Parallel Computing 36 (2010) 449–468

Contents lists available at ScienceDirect

Parallel Computing

journal homepage: www.elsevier.com/locate/parco

Acceleration of hierarchical Bayesian network based cortical models on multicore architectures

Pavan Yalamanchili [a], Sumod Mohan [a], Rommel Jalasutram [a], Tarek Taha [b,*]

[a] Electrical and Computer Engineering, Clemson University, Clemson, SC 29634, USA
[b] Electrical and Computer Engineering, University of Dayton, Dayton, OH 45469, USA

ARTICLE INFO

Article history:
Received 31 December 2008
Received in revised form 9 April 2010
Accepted 9 April 2010
Available online 20 April 2010

Keywords:
Multicore
Parallelization
Cortical models

ABSTRACT

We examine the parallelization of two biologically inspired hierarchical Bayesian cortical models onto recent multicore architectures. These models have been developed recently based on new insights from neuroscience and have several advantages over traditional neural network models. In particular, they need far fewer network nodes to simulate a large scale cortical model than traditional neural network models, making them computationally more efficient. This is the first study of the parallelization of this class of models onto multicore processors. Our results indicate that the models can take advantage of both thread and data level parallelism to provide significant speedups on multicore architectures. MPI implementations on clusters of multicore processors were also examined, and showed that the models scaled well with the number of machines in the clusters.

1. Introduction

Recent scientific studies of the primate brain have led to new neuromorphic computational models [3,9,13,18,23,38] of the information processing taking place in the cortex. These cortical models provide insights into the workings of the brain and concur well with experimental results. The models differ significantly from traditional neural network models in that they are generally at a higher level of abstraction than neural network models and they consider several new biological details about the organization and processing in the cortex. Some of these newer cortical models [8,14] are based on hierarchical Bayesian networks and incorporate several of the recently suggested properties of the neocortex [23,29]. These include a hierarchical structure of uniform computational elements, invariant representation and retrieval of patterns, auto associative recall, and sequence prediction through both feed-forward and feed-back inference between layers in the hierarchy.

These new models utilize several recent finings from neuroanatomits. In particular, neuroanatomists have identified that a collection of about 80–100 neurons form into regular patterns of local cells running perpendicular to the cortical plane [17]. These collections of neurons are called mini-columns. Mountcastle [28] states that the basic unit of cortical operation is the mini-column and that a collection of mini-columns are grouped into a cortical column. He also states that the mini-columns within a cortical column are bound together by a common set of inputs and short-range horizontal connections.

Hierarchical Bayesian network based cortical models have a significant computational advantage over traditional neural network models. Each node in the former models a cortical mini-column or a cortical column, while in the latter each node models only a single neuron. Thus to model a large collection of neurons, a hierarchical Bayesian network based model

* Corresponding author. Tel.: +1 937 229 3119.
E-mail addresses: pyalama@clemson.edu (P. Yalamanchili), sumodm@clemson.edu (S. Mohan), rjalasu@clemson.edu (R. Jalasutram), ttaha@ieee.org (T. Taha).

0167-8191/$ - see front matter © 2010 Published by Elsevier B.V.
doi:10.1016/j.parco.2010.04.001

would require far fewer nodes than a traditional neural network based model. Additionally, the number of node-to-node connections is greatly reduced in hierarchical Bayesian network based cortical models. Anatomical evidence suggests that most of neural connections in the cortex are within a column as opposed to being between columns.

The brain utilizes a large collection of slow neurons operating in parallel to achieve very powerful cognitive capabilities. There has been a strong interest amongst researchers to develop large parallel implementations of cortical models on the order of animal or human brains. At this scale, the models have the potential to provide much stronger inference capabilities than current generation computing algorithms [7]. A large domain of applications would benefit from the stronger inference capabilities including speech recognition, computer vision, textual and image content recognition, robotic control, and making sense of massive quantities of data. Several research groups are examining large scale implementations of neuron based models [1,24] and cortical column based models [21,35]. Such large scale implementations require high performance resources to run the models at reasonable speeds. IBM is utilizing a 32,768 processor Blue Gene/L system to simulate a spiking network based model [1], while EPFL and IBM are utilizing a 8192 processor Blue Gene/L system to simulate a sub-neuron based cortical model [24]. The PetaVision project announced recently at the Los Alamos National Laboratory in June 2008 is utilizing the Roadrunner supercomputer (currently ranked as the world's fastest supercomputer) to model the human visual cortex [31].

In this paper we examine optimizations and parallel implementations on multicore architectures of the recognition phase of two recent hierarchical Bayesian network cortical models. Implementations on clusters of multicore processors using MPI are also examined. The two models examined are Hierarchical Temporal Memories (HTM) [18] and Dean's Hierarchical Bayesian model (to be referred to as the Dean model in the rest of the paper) [8]. At present we are not aware of any other hierarchical Bayesian cortical models (other than updates to these models and Lee and Mumford's work [23], on which Dean's model is based). The training of the models are generally carried out in a longer offline process, only the acceleration of the recognition phase is considered.

With the limited scaling in processor clock frequencies, multicore processors have become the standard industrial approach to improve processor performance. However we are not aware of any studies examining the parallelization or implementation of any hierarchical Bayesian cortical models onto multicore processors. Lansner and Johannson [21] have shown that mouse sized cortical models developed on a cluster of commodity computers are computationally bound rather than communication bound. Therefore the acceleration of the computations of these models on multicore architectures can provide significant performance gains to enable large scale implementations. The multicore architectures examined in this study are the 8 + 1 core IBM/Sony/Toshiba Cell broadband engine [16], the quadcore Intel Xeon processor, and the eight core Sun UltraSPARC T2 Plus processor [33]. Additionally, two clusters based on the Cell processor and the Intel Xeon processor were also examined. The Cell processor has attracted significant attention recently because of its large number of high performance processing cores. The fastest supercomputer at present, the IBM Roadrunner supercomputer installed at Los Alamos National Laboratory, utilizes 12,240 Cell processors and 6912 AMD Opteron processors. The PetaVision project at that laboratory is modeling "1 billion visual neurons and trillions of synapses" [31] on this machine. Details of the project are not publicly available yet, however, this appears to be a neuron level model.

The main contributions of this work are:

1. A study of the parallelization of two hierarchical Bayesian cortical models. We examine both thread level parallelization and data level parallelization of the models.
2. A study of different optimizations and parallelization strategies for the models.
3. An evaluation of the multicore implementations of the models. We examine the performance of the models on three multicore processors using four platforms (a Sony Playstation 3, an IBM QS20 blade, a Sun Enterprise 5140 server, and a dual processor Intel Xeon blade). Several sizes of the model networks were implemented to examine the effect of scaling.
4. A preliminary study of the cluster implementation of the models. Two clusters were utilized for: the Palmetto Cluster at Clemson University containing dual processor Intel Xeon blades and a cluster of Sony Playstation 3s at the Arctic Region Supercomputing Center (ARSC).

Our results indicate that optimized parallel implementations of the model can provide significant speedups on multicore architectures. Using all eight cores on of Cell processor on an IBM QS20 blade provided a speedup of 107 times for the Dean model and 93 times for the HTM model over a serial implementation on the Power Processor Unit of the Cell processor. The quadcore Intel Xeon processor provided a speedup of 36 times for the Dean model and 43 times for the HTM model. The Sun UltraSPARC T2 Plus processor provides a speedup of 17 times over the serial implementation for both the Dean and HTM models. The MPI versions were able to provide near linear speed ups. A cluster of eight Intel Xeon blades was able to provide a speed up of 7.6 times over a single blade for the HTM model and up of 6.8 times for the Dean model. The cluster of eight Playstation 3s provided speedups of 7.8 for the HTM model, and 6.5 for the Dean model over a single Playstation 3. We noticed that the speedups on all the processing platforms increased as we increased the network size. These speedup numbers were for the largest networks we tested.

Section 2 of this paper describes the two models, Section 3 examines the multicore architectures examined, and Section 4 details related work in the area. Sections 5 and 6 discuss the parallelization of the models and their implementation on multicore architectures respectively. Section 7 discusses the experimental setup for the evaluations, Section 8 examines the results on the study, and Section 9 concludes the paper.