LA-UR-16-22253

Title: An Early Quantum Computing Proposal

Author(s): Lee, Stephen Russell; Alexander, Francis Joseph; Barros, Kipton Marcos; Daniels, Marcus G.; Gattiker, James R.; Hamada, Michael Scott; Howse, James Walter; Loncaric, Josip; Pakin, Scott D.; Somma, Rolando Diego; Vernon, Louis James

Intended for: Report

Issued: 2016-04-04

An Early Quantum Computing Proposal

Frank Alexander, Kipton Barros, Marcus Daniels, Jim Gattiker, Mike Hamada, James Howse, Stephen Lee, Josip Loncaric, Scott Pakin, Rolando Somma, Louis Vernon

Los Alamos National Laboratory
P,O. Box 1663
Los Alamos, NM 87545

November 2014, *revised* March 2016

Introduction—What is a D-Wave?

The D-Wave 2X is the third generation of quantum processing created by D-Wave. NASA (with Google and USRA) and Lockheed Martin (with USC), both own D-Wave systems. Los Alamos National Laboratory (LANL) purchased a D-Wave 2X in November 2015. The D-Wave 2X processor contains (nominally) 1152 quantum bits (or *qubits*) and is designed to specifically perform quantum annealing, which is a well-known method for finding a global minimum of an optimization problem. This methodology is based on direct execution of a quantum evolution in experimental quantum hardware. While this can be a powerful method for solving particular kinds of problems, it also means that the D-Wave 2X processor is not a general computing processor and *cannot* be programmed to perform a wide variety of tasks. It is a highly specialized processor, well beyond what NNSA currently thinks of as an "advanced architecture."

A D-Wave is best described as a quantum optimizer. That is, it uses quantum superposition to find the lowest energy state of a system by repeated doses of power and settling stages. The D-Wave produces multiple solutions to any suitably formulated problem, one of which is the lowest energy state solution (global minimum). Mapping problems onto the D-Wave requires defining an objective function to be minimized and then encoding that function in the Hamiltonian of the D-Wave system. The quantum annealing method is then used to find the lowest energy configuration of the Hamiltonian using the current D-Wave Two, two-level, quantum processor. This is not always an easy thing to do, and the D-Wave Two has significant limitations that restrict problem sizes that can be run and algorithmic choices that can be made. Furthermore, as more people are exploring this technology, it has become clear that it is very difficult to come up with general approaches to optimization that can both utilize the D-Wave and that can do better than highly developed algorithms on conventional computers for specific applications. These are all fundamental challenges that must be overcome for the D-Wave, or similar, quantum computing technology to be broadly applicable.

As a result, D-Wave is risky technology and somewhat controversial in the computational science community. This makes this proposition a truly exploratory one, and the potential research possibilities and recommendations below reflect these realities.

Research Possibilities for D-Wave Technology

There are multiple research avenues to explore with this technology, which are binned into three categories below: traditional NNSA application space, computer science space, and other non-NNSA application space.

Various potential research threads aside, perhaps the most important aspect such work may provide is that it will build an understanding of both the promise and limitations of current D-Wave technology, and guide future developments (an excellent co-design opportunity). Critical, then, to any efforts discussed here is a partnership between D-Wave scientists and LANL scientists to explore the opportunity space, refine approaches (and, perhaps, refine future D-Wave technologies). This, therefore, may represent an excellent avenue for Los Alamos to attract interest and research projects in quantum computing, and re-affirm our expertise in computational co-design and as architecture pioneers at scale (as demonstrated through our experiences throughout the years in this regard, often getting the first versions of a variety of machines including Cray, Connection Machine, and Roadrunner).

This must be tempered with a realistic view of what this opportunity really is and how it may unfold: this is very much exploratory research, in which negative results are extraordinarily valuable as they suggest technology improvements and strategies for the computational community at large. The computational abstraction for D-Wave hardware is not validated (or even discovered in some cases). This requires hardware to resolve. Figuring out what to do with potential computational abstractions (mapping problems to the restricted Hamiltonian) is possible without hardware (mathematical/algorithmic exploration with emulators). In any case, this work will not result in a rapid uptake of quantum computers for general-purpose computational physics problems of interest to NNSA.

Possible Traditional NNSA Mission Research

This section briefly describes ideas for exploiting D-Wave technology that directly impact traditional NNSA missions. This list is not meant be inclusive nor to suggest that these are guaranteed to pan out. This is research. The results are unknown.

Boundary Value Problems, Solver Pre-conditioner Optimization

Explore possibilities, using the D-Wave, to solve a one-dimensional boundary value problem. This would demonstrate that we can use the D-Wave to solve an ordinary differential equation (which is a bare minimum representation of the ASC workload). It is a problem that can likely be formulated as an optimization problem (variational formulation of a boundary value problem), but the challenge would be to derive a Hamiltonian for this problem that can be handled by the current D-Wave architecture.

The D-Wave could potentially be used to design optimal pre-conditioners for PDE solvers, which would then be useful for a variety of computational physics applications. Additionally, if we can solve ordinary differential equations on the D-Wave, we may be able to use it to design optimal discretization strategies for differential equations. Since the optimal pre-conditioner depends on the discretization, another possibility is the co-design of discretization and pre-conditioner strategies using the D-Wave.

Mesh Generation and Optimization

It may be possible to use D-Wave technology to explore the optimization of generating complex computational meshes. For example, most ASC codes do some sort of mesh partitioning which can be considered a graph partitioning algorithm. Exploring such approaches may be consistent with current instantiations of D-Wave technology. Additionally, there may be applicability in mesh optimization issues for a deterministic (Sn) thermal radiation transport code.

Statistical Experimental Design

The design of experiments can be cast as an optimization problem. Consider, for example, the possibility of conducting n experiments in order to explore p variables. A design protocol can be thought of as an $n \times p$ matrix consisting of symbols -1 and $+1$. Orthogonality is desirable, which occurs if all pairs of columns have inner products that equal 0. Often, though, such an orthogonal array may not exist. In that case one would like to find a matrix with the minimum sum of absolute values of inner products between all pairs of columns. Minimizing such an objective function would then optimize the design of the experiment under consideration. The goal (and technical challenge) would be to map the objective function to the set of allowable D-Wave Hamiltonians. At the moment it is unclear how to do this. Moreover, the problem becomes more difficult when there are more than two symbols in each column.

Particle Scheduling

Imagine a set of particles distributed in space and moving on random trajectories in two dimensions. Computationally, assume the goal is to advance the particles particles in parallel. Suppose particle A and particle B cross points at location X at the same time. Each needs to read out the current tally at the location, add one to it, and write it back, which means they can overwrite one another's update. One remedy is to use atomic operations, which require a lot of circuitry to implement. Suppose a qubit is responsible for each tally cell, and that a particle's location can be approximated by the nearest tally cell. Further, perhaps several qubits discretize an angle for a particle (e.g., there could be a qubit for north, south, west, and east). The constraint problem is to find which bits should be "on" (i.e., which particles fire and in which direction) to avoid particles ever possibly hitting one another (that is, needing to update the same tally at the same time). For example, a particle headed north from the southern-most part of the space can potentially hit a particle headed east from the west side of the space. The couplings between these qubits should be strong. On the other hand, if a particle is headed due south from the north west side of the space and another particle is headed due north from the south east side of the space, it is impossible for them to collide. Such scheduling ideas may have applications to physics-based particle motion to optimize distribution of workload or sorting algorithms for "vectorized stacks" of particles to ensure efficient computation.

Materials Microstructure Mapping

Given experimental images of various sources, if an estimate could be made of three-dimensional density variations at various location features, identifying how a material might crack under impacts at different locations might be thought of as a traveling salesman problem (which is adaptable to D-Wave technology). A classical way to implement this would be to index the space into an octet tree then move from a point to N neighboring points recursively (the indices might be very large and the recursive search very deep). A D-Wave machine might be able to trace out paths much more efficiently, and ultimately output a set of location vertices that represent the weakest paths through the material.

Description of Alloy Materials

At first glance the challenge of on-lattice atomic configuration optimization seems well suited to the D-Wave. With atoms as qubits (or sets of qubits) the short-range nature of classically simulated atomic interactions maps conveniently to the low connectivity of qubits on the D-Wave chip, while the qubit coupling field translates well to the pairwise nature of many empirical potentials (functional form aside). Using this basis it seems possible to optimize the configuration of specific canonical and grand canonical atomic ensembles using the D-Wave; such problems might be tackled through Metropolis and grand canonical Monte Carlo classically. To illustrate this idea we can consider the classically complex problem of dopant/defect concentration and distribution in metals; for example the concentration of Gallium in Ga stabilized Pu. In our on-lattice description of the problem, Ga atoms occupy interstitial sites on a sub-lattice that does not perturb the parent Pu lattice (achieved through positive bias and strong Pu-Pu coupling). In this naive example, the Ga-Ga interaction is tuned to be repulsive (representing their physical interaction at the representative displacement in the on-lattice description) while the Ga-Pu interaction is attractive. With the on-chip representation carefully defined and parameterized we could potentially use the D-Wave to solve for both the minimum energy Ga concentration and distribution in a single pass—a problem that is classically computationally expensive to solve. Even if the on-lattice description and pair interaction is non-physical, the D-Wave can provide candidate sets of solutions that we can validate with classical empirical or ab initio methods. This idea of on-lattice configuration optimization with qubits representing the occupancy of sites and coupling fields representing the pairwise attraction or repulsion of relative atomic species at the on-lattice defined displacements could be extended to other interesting materials problems.

Nuclear Nonproliferation

It is conceivable that the D-Wave could be programmed to map sensor measurements from an illegal detonation to the set of devices or device characteristics that could have produced such readings. While this problem is more obviously solved using machine-learning techniques (e.g., neural networks), the D-Wave's ability to "run problems backward" may make it the more suitable vehicle for addressing such challenges.

Possible Computer Science Research

There are a number of interesting computer science research activities that may be possible that would have a direct impact on traditional NNSA missions, as well as the computational community in general.

Identifying problems for which the D-Wave may eventually become superior to current digital computing are potentially important. In fact, beyond D-Wave considerations, if the rate of innovation in shrinking component sizes in lithography processes continues to slow (as predicted through 2020), there will be pressure on the semiconductor industry to diversify their offerings. It is therefore possible to envision products designed to help with various aspects of particular algorithms, where quantum annealing is one example of a possible specialized data processing technology. We could end up with quite a mix of computing technologies that will need different algorithms to exploit efficiently, and post-CMOS technologies as well as non-digital technologies may enter this mix gradually. If computing evolves in this manner, scientific computing will need to further adapt to a diverse computational environment in the 2020–2030 time frame. Early exploration of D-Wave (and other emerging technologies) provides one early vehicle to aid that transformation.

Augmentation of Compiler Optimization/Hardware-enabled Compilation Technology

Compilers represent programs as a graph. At any point in a program it is simple to create a stack trace. Such traces make it possible to categorize a particular statement of expression in a program as a set of paths (which describes what is being accessed and how one arrives at that point in the program). If we consider the location of the program as a "site," and what memory object is being read or written as an "object," these can be thought of as two types of vertices on a bipartite graph. Different sites are connected by edges according to the observed call graph. Sites are connected to objects (memory accesses) on the basis of the meaning of the abstract syntax tree. If a site (line in a program) indicated a read or write to an object, connect them. If a data item is read n times, pull that object node n times "closer" to the requesting site compared to another site that made a single request. This approach may map well to the D-Wave system, and can lead to graph matching algorithm data flow to machine topology, optimized pre-fetch, data layout for sparse solvers, and other compile-time optimizations on conventional architectures.

One specific example would be to examine dynamic polyhedral code transformations and optimizations for a code that uses adaptive mesh refinement, which may be amenable to a quantum annealing approach. Using open source compilers, ask to what extent are optimizations not considered due to computational cost (both static compilation costs and dynamic compilation costs). Would quantum annealing hardware, like the D-Wave, help improve feasibly in either case? This may also help lay groundwork for deeper mathematical optimization in NNSA code improvement tasks.

Job Scheduling

Network fabrics are designed around various constraints and assumptions; Cielo has a 3D torus, while Sequoia is a 5D torus. A fast mapping of a computational problem should have a logical topology similar to the underlying physical topology. It is common practice however, to run many small jobs on such systems which share this fabric, and many such jobs wind up using the interconnect in a sub-optimal manner. Formulating a penalty function to rank-order good versus bad candidate node sets for a given application should be possible, with a "bad" node set defined to be one that utilizes communications over slow links for long distances and a "good" node set defined to be one that has only fast links. Costs like checkpoint/restart, memory hierarchy capacity and bandwidth, network bandwidth, storage, energy utilization, and so on could be included in this penalty function. If such an approach could be mapped to a D-Wave, the optimal solution might yield better multi-resource scheduling as well as improvements in the applications themselves. Such fine-grained scheduling will drive us away from bulk-synchronous programming, which is necessary for most future architectures ("conventional" ones, independent of the D-Wave). This work might serve to address the question of doing computation efficiently from a system perspective, which is also a key issue for future exascale systems.

Improved Utilization of Accelerated Conventional Computing

A good example of such systems today is the use of GPUs to augment conventional microprocessor architectures. GPUs can perform simple operations simultaneously across high dimensional index spaces. Using a GPU efficiently can be thought of as a scheduling problem because the data must be sent to the device, then returned from the device, and these transfers have a cost penalty. It is obviously best to do such transfers during computation. Additionally, memory transfers and calculations can be done in different streams on a GPU. A stream executes transfers and calculations in the order in which the work was queued. It would be easier for the programmer if the dependencies could be declared and the system software ensured an optimally efficient execution plan.

Additionally, it is common with ASC codes to compute stencils across a computational mesh. Given how quantities on the mesh are calculated from neighboring information, and given the simultaneous execution of neighbors on a GPU, there is no way to ensure what version of the quantity is being read or written. Locking does not scale, as there is too much memory and coordination needed. It may be possible to address this issue through optimized scheduling, such that a transformation of the traversal order results in independent mesh cell reads and writes (which is labor intensive). Current compiler optimizations cannot optimize nested loops when the loop ranges are unknown at compile time (which is often the case). However, if such integer programming problems could be simultaneously solved "instantaneously" with quantum annealing (on a technology like the D-Wave), it may be possible to consider many more complex computational kernels as candidates for GPU execution without introducing a large labor cost, which is the case today.

Memory Partitioning

ASC supercomputers like Trinity phase II comprise different types of memory with different performance properties. Determining which fields of which data structures should be collocated is an optimization problem. We can potentially address this problem by constructing a graph of data accesses. Data records can be represented by nodes, weighted by their memory footprint, and consecutive occurrences of a record can be represented by edges, weighted negatively if they access the same mesh field or positively if not. This graph has a natural mapping to the D-Wave's Hamiltonian and can therefore possibly be run on the D-Wave concurrently to a physics simulation running on the classical supercomputer, occasionally feeding back updated suggestions for data layout.

Programming Models

Physically, the D-Wave solves a single problem—finding ground states in Ising systems with up to two-spin interactions. Most, if not all, programming models provided with the system are low-level abstractions of this physical reality. We have identified a number of directions for programming-model research that can be performed to investigate higher-level abstractions to more easily map a wider variety of problems to the D-Wave than overt optimization problems. This effort could potentially open up the D-Wave to a greater set of NNSA mission needs than optimization alone.

Probabilistic Computing

The D-Wave produces a set of "answers" to a given optimization problem, one of which represents the "correct" minimum energy configuration. This can therefore also be thought of as a form of probabilistic computing, in which quick, but possibly incorrect, answers are provided. Such computing may be exploitable for image processing and analysis for a variety of areas, including those of interest to NNSA.

Possible Nontraditional NNSA Research

Finally, there are multiple possibilities for D-Wave technology that lie outside the traditional NNSA mission area, but which are potentially useful and may also positively (if indirectly) directly impact NNSA mission space.

There are also a number of other application areas that are not appropriate to discuss in this document.

Finding Weaknesses in Viruses

One may be able to use D-Wave technology to discover weaknesses in biological or cyber viruses. This is an alternative to structural reconstruction. Here, rather than worrying about the structure, one gathers together the aligned sequences and does a search for the combination of (for a biological virus) amino acid states that best fit the neutralization level related to the sequence (or whether the patient lived or died). The minimal set of states in the sequence that is shared by the "live" sequences versus the "die" sequences is likely the relevant part of an antibody response. Given that the D-Wave returns multiple answers as it progresses, there could be multiple interacting epitopes where one result would name one of the states and another result could be an orthogonal set of states (perhaps representing a weaker antibody response). Direct analogies to cyber viruses apply, but the problem is that they are "longer" thereby requiring more qubits than available at present to find relevant instructions, memory addresses, or signatures.

Connected Graphs with Locality

There are broad classes of complex optimization problems that may be able to exploit D-Wave technology. The problem can be described in a general manner as follows. Assume you have a set of nodes with unknown connections between them, but you can collect data at the nodes. The data at the nodes follows a distribution with parameters that are a function of the unknown topology of the connections. We want to maximize the likelihood over the combinatorial space of possible links between the nodes. In an electrical grid example, you want to optimize the placement of data collection units (called Phasor Measurement Units, or PMUs, on the grid) on the nodes such that you maximize the probability of correctly identifying downed power lines. This is a complex problem, as each candidate placement is an integral over many continuous unknowns (such as loads at the nodes). There are similar applications in transportation, communication, social, and biological networks. It is not clear at present how data, network information, or prior distribution on unknowns would be represented on the D-Wave machine, nor how it would evaluate the objective function (the maximum of a marginal posterior distribution).

Image Processing

Image segmentation—isolating "interesting" objects in a photographic image—is an important an rapidly growing field. The D-Wave can represent adjacent-pixel similarity in terms of coupler strengths between qubits and thereby group together blocks of pixels that likely belong to the same object. The results can be used, for example, to identify treaty-violating facilities in foreign countries.

Long-term Vision

This specific exploration of D-Wave technology is intended to lead to a sustained effort in unconventional computing. The vision is the establish a combined science and technology exploration activity to evaluate and co-design computing technologies, algorithms, and codes that may be decades away (or be eliminated entirely), thus enabling the long-term viability of such technologies and the NNSA assets that will need to take advantage of them. Los Alamos is uniquely suited to pioneer such efforts.