Zero copy memory

Zero copy memory

Zero-copy memory is pinned memory that is mapped into the device address space. So that both device and host have direct access to that memory

Advantage of zero copy memory

- Leveraging host memory when there is insufficient device memory
- Avoiding explicit data transfer between the host and device
- Improving PCIe transfer rates

Zero copy memory

```
cudaError_t cudaHostAlloc
      (void ** pHost, size_t  count, unsigned int
flags);

 cudaFreeHost ( void ** pHost)
```

cudaHostAllocDefault

Same as pinned memory

cudaHostAllocPortable

pinned memory that can be used by all CUDA contexts

cudaHostAllocWriteCombined

written by the host and read by the device

cudaHostAllocMapped

host memory that is mapped into the device

```
cudaError_t cudaHostGetDevicePointer
        (void ** pDevice, void * pHost, unsigned
int flags);
```

SIZE	DEVICE MEMORY (ELAPSED TIME)	ZERO-COPY MEMORY (ELAPSED TIME)	SLOWDOWN
1 K	1.5820 us	2.9150 us	1.84
4 K	1.6640 us	3.7900 us	2.28
16 K	1.6740 us	7.4570 us	4.45
64 K	2.3910 us	22.586 us	9.45
256 K	7.2890 us	82.733 us	11.35
1 M	28.267 us	321.57 us	11.38
4 M	104.17 us	1.2741 ms	12.23
16 M	408.03 us	5.0903 ms	12.47
64 M	1.6276 ms	20.347 ms	12.50

Warning

when using zero-copy memory to share data between the host and device, you must synchronize memory accesses across the host and device