

Faster identification of optimal contraction sequences for arbitrary tensor networks

Robert N. C. Pfeifer, Jutho Haegeman, Frank Verstraete

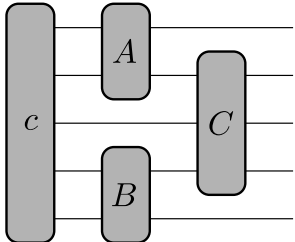
The efficient evaluation of tensor expressions involving sums over multiple indices is of significant importance to many fields of research, including quantum information, quantum many-body physics, loop quantum gravity, and quantum chemistry. In quantum information science the evaluation of tensor expressions is required when classically simulating the behaviour of a quantum circuit, either for comparison with experiment or to facilitate an understanding of the behaviour of a complex system. The inherent challenge of classically simulating a quantum system is well-understood, with the dimension of the Hilbert space growing exponentially with system size. Less widely-recognised, however, is the importance to computational efficiency of the order in which gates in the circuit are either applied or combined.

In general, the task of determining the optimal sequence for evaluating multiple gate applications is NP-hard, being isomorphic to the problem of contracting an arbitrary tensor network [1, 2]. Equivalent problems in quantum chemistry and quantum many-body physics have been the subject of intense study at least since 1997 [see e.g. 1–18], and it has long been acknowledged within the quantum chemistry community that determination of optimal evaluation sequences represents a significant bottleneck in the development of new algorithms [3, 4, 6, 15]. The search for optimal evaluation sequences may be automated through the use of software such as the Tensor Contraction Engine (TCE) [3, 14], but this task represents a significant computational burden for large quantum circuits.

In this talk we present a novel approach to the search for an optimal gate application sequence which performs several orders of magnitude faster than existing search algorithms, while still guaranteeing identification of an optimal evaluation sequence for a given quantum circuit. Our work is available on arXiv [19] and includes a reference implementation of the algorithm written in MATLAB and C++ for immediate practical application.

In its most general form, the problem of evaluating the action of a quantum circuit

is that of evaluating the corresponding multidimensional tensor sum, for example



$$\equiv \sum_{i,j,l,m,o,p} c_{ijklm} A_{no}^{ij} B_{pq}^{lm} C_{qrs}^{okp}, \quad (1)$$

as rapidly as possible subject to the constraints of available computing hardware. (A familiar interpretation may be placed on this example by assuming that c represents a quantum state and A , B , and C represent quantum gates.) This problem may be seen as a generalisation of the matrix-chain multiplication problem, where a string of matrices are to be multiplied together as efficiently as possible. Unlike the matrix-chain multiplication problem, however, this problem cannot be solved in polynomial time through the use of dynamic programming techniques [1].

While this optimisation problem is intrinsically multidimensional, balancing available memory and (for multi-node machines) inter-node communication delays against the number of floating-point operations which must be performed, the predominant approach to this problem is first to identify the ideal contraction procedure which would be performed on a single node with infinite resources, minimising the number of floating point operations to be performed (a process known as *operation minimisation*), before trading off performance against memory constraints and distributing the problem across multiple nodes [4, 9, 10, 14, 16, 18]. Consequently, the task of operation minimisation is of fundamental importance.

The dominant approach presently employed may be described as a breadth-first constructive approach. In this approach, for a network of n tensors we create n sets, denoted S_1, \dots, S_n , and place all of the original tensors (gates and states) in set S_1 . A set S_i is then defined to contain all tensors which may be constructed by contracting together i tensors from set S_1 . (For example, a tensor $(cA)_{noklm} = \sum_{i,j} c_{ijklm} A_{no}^{ij}$ is found in S_2 and is computed from tensors c and A in S_1 by summing over indices i and j .) With each tensor is also stored the minimum number of floating point operations required to construct this tensor, and a sequence of operations which yields the tensor for this cost.

It may be shown that the optimal contraction sequence is always a series of pairwise tensor contractions, and thus for tensors in S_i we need only consider pairwise contractions between elements of S_j and S_{i-j} for $1 \leq j \leq \lfloor \frac{i}{2} \rfloor$. The cheapest means of constructing an element in S_i is therefore the cheapest such pairwise contraction, taking into account the computational cost of constructing the two elements of S_j and S_{i-j} . Note that it is *not* required that the elements of S_j and S_{i-j} share any common index, and indeed there exist quantum circuits for which the optimal evaluation sequence necessarily involves contraction of tensor pairs not sharing any index, or sharing only indices of dimension 1.

The search is completed on iterating over all possible pairwise contractions yielding S_n , with the cheapest identified construction then corresponding to the cheapest

sequence for contracting the tensor network as a whole, and thus evaluating the action of the quantum circuit.

We improve on this method in two areas:

- First, we recognise that there are frequently elements in intermediate sets S_i , $1 < i < n$, for which the cost of computing the element exceeds the minimum cost of evaluating the quantum circuit as a whole. By prioritising the construction of cheaper elements in the intermediate sets, we are able to greatly restrict the number of tensor contraction sequences which must be considered before a known optimal-cost sequence is identified.
- Second, while it may sometimes be necessary to contract together tensors sharing no common index as a part of the optimal contraction sequence [e.g. $(AB)_{ab} = A_a B_b$], we obtain a number of analytical proofs which substantially constrain the circumstances under which this is necessary, again substantially reducing the number of sequences which need be explored.

The outcome of these refinements is a search algorithm which we describe as cheapest-first constructive, in which sequences having a cost bounded by some number of floating point operations ξ_{\max} are explored first, with bound ξ_{\max} being progressively increased until the minimum-cost sequence for evaluating the entire tensor sum is obtained.

We demonstrate the usefulness of this search algorithm for a number of quantum circuits commonly employed in condensed matter simulations, involving between 5 and 27 tensors. Our algorithm showed superior performance for all circuits examined, most dramatically for quantum circuits involving larger numbers of gates, with results being obtained as much as 10^5 times faster than the standard breadth-first algorithm:

Number of tensors	Time for breadth-first (s)	Time for cheapest-first (s)
5	0.0014	0.0013
6	0.0016	0.0015
7	0.0025	0.0019
9	0.0152	0.0036
11	0.0946	0.0048
19	7298	0.069
27	*	36

* Insufficient memory to perform calculation without swapping to disk (48Gb node).

Based on these results, we hope that our algorithm will be of substantial benefit to the scientific community, both directly to those whose work requires them to contract tensor networks in the simulation of quantum information or condensed matter systems, and to those implementing software for the efficient simulation of quantum systems, and indirectly to the quantum chemists, condensed matter physicists, quantum information scientists, and others who will make use of these software systems.

References

- [1] Chi-Chung Lam, P. Sadayappan, and Rephael Wenger. On optimizing a class of multi-dimensional loops with reduction for parallel execution. *Parallel Processing Letters*, 07(02):157–168, 1997. doi: 10.1142/S0129626497000176. URL <http://www.worldscientific.com/doi/abs/10.1142/S0129626497000176>.
- [2] I. Markov and Y. Shi. Simulating quantum computation by contracting tensor networks. *SIAM Journal on Computing*, 38(3):963–981, 2008. doi: 10.1137/050644756. URL <http://dx.doi.org/10.1137/050644756>.
- [3] Alexander A. Auer, Gerald Baumgartner, David E. Bernholdt, Alina Bibireata, Venkatesh Choppella, Daniel Cociorva, Xiaoyang Gao, Robert Harrison, Sriram Krishnamoorthy, Sandhya Krishnan, Chi-Chung Lam, Qingda Lu, Marcel Nooijen, Russell Pitzer, J. Ramanujam, P. Sadayappan, and Alexander Sibiryakov. Automatic code generation for many-body electronic structure methods: the tensor contraction engine. *Molecular Physics*, 104(2):211–228, 2006. doi: 10.1080/00268970500275780. URL <http://www.tandfonline.com/doi/abs/10.1080/00268970500275780>.
- [4] Gerald Baumgartner, David E. Bernholdt, Daniel Cociorva, Chi-Chung Lam, J. Ramanujam, Robert Harrison, Marcel Nooijen, and P. Sadayappan. A performance optimization framework for compilation of tensor contraction expressions into parallel programs. In *Proceedings of the 16th International Parallel and Distributed Processing Symposium, IPDPS '02*, page 33, Washington, DC, USA, 2002. IEEE Computer Society. ISBN 0-7695-1573-8. URL <http://dl.acm.org/citation.cfm?id=645610.661695>.
- [5] G. Baumgartner, D.E. Bernholdt, D. Cociorva, R. Harrison, So Hirata, C. Lam, M. Nooijen, R. Pitzer, J. Ramanujam, and P. Sadayappan. A high-level approach to synthesis of high-performance codes for quantum chemistry. In *Supercomputing, ACM/IEEE 2002 Conference*, pages 5–5, 2002. doi: 10.1109/SC.2002.10056.
- [6] G. Baumgartner, A. Auer, D.E. Bernholdt, A. Bibireata, V. Choppella, D. Cociorva, X Gao, R.J. Harrison, S. Hirata, S. Krishnamoorthy, S. Krishnan, C. Lam, Qingda Lu, M. Nooijen, R.M. Pitzer, J. Ramanujam, P. Sadayappan, and A. Sibiryakov. Synthesis of high-performance parallel programs for a class of ab initio quantum chemistry models. *Proceedings of the IEEE*, 93(2):276–292, 2005. ISSN 0018-9219. doi: 10.1109/JPROC.2004.840311.
- [7] D. Cociorva, J. Wilkins, G. Baumgartner, P. Sadayappan, J. Ramanujam, M. Nooijen, D. Bernholdt, and R. Harrison. Towards automatic synthesis of high-performance codes for electronic structure calculations: Data locality optimization. In Burkhard Monien, ViktorK. Prasanna, and Sriram Vajapeyam, editors, *High Performance Computing — HiPC 2001*, volume 2228

of *Lecture Notes in Computer Science*, pages 237–248. Springer Berlin Heidelberg, 2001. ISBN 978-3-540-43009-4. doi: 10.1007/3-540-45307-5_21. URL http://dx.doi.org/10.1007/3-540-45307-5_21.

- [8] D. Cociorva, J. W. Wilkins, C. Lam, G. Baumgartner, J. Ramanujam, and P. Sadayappan. Loop optimization for a class of memory-constrained computations. In *Proceedings of the 15th International Conference on Supercomputing, ICS '01*, pages 103–113, New York, NY, USA, 2001. ACM. ISBN 1-58113-410-X. doi: 10.1145/377792.377814. URL <http://doi.acm.org/10.1145/377792.377814>.
- [9] Daniel Cociorva, Gerald Baumgartner, Chi-Chung Lam, P. Sadayappan, J. Ramanujam, Marcel Nooijen, David E. Bernholdt, and Robert Harrison. Space-time trade-off optimization for a class of electronic structure calculations. In *Proceedings of the ACM SIGPLAN 2002 Conference on Programming Language Design and Implementation, PLDI '02*, pages 177–186, New York, NY, USA, 2002. ACM. ISBN 1-58113-463-0. doi: 10.1145/512529.512551. URL <http://doi.acm.org/10.1145/512529.512551>.
- [10] Daniel Cociorva, Gerald Baumgartner, Chi-Chung Lam, P. Sadayappan, J. Ramanujam, Marcel Nooijen, David E. Bernholdt, and Robert Harrison. Space-time trade-off optimization for a class of electronic structure calculations. *SIGPLAN Not.*, 37(5):177–186, May 2002. ISSN 0362-1340. doi: 10.1145/543552.512551. URL <http://doi.acm.org/10.1145/543552.512551>.
- [11] D. Cociorva, X Gao, S. Krishnan, G. Baumgartner, C. Lam, P. Sadayappan, and J. Ramanujam. Global communication optimization for tensor contraction expressions under memory constraints. In *Parallel and Distributed Processing Symposium, 2003. Proceedings. International, 2003*. doi: 10.1109/IPDPS.2003.1213121.
- [12] Albert Hartono, Alexander Sibiryakov, Marcel Nooijen, Gerald Baumgartner, David E. Bernholdt, So Hirata, Chi-Chung Lam, Russell M. Pitzer, J. Ramanujam, and P. Sadayappan. Automated operation minimization of tensor contraction expressions in electronic structure calculations. In Vaidy S. Sunderam, Geert Dick Albada, Peter M. A. Sloot, and Jack J. Dongarra, editors, *Computational Science – ICCS 2005*, volume 3514 of *Lecture Notes in Computer Science*, pages 155–164. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-26032-5. doi: 10.1007/11428831_20. URL http://dx.doi.org/10.1007/11428831_20.
- [13] Albert Hartono, Qingda Lu, Xiaoyang Gao, Sriram Krishnamoorthy, Marcel Nooijen, Gerald Baumgartner, David E. Bernholdt, Venkatesh Choppella, Russell M. Pitzer, J. Ramanujam, Atanas Rountev, and P. Sadayappan. Identifying cost-effective common subexpressions to reduce operation count in tensor contraction evaluations. In Vassil N. Alexandrov, Geert Dick Albada, Peter M. A. Sloot, and Jack Dongarra, editors, *Computational Science – ICCS 2006*, volume 3991 of *Lecture Notes in Computer Science*, pages 267–275. Springer Berlin

- Heidelberg, 2006. ISBN 978-3-540-34379-0. doi: 10.1007/11758501_39. URL http://dx.doi.org/10.1007/11758501_39.
- [14] So Hirata. Tensor contraction engine: abstraction and automated parallel implementation of configuration-interaction, coupled-cluster, and many-body perturbation theories. *J. Phys. Chem. A*, 107(46):9887–9897, 2003. doi: 10.1021/jp034596z. URL <http://pubs.acs.org/doi/abs/10.1021/jp034596z>.
- [15] So Hirata, Peng-Dong Fan, Alexander A. Auer, Marcel Nooijen, and Piotr Piecuch. Combined coupled-cluster and many-body perturbation theories. *J. Chem. Phys.*, 121(24):12197–12207, 2004. doi: <http://dx.doi.org/10.1063/1.1814932>. URL <http://scitation.aip.org/content/aip/journal/jcp/121/24/10.1063/1.1814932>.
- [16] Chi-Chung Lam, P. Sadayappan, and Rephael Wenger. Optimization of a class of multi-dimensional integrals on parallel machines. In *Proc. of Eighth SIAM Conf. on Parallel Processing for Scientific Computing*, 1997.
- [17] C. Lam, P. Sadayappan, D. Cociorva, M. Alouani, and J. Wilkins. Performance optimization of a class of loops involving sums of products of sparse arrays. In *Ninth SIAM Conference on Parallel Processing for Scientific Computing*, San Antonio, TX, March 1999.
- [18] Chi-Chung Lam, Daniel Cociorva, Gerald Baumgartner, and P. Sadayappan. Optimization of memory usage requirement for a class of loops implementing multi-dimensional integrals. In Larry Carter and Jeanne Ferrante, editors, *Languages and Compilers for Parallel Computing*, volume 1863 of *Lecture Notes in Computer Science*, pages 350–364. Springer Berlin Heidelberg, 2000. ISBN 978-3-540-67858-8. doi: 10.1007/3-540-44905-1_22. URL http://dx.doi.org/10.1007/3-540-44905-1_22.
- [19] Frank Verstraete Robert N. C. Pfeifer, Jutho Haegeman. Faster identification of optimal contraction sequences for tensor networks. URL <http://arxiv.org/abs/1304.6112>.