Cognitive Domain Ontologies: HPCs to Ultra Low Power Neuromorphic Platforms

Tarek M. Taha^{1*}, Chris Yakopcic¹, Nayim Rahman¹, Tanvir Atahary¹, and Scott Douglass²

¹Dept. Of Electrical and Computer Engineering, University of Dayton, Dayton, OH, USA

²Air Force Research Laboratory, 711 Human Performance Wing, Wright-Patterson AFB, OH, USA

*tarek.taha@udayton.edu

CCS CONCEPTS

CCS → Hardware → Emerging technologies → Analysis and design of emerging devices and systems → Emerging architectures

KEYWORDS

Neuromorphic, Hardware, Cognitive Domain Ontology, Low Power

1 Introduction

The Cognitively Enhanced Complex Event Processing (CECEP) agent-based decision-making architecture is being developed at AFRL/RHCI [1]. Within this agent, the Cognitive Domain Ontology (CDO) component is the slowest for most applications of the agent. We show that even after acceleration on a high performance server based computing system enhanced with a high end graphics processing unit (GPU), the CDO component does not scale well for real time use on large problem sizes. Thus, to enable real time use of the agent, particularly in power constrained environments (such as autonomous air vehicles), alternative implementations of the agent logic are needed. These alternative implementations need to utilize different algorithms that implement the CDO logic and need to be targeted to much lower power (and weight) computing systems than GPU enabled servers (which can consume over 500W and weigh over 50lbs).

The objective of this work was to carry out an initial design space search of algorithms and hardware for decision making through the domain knowledge component of CECEP (the CDO) [2-5]. Several algorithmic and circuit approaches are proposed that span across six hardware options of varying power consumption and weight (ranging from over 1000W to less than 1W). The algorithms range from exact solution producers optimized for running on a cluster of high performance computing systems [1] to approximate solution producers running fast on low power neuromorphic hardware [6-9]. The loss in accuracy for the approximate approaches is minimal, making them well suited to SWaP constrained systems, such as UAVs. The exact solution approach on an HPC will give confidence that the best answer has been evaluated (although this may take some time to generate).

2 Cognitive Domain Ontologies

The CECEP framework (see Fig. 1) consists of a coordinated set of net-centric event processing components each of which DOI: 10.1145/3381755.3381781

provides a unique information/knowledge processing capability. The computationally heavy tasks in CDO solving can be accelerated via algorithm parallelization onto a high-performance computing cluster to achieve exact answers quickly. This approach gives exact solutions from the CDO but requires significant computational energy (a high-performance cluster could consume several kW and weigh well over 100lbs). Given that autonomous systems are generally highly Size, Weight, and Power (SWaP) constrained, CDOs will have to run on lower power computing systems at real time in order to be usable on these platforms. By real time, we mean answers should be available within a few seconds at the most, though depending on the speed of a UAV and task being performed, faster processing may be needed.

Fig. 1. The CECEP Agent

Fig. 2 shows the run time of an $M \times N$ task allocation problem (a problem commonly tasked to CDOs) on a high-performance computing system with an NVIDIA K80 GPU. The algorithm was highly optimized and was able to fully exploit the parallelism on the K80 GPU. As the problem size grows, the run time on this system increases exponentially to the point that solution processing

time is intractable. For example, a 15×15 problem will take around 607 years to solve using our parallelized exact solution algorithm.

Fig. 2. Estimated Run Time for Exact Answers from The $M \times N$ Task Allocation CDO (where $N=M$) Using a Highly Parallelized and Optimized System Using a Server with A High-End NVIDIA K80 GPU.

3 Results and Discussion

We developed a novel parallel path based forward checking algorithm to solve CDOs through a traditional exhaustive search approach [1]. The key benefit of this approach is that it is guaranteed to find the best answer for any task. Our approach first pruned the search space very efficiently to reduce the solution space that had to be search exhaustively. We were able to implement this approach on 128 compute nodes at the Ohio Supercomputing Center (768 cores) to achieve a speedup of over 200 times compared to a serial implementation of the algorithm. Our neural approach for solving CDOs is compared to this parallel path based solver running on a GPU enabled server.

To alleviate SWaP constraints of an HPC based solution engine, we developed several approximate algorithms to solve CDOs [2-9]. The approximate algorithms generate solutions that are close to the best answer but require significantly less computations than needed to generate the best answer. Thus, we trade off answer quality for the computational complexity needed to enable real time implementation on low SWaP systems.

Fig. 3. Diagram of the spiking algorithm for asset allocation and plot showing neuron accumulations during runtime.

Two of our approximate algorithms are based on spiking neural networks [8,9]. The algorithm in [8] takes advantage of neuron accumulation and matrix-vector processing within a spiking feedback loop to quickly obtain a winning set of neurons. Fig. 3 displays a block diagram for the solution finding process. An input neuron layer is connected to an output/feedback layer through a set

Exact Solution			Approximate Solutions		
Traditional Power Efficient Computers			Neuromorphic Computers (Spiking Neural Network Solutions)

Server with GPUs	Intel NUC	ASUS Tinkerboard	IBM TrueNorth	Intel Loihi	Memristors
1500 W	70 W	5 W	0.1 W*	0.1 W*	0.1 mW*
50 lbs	3 lbs	2 oz	2 oz*	2 oz*	1 oz*
0.95s	0.01s	0.4s	0.04s	0.0007s	0.000005s
100%	99.99%	99.99%	99.99%	99.99%	99.99%

Time and accuracy for solving 8x8 asset allocation problem

*Approximate weight and power

Fig. 4. System study for the execution of low-SWaP cognitive decision making.

of weights. This winning set of neurons then denotes the $M \times N$ vehicle allocation result. Properties such as vehicle to target distance, target priority, and probability of success for a given scenario are directly programmed into a weight matrix that feeds the inputs to the neurons. The plot in Fig. 3 displays the potential for all of the neurons in an example vehicle allocation problem. The four bold highlighted neuron outputs were the first four to fire, and have thus generated an optimized allocation pattern based on a given allocation problem. This algorithm does not require training. Therefore, it can be implemented very efficiently by directly programing the simplified weight matrices into the neuron layers. Furthermore, this algorithm is well suited to low power neuromorphic platforms such as the Intel Loihi and IBM True North systems.

Fig. 4 displays a SWaP comparison of six different systems studied for use in CDO implementation, and some of the greatest efficiency gains were obtained using the neuromorphic systems. When running the presented neural algorithm for vehicle allocation on more traditional systems such as the Intel NUC and the ASUS Tinkerboard, significant SWaP advantages are observed relative to the HPC system. However, the Loihi [10,11] and True North many-core neuromorphic processors provide even more significant gains. Finally, the potential of a memristor based system running this spiking algorithm was also examined [4]. While this experiment was performed in simulation only, it is expected to provide additional orders of magnitude SWaP reduction. The full-scale GPU system requires drastically more resources than all of the portable systems. However, the GPU system is guaranteed to produce the absolute best solution (as long as the problem size is small enough so that a complete solution search can be performed in a reasonable amount of time). In summary, the system analysis in this work presents a range of technology options for deploying decision making in the CECEP engine at low SWaP.

REFERENCES

- [1] T. Atahary, T. Taha, F. Webber, and S. Douglass, "Knowledge mining for cognitive agents through path based forward checking," 16th IEEE/ACIS SNPD, pp. 1-8, June, 2015.
- [2] C. Yakopcic, N. Rahman, T. Atahary, T. M. Taha, and S. Douglass, "Cognitive Domain Ontologies in a Memristor Crossbar Architecture," IEEE National Aerospace and Electronics Conference (NAECON), pp. 76-83, Dayton, OH, June 2017.
- [3] N. Rahman, T. Atahary, T. Taha, S. Douglass, "A pattern matching approach to map cognitive domain ontologies to the IBM TrueNorth Neurosynaptic System." 2017 Cognitive Communications for Aerospace Applications Workshop (CCAA). IEEE, 2017.
- [4] N. Rahman, C. Yakopcic, T. Atahary, R. Hasan, T. M. Taha, and S. Douglass, "Cognitive Domain Ontologies in Lookup Tables Stored in a Memristor String Matching Architecture," 30th annual IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 1-4, Windsor, Ontario, April 2017.
- [5] N. Rahman, T. Atahary, C. Yakopcic, T. M. Taha, Scott Douglass, "Task Allocation Performance Comparison for Low Power Devices," IEEE National Aerospace and Electronics Conference (NAECON), pp. 247-253, Dayton, OH, July, 2018.
- [6] C. Yakopcic, T. Atahary, T. M. Taha, A. Beigh, and S. Douglass, "High Speed Approximate Cognitive Domain Ontologies for Asset Allocation based on Isolated Spiking Neurons," IEEE National Aerospace and Electronics Conference (NAECON), pp. 241-246, Dayton, OH, July, 2018.
- [7] C. Yakopcic, N. Rahman, T. Atahary, T. M. Taha, A. Beigh, and S. Douglass, "High Speed Approximate Cognitive Domain Ontologies for Constrained Asset Allocation based on Spiking Neurons," IEEE National Aerospace and Electronics Conference (NAECON), 2019.
- [8] C. Yakopcic, T. Atahary, N. Rahman, T. M. Taha, A. Beigh, and S. Douglass, "High Speed Approximate Cognitive Domain Ontologies for Asset Allocation Using Loihi Spiking Neurons," IEEE/INNS International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, July, 2019.
- [9] C. Yakopcic, J. Freeman, T. M. Taha, S. Douglass, and Q. Wu, "Cognitive Domain Ontologies Based on Loihi Spiking Neurons Implemented Using a Confabulation Inspired Network," IEEE Cognitive Communications for Aerospace Applications Workshop, June, 2019.
- [10] C. Yakopcic, N. Rahman, T. Atahary, T. M. Taha, and S. Douglass, "Solving Constraint Satisfaction Problems Using the Loihi Spiking Neuromorphic Processor," IEEE Design, Automation and Test in Europe, Grenoble, France, March, 2020 (Accepted).
- [11] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, Y. Liao, C.-K. Lin, A. Lines, R. Liu, D. Mathaikutty, S. McCoy, A. Paul, J. Tse, G. Venkataramanan, Y.-H. Weng, A. Wild, Y. Yang, H. Wang, "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," IEEE Micro, vol. 38, no. 1, pp. 82-99, 2018.