Abstract

Manal Helal

Indexing and Partitioning Schemes For Distributed Tensor Computing With Application To Multiple Sequence Alignment

This thesis investigates indexing and partitioning schemes for high dimensional scientific computational problems. Building on the foundation offered by Mathematics of Arrays (MoA) for tensor-based computation, the ultimate contribution of the thesis is a unified partitioning scheme that works invariant of the dataset dimension and shape. Consequently, portability is ensured between different high performance machines, cluster architectures, and potentially computational grids. The Multiple Sequence Alignment (MSA) problem in computational biology has an optimal dynamic programming based solution, but it becomes computationally infeasible as its dimensionality (the number of sequences) increases. Even sub-optimal approximations may be unmanageable for more than eight sequences. Furthermore, no existing MSA algorithms have been formulated in a manner invariant over the number of sequences. This thesis presents an optimal distributed MSA method based on MoA. The latter offers a set of constructs that help represent multidimensional arrays in memory in a linear, concise and efficient way. Using MoA allows the partitioning of the dynamic programming algorithm to be expressed independently of dimension. MSA is the highest dimensional scientific problem considered for MoA-based partitioning to date. Two partitioning schemes are presented: the first is a master/slave approach which is based on both master/slave scheduling and slave/slave coupling. The second approach is a peer-to-peer design, in which the scheduling and dependency communication are calculated independently by each process, with no need for a master scheduler. A search space reduction technique is introduced to cater for the exponential expansion as the problem dimensionality increases. This technique relies on defining a hyper-diagonal through the tensor space, and choosing a band of neighbouring partitions around the diagonal to score. In contrast, other sub-optimal methods in the literature only consider projections on the surface of the hyper-cube. The resulting massively parallel design produces a scalable solution that has been implemented on high performance machines and cluster architectures. Experimental results for these implementations are presented for both simulated and real datasets. Comparisons between the reduced search space technique of this thesis with other sub-optimal methods for the MSA problem are presented.