

Array add is completely memory bound. All We're doing is timing a local array to read/write (and it's so small: completely in-cache) versus read from memory, a copy across the PCIe bus, then an array read/write in device memory, and then a copy back across the PCIe bus and write to memory again. (Not to mention the queuing up of a job and launching of a task on the GPU - this alone probably takes more than adding elements). Of course, it will be slower using the GPU to do this - We have much more additional memory movement and such a tiny bit of work to do.