Commodity Graphics Cards for Quantum Monte Carlo and Teraflop Scale Scientific Computing

Amos Anderson, Dan Fisher, Albert Cervantes, Mario Blanco, Peter Schröder, William A Goddard
California Institute of Technology

Abstract
Graphical Processing Units (GPUs) commonly referred to as "graphics cards" are finding increasing usefulness in general purpose calculations as they have become programmable. GPUs have far more raw processing power than CPUs (approximately 50 GFLOPS versus less than 10 GFLOPS) and thus, if one can translate the calculation into the expressions of graphics and the SIMD paradigm (and not demand double precision), one can expect very large performance boosts. This idea is further boosted by the multibillion dollar gaming market, which pushes graphics hardware for higher performance gains per generation than the Moore's Law for CPUs.

Current estimates are that GPUs will double their processing power every 6 to 12 months, as opposed to every 18 months. Simple timing results based on the Quantum Monte Carlo (QMC) algorithm have produced encouraging results, and proof of concept QMC results have been completed.

We are working to piece together a supercomputer composed of these graphics cards on which to run our QMC program. A simple cluster with 2 GPUs for all 100 CPUs would have calculation power on the teraflop scale. Ignoring the specialized nature of such a computer, this would "rival" Top 500 machines at almost negligible cost per flop. Our estimate is that using a the configuration described, we should be able to reach 4 TFLOPS per $100k. Recent developments in our research group have enabled our QMC software to automatically load balance on even heterogeneous networks. This means that it will be a fairly simple task to add CPU/GPU nodes as funds become available, using the most cost effective GPUs available at a given time of purchase. Furthermore, upgrading the machine would be trivially simple, involving simply the replacement (or potentially addition) of GPUs.