total Size: 16MB)	2^{24} (Total Size: 64MB)
256	0.350842ms	1.425ms	5.602ms
512	0.347134ms	1.396441ms	5.600768ms
1024	0.347902ms	1.404824ms	5.580096ms

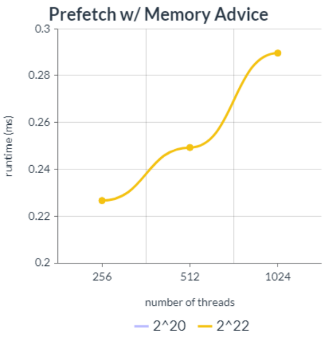
Device to Host execution time	2^{20} (Total Size: 4MB)	2^{22} (Total Size: 16MB)	2^{24} (Total Size: 64MB)
256	0.322074ms	1.288ms	5.153ms
512	0.322078ms	1.288058ms	5.162405ms
1024	0.335166ms	1.288282ms	5.152165ms

Compute Method Graphs:

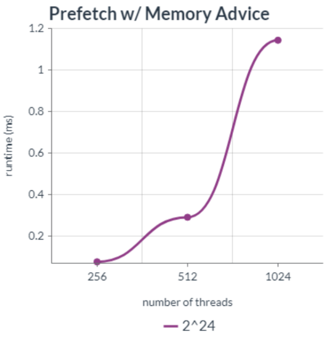




Compute Method 2^{20}



Compute Method 2^{22}



Compute Method 2^{24}

C. Data transfer or initialization as a CUDA kernel

Compute method execution time	2^{20}	2^{22}	2^{24}
256	1.4100ms	6.6046ms	48.332ms
512	1.5266ms	5.9278ms	29.895ms
1024	1.4160ms	7.5558ms	26.290ms

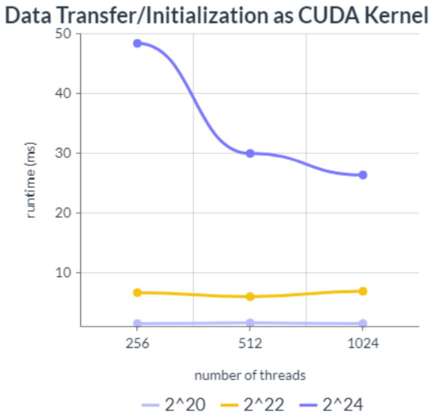
Transfer method execution time	2^{20}	2^{22}	2^{24}
256	0.40287ms	0.30399ms	0.31340ms
512	0.33411ms	0.38109ms	0.37981ms
1024	0.35661ms	0.34896ms	0.30399ms

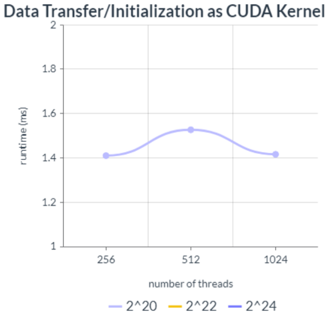
Device to Host execution time	2^{20}	2^{22}	2^{24}
256	0.017663ms	0.017535ms	0.017344ms
512	0.016928ms	0.01712ms	0.017056ms
1024	0.017088ms	0.017888ms	0.017024ms

GPU Page faults execution time	2^{20}	2^{22}	2^{24}
256	1.733ms	6.622ms	47.571ms

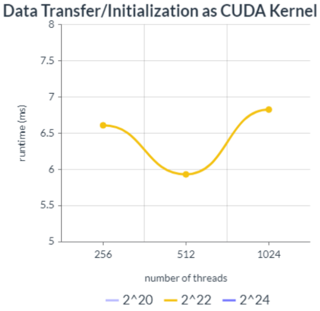
512	1.509335ms	7.666420ms	29.57955ms
1024	1.636470ms	7.028473ms	34.54656ms

Compute Method Graphs:

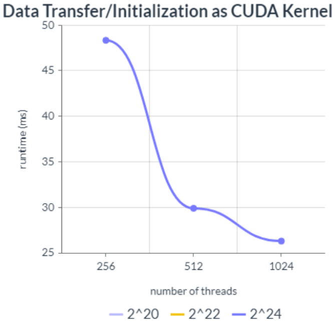




Compute Method 2^{20}



Compute Method 2^{22}



Compute Method 2^{24}

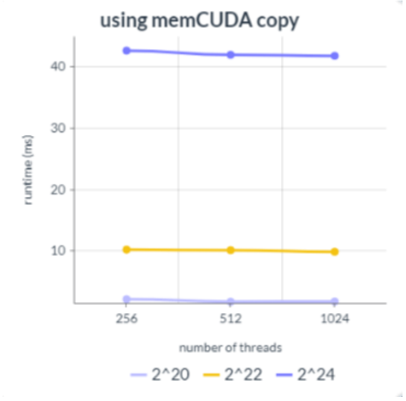
D. Old method of transferring data between CPU and memory (memCUDA copy)

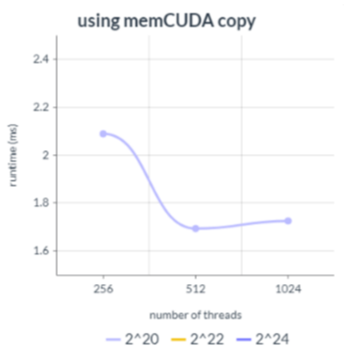
Compute method execution time	2^{20}	2^{22}	2^{24}
256	0.073214ms	0.27875ms	1.1004ms
512	0.083615ms	0.32256ms	1.2832ms
1024	0.095392ms	0.36758ms	1.4496ms

CUDA memCopy	2^{20}	2^{22}	2^{24}

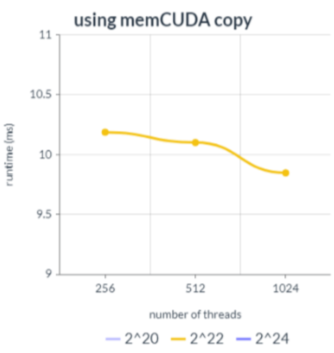
DtoH execution time			
256	2.0885ms	10.185ms	42.671ms
512	1.6930ms	10.099ms	41.991ms
1024	1.7249ms	9.8457ms	41.817ms

CUDAMemCpy Method Graphs:

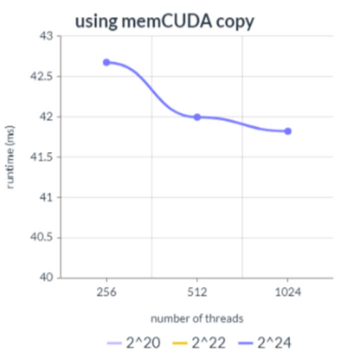




CUDAMemCpy 2^{20}



CUDAMemCpy 2^{22}



CUDAMemCpy 2^{24}