D-Wave
The Quantum Computing Company™

Computational Power Consumption and Speedup

WHITEPAPER

Summary

Power consumption for computation is a serious and growing issue for the world. We rely more and more on computing in everything we do as we try to satisfy our ever-increasing thirst for mobile computing, automation, machine intelligence, cloud computing, and increasingly powerful supercomputers. Highly specialized coprocessors such as D-Wave's quantum processing units (QPUs) show promise in significantly increasing the power efficiency of computing. In a recent study, D-Wave's 2000-qubit system was shown to be up to 100 times more energy efficient than highly specialized algorithms on state-of-the-art classical computing servers when considering pure computation time, suggesting immediate relevance to large-scale energy efficient computing.

Much academic discussion of D-Wave™ quantum processing units (QPUs) has centered around various definitions of quantum speedup[1,2] from a theoretical computer science point of view. In theoretical computer science, a constant prefactor (e.g., being 1,000,000 times faster than an alternative) is viewed as offering no scaling advantage because the computational advantage remains the same as the problem size increases. While interesting for abstract comparisons of quantum versus classical, such analyses ignore the fact that, in the real world, a 1,000,000 times constant prefactor improvement offers an enormous advantage in application to practical problems. In this paper we examine observed constant prefactors in relation to the serious issue of power consumption that is of immediate relevance to high-performance computing (HPC) and hyperscale cloud computing as employed by Google, Amazon, Microsoft, and others.

In the full study[3] on which this document is based, D-Wave QPUs were evaluated on a class of synthetic inputs that are both challenging and relevant to real-world applications. In absolute terms, the 2000-qubit D-Wave QPU outperformed competitive software solvers on classical processors by factors of roughly 1000 to 10,000 in pure computation time. We discuss two software solvers for classical processors considered strong competition for D-Wave QPUs. The first is quantum Monte Carlo (QMC), a classical approximation of quantum annealing. QMC is used for molecular simulation and can also be used effectively for optimization applications. The second is the Hamze-de Freitas-Selby (HFS) algorithm. HFS is recognized as the most competitive classical algorithm for current-generation D-Wave QPUs, but is not expected to remain competitive against D-Wave QPUs in development with greater numbers of couplings between qubits. The other algorithms considered in the full paper failed to perform well on the harder inputs.

Energy consumption and exascale computing The US Department of Energy's Exascale Computing Initiative has the goal of deploying an exascale supercomputer—one capable of 1 exaflops, or 10^{18} floating-point operations per second (FLOPS)—that draws only 20 to 30 MW of power.[4] This translates to an efficiency of up

[1] T. F. Rønnow et al., Science 345(6195):420–424, 2014.
[2] S. Mandrà et al., Phys. Rev. A 94(022337), 2016.
[3] J. King et al., D-Wave Technical Report 14-1003A-C, 2017.
[4] *Preliminary conceptual design for an exascale computing initiative*, http://science.energy.gov/~/media/ascr/ascac/pdf/meetings/20141121/Exascale_Preliminary_Plan_V11_sb03c.pdf

to 50 gigaflops per watt. By contrast, the world's most powerful supercomputer as of 2017—China's Sunway TaihuLight—achieves an efficiency of 2.2 gigaflops per watt including cooling power, a factor of 20 from the exascale goal.

Including the cooling, TaihuLight requires 42 MW of power, only slightly less than the 57 MW generated by the average hydroelectric facility in the US. Extrapolating based on its efficiency, TaihuLight would require 450 MW of power if extended to exascale (1000 petaflops). Using the average price for industrial power in the US, the operating costs would be a staggering $270 million[5] per year. In contrast, the power draw of a D-Wave system is only 16 kW resulting in an electricity cost of $10,000 per year. Due to our use of superconducting processors, this is not expected to increase significantly as we scale to larger systems.

Improving efficiency with specialized coprocessors

More efficient computation is needed and can be achieved using specialized coprocessors. As an example, we consider the 2017 state-of-the-art NVIDIA DGX-1, a highly optimized graphics processing unit (GPU) server costing $129,000. An NVIDIA DGX-1 server is capable of 170 teraflops at an efficiency of 53 gigaflops per watt. While certain tasks can benefit from such a specialized processor, not all algorithms can be parallelized and more efficient computation comes at the expense of generality; for example, HFS cannot be implemented efficiently on GPU platforms like the NVIDIA DGX-1.

If we go to an even more specialized coprocessor, the D-Wave QPU, the benefits in terms of energy efficiency can be massive. While D-Wave QPUs do not perform floating-point operations, we can compare efficiency using equivalent problem solving performance. In the full study, the D-Wave QPU solves problems 10,000 times faster than QMC run on an NVIDIA GTX 1080, a high-end consumer graphics card. Extrapolating based solely on FLOP rate and computation time, this would be equivalent to roughly 500 NVIDIA DGX-1 servers.[6] The power draw of the D-Wave system is nominally 16 kW whereas 500 NVIDIA DGX-1 servers draw 1.6 MW; i.e., the state-of-the-art GPUs would consume 100 times more power than the D-Wave system (see Figure 1).

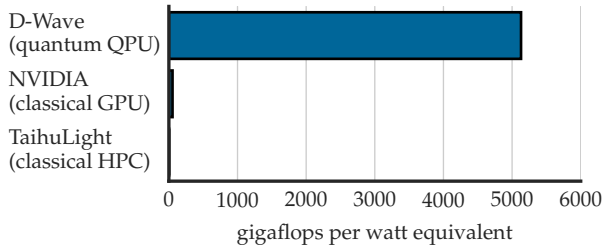

Figure 1: Power efficiency of computing systems measured in equivalent computational performance per watt for D-Wave quantum computing versus classical simulation.

In the full study, HFS was more energy efficient than QMC because a) it was faster than QMC, and b) it was run on a single CPU core drawing 20 W rather than on a GPU drawing 180 W. Considering pure annealing time (i.e., computation time), HFS is roughly on par with the 2000-qubit D-Wave QPU in terms of ground state throughput per watt. However, we have noted that HFS is not expected to be effective when applied to denser topologies that are in development at D-Wave.

Almost all of the power drawn by D-Wave systems is used by the cryogenic refrigeration. This has remained constant since the introduction of the first generation of D-Wave system in 2011 and is expected to stay constant as the computational power continues to grow dramatically with successive generations of QPU. As a result, the computing power per watt is expected to increase much more rapidly for QPUs than for classical systems.

Conclusions The future of computing requires increased energy efficiency, which in turn will depend on increased reliance on specialized coprocessors including GPUs as we near the end of Moore's law. D-Wave QPUs have demonstrated significant improvements over classical alternatives in terms of computing power per watt due to their ability to leverage quantum computational resources and their use of superconducting electronics, thereby showing great promise as components in future hybrid classical/quantum supercomputers.

[5] All dollar amounts are in USD.

[6] This back-of-the-envelope extrapolation unfairly benefits classical solvers for two reasons. First, it is valid only in the case where we have a large number of independent jobs to run in parallel. In practice, the ability to parallelize across devices is limited by the number of concurrent jobs that can be run since all of the algorithms we consider are dominated by sequential loops. Second, our calculation ignores communication time between devices.