Accelerating Path Planning for Autonomous Driving with Hardware-Assisted Memorization

Mulong Luo
Cornell University
Ithaca, NY, USA
ml2558@cornell.edu

G. Edward Suh
Cornell University
Ithaca, NY, USA
suh@ece.cornell.edu

Abstract—Path planning for autonomous driving with dynamic obstacles poses a challenge because it needs to perform a higher-dimensional search (including time) while still meeting real-time constraints. This paper proposes an algorithm-hardware co-optimization approach to accelerate path planning with high-dimensional search space. First, we reduce the time for a nearest neighbor search and collision detection by mapping nodes and obstacles to a lower-dimensional space and memorizing recent search results. Then, we propose a hardware extension for efficient memorization. The experimental results on a modern processor and a cycle-level simulator show that the execution time is reduced significantly.

I. INTRODUCTION
Path planning for autonomous driving faces challenges including strict real-time constraints as well as dynamic obstacles. First, in autonomous driving, path planning can take 100 ms on average on a commercial autonomous driving development platform [15] like Baidu Apollo, while the safety requirement assumes the sensor-to-actuator latency must be less than 100 ms [9] to ensure the responsiveness. Second, modern autonomous vehicle has to deal with dynamic obstacles. In real-world streets, there are other vehicles moving at high speed, and there are pedestrians walking at low speed. Path planning algorithm has to deal with these moving objects and avoid collisions. It has to take into account the locations of the obstacles at different timesteps. Recent work on path planning algorithms such as PUMP [7] and contingency planning [6] consider dynamic obstacles. However, it could take as long as 500 ms to compute a path in these scenarios on a CPU. Even on a GPU, PUMP still need around 100 ms to find a path.

On-time execution of path planning is important to the safety and efficiency of vehicle. Pruning [11] has been used to accelerate it in software. Algorithm-specific accelerators have also been proposed for rapid-exploring random tree (RRT) [14], probabilistic roadmap (PRM) [10], and A* [8], using different hardware technologies including FPGA or ASICs. While these accelerators for path planning do address the individual needs in the specific scenarios, they do not target the autonomous driving and do not consider dynamic obstacles.

Besides, flexibility to choose and modify the path planning algorithm to be accelerated is also very important, as the regulations and algorithms are evolving continuously. However, many of the existing path planning accelerator designs only consider one specific algorithm [14] at a time with most of the hyper-parameters baked into the accelerator hardware at design/program time. While these hardwired algorithms do perform well and provide good performance on the targeted scenario at the time of launch, they may not be flexible enough to adapt to new or modified algorithms.

In this paper, we address path planning for autonomous driving with dynamic obstacles. The major contributions of this work is as follows:

1) We introduce the space-filling curve for efficient indexing, memorization and pruning of time-consuming nearest neighbor search and collision detection in path planning.
2) We propose a high-performance hardware implementation and the programming interface for memorization of the space-filling curve-indexed tree nodes, which is for nearest neighbor search and collision detection.
3) We demonstrate that the proposed hardware-assisted approach can lead to significant performance improvement for path planning with dynamic obstacles in both synthetic and realistic benchmarks.

II. APPROACH

Handling Dynamic Obstacles: We consider the autonomous vehicle is performing path planning on a 2D plane. At different time, the obstacles will be at different positions on the 2D plane. A dynamic obstacle in a 2D plane will become a set of obstacles whose coordinates in the 3D space can be described by \((x, y, t)\). We can effectively regard the 2D planning problem with dynamic obstacles as a 3D planning problem. However, the time dimension is different from a normal spatial dimension, which requires additional constraint for it. Since time only goes in one direction, for any two nodes \((x_i,y_i,t_i)\) and \((x_j,y_j,t_j)\) on a path \(P\) where \(i < j\), we must guarantee \(t_i < t_j\). The additional constraint is enforced when selecting nearest neighbors in RRT.

Software-based Memorization: We use memorization to reduce the total execution time of path planning. In traditional sampling-based planning, much of the execution time is spent on finding the nearest neighbor (NN) and detecting collisions. To accelerate these procedures, we memorize the recent accessed nodes and the collision state so that some of similar queries can be skipped. Figure 1 shows the high-level flow of our approach. In the baseline RRT, the time consuming nearest
neighbor search and collision detection are performed for all iterations. However, we use an small-size data store named Morton store (name after Morton space-filing curve [4]) to memorize the key information from baseline nearest neighbor search and collision detection. For each iteration, we always check the Morton store first to opportunistically skip the time-consuming baseline NN search and collision detection.

![Image of Morton store](image)

**Algorithm 1 RRT with Morton code-based memorization**

```plaintext
0: while Xgoal not reached do
0:  Xrand ← rand()
0:  if Xsource ← Morton_mem(Xrand) == 0 then
0:    Xnearest ← nearest(Xsource)
0:  end if
0:  Xnew ← steer(Xsource, Xrand)
0:  if state = NO_COLLISION then
0:    state ← collision(Xsource)
0:  end if
0:  morton_update(state, Xsource)
0:  if collision(Xnew, Xsource) == NO_COLLISION then
0:    nn.add(Xnew)
0:  end if
0:  end while
```

**Hardware-assisted Memorization:** The software-based Morton store reduces the execution time significantly by skipping time-consuming baseline NN search and collision detection, and it can be further accelerated by hardware-assisted memorization. Since Morton code calculation and matching is processed by iterating all the existing Morton codes sequentially in software, it can take multiple instructions. To estimate how many dynamic instructions the Morton store takes, we use Valgrind [13] to count the number of dynamic instructions of the Morton store lookups and updates.

![Image of percentage of dynamic instruction count](image)

Figure 2 shows the profiled results of different testcases with different map edge length, timesteps, and number of dynamic obstacles. On average the Morton store-related dynamic instruction count is made up of 20% of the total dynamic instructions, and in the worst case, it can compose of more than 45% of the total number of dynamic instructions. However, with hardware-assisted memorization, the dynamic instruction counts can be reduced dramatically. We can use content-addressable memory for the Morton store lookup and update directly in hardware to reduce the dynamic instruction count and the correspondingly the execution time.

The hardware-based Morton store is implemented by a fully-associative content-addressable memory whose memory-line size is 64 bytes, which can maintain eight 8-byte addresses of RRT nodes. Since these nodes are allocated on the stack, the MSB are always 0 in our settings, and we use the most significant 8 bits of the address to indicate whether there is a collision or not. Figure 3 shows the example of the hardware Morton store implemented in content-addressable memory. The tagstore keeps the Morton codes, and each row contains the state (collision/no collision) information as well as the probabilities.
addresses of the nodes. For collision state, as long as there is one state that indicates collision among all states in one memoryline, the output will be a collision.

The content-addressable memory is connected directly to the CPU, as shown in Figure 4. Upon a hit, the corresponding memoryline used for extracting collision state or node address is processed. Upon a read miss, nothing is modified in the Morton store. While upon a write miss, the oldest referenced memoryline is evicted. In order to access this content-addressable memory in software and provide flexibility to different planning algorithms, we define the following ISA extensions. \texttt{morton\_update}, \texttt{morton\_col} and \texttt{morton\_nn} which are listed in Table I. \texttt{morton\_update} takes in the coordinates of a node and its collision state and updates this information in the Morton store. \texttt{morton\_col} takes in the coordinates of a node, look it up in the content-addressable memory and decides whether there is a collision. \texttt{morton\_nn} finds the memory address of the approximate nearest neighbor in the Morton store by looking up entry with the same Morton code. In this work, we use these instructions to accelerate RRT. However, it can also be used in other sampling-based planning algorithms such as PRM.

III. Evaluation

In this section, we evaluate the effectiveness of the proposed memorization scheme based on the Morton code for both software and hardware implementations.

**Synthetic Test Cases:** For the synthetic test cases, they are on a square map, with dynamic obstacles running on it. The synthetic test cases are characterized by three parameters, i.e., the number of dynamic obstacles, the edge length of the square map, and the number of time steps. The solution path should start at (0,0) and destination is (l,l) where l is the edge length of the square map. We randomly generate the starting and ending locations of the obstacles at the beginning and the ending of the simulation period. The intermediate positions of the obstacles are linearly interpolated. Figure 5 shows an example of the test case with 5 dynamic obstacles, the edge length of 100, and 20 time steps. The solution path from (0,0) to (100,100) is marked in blue.

The number of obstacles determines the difficulty of the path planning problem. In general, with more obstacles, it takes more time to find a safe path and the length of the path may also be longer because more space is occupied by the obstacles. The edge length of the map and the number of time steps mainly increases the execution time for path planning. On the other hand, a longer edge length and more time steps will provide more accurate solutions with more intermediate points along the path.

**Software Performance Evaluation:** We evaluate the software-based Morton store on a desktop with Intel i7-6700 CPU 3.40GHz and 16GB DDR4 memory at 2133MHz. The baseline RRT algorithm is implemented in C++. The baseline nearest neighbor search is based on kd-tree [12]. We use libmorton [3] for Morton code calculation.

We simulate 12 configurations with the map edge length in \{100, 200\}, the number of time steps in \{10,100\}, the number of obstacles in \{5, 10, 20\}. We compare the total execution time of the baseline RRT and the RRT with the software Morton-store (sw-morton). For each configuration, we randomly generate 10 different test cases. Since the algorithm is a randomized algorithm, for each test case, we repeat the experiment 10 times and show the average execution time. Figure 6 shows the execution time. On average, the software Morton store-based path planning reduces the execution time by 51% compared to the baseline RRT.

**Hardware Evaluation:** The proposed hardware-assisted approach uses content-addressable memory for a fast look-up and update of recently-accessed tree nodes for the nearest neighbor search and collision detection. We evaluate the performance
TABLE I
ISA INTERFACE FOR MORTON STORE.

| ISA Instruction   | Operation                                      |
|-------------------|------------------------------------------------|
| morton_update <x|y>, <t>, <addr>          | update a node with coordinate \((x,y,t)\) and memory address \(addr\) in Morton store |
| morton_col <x|y>, <t>, <st>            | check whether there is a collision for node \((x,y,t)\) using Morton store, returns result in \(st\) |
| morton_nn <x|y>, <t>, <addr>          | find an approximate near neighbor of node at \((x,y,t)\) returns the address in \(addr\) |

TABLE II
GEM5 CONFIGURATION.

| Core               | Simple Timing CPU @ 2GHz          |
|--------------------|-----------------------------------|
| L1,i               | 32KB, 4-way, latency=2 cycles     |
| L1,d               | 32KB, 8-way, latency=2 cycles     |
| Morton Store       | 32KB, fully-associative, memoryline=64 bytes, latency=2 cycles |
| L2                 | 256KB, 8-way, latency=20 cycles   |
| L3                 | 512KB, 16-way, latency=20 cycles  |
| Memory             | 4GB, DDR3@1600MHz                 |

Fig. 7. Solution path on the map by baseline RRT with length 175.

Fig. 8. Solution path by the HW-based Morton store with length 217.

of this hardware-based Morton store using gem5 simulator [5]. The simulator configuration is shown in Table II. We implemented HW Morton store and a customized port between CPU and HW Morton store store for the look-up and update in gem5.

We use the same 12 configurations that are used in the software evaluation in the previous subsection. Figure 9 shows the execution time of the baseline, the software-based Morton store (sw-morton) and the hardware-based Morton (hw-morton) store. The performance improvements over the baseline are 2.28X for software-based Morton store and 8X for hardware-based Morton store on average. For the best configuration, the performance is improved by 6.74X and 56.3X for software and hardware Morton stores, respectively. Figure 10 shows the solution path lengths for baseline, the software-based Morton store and the hardware-based Morton store. The average solution length of the software Morton and hardware Morton stores are 1.42X and 1.65X that of the baseline solution. This is due to the fact that the SW or HW Morton store can only find approximate nearest neighbor instead of the exact nearest neighbor. Thus, the solution path may fluctuate more compared to the baseline. As shown in Figure 8, around position \((80,80)\) of the map, path given by the HW-based Morton store moves back and forth before reaching \((100,100)\), which increases the solution length compared to the baseline solution shown in Figure 7. Such path can use refinement [1] to reduce the overall length.

Commonroad Evaluation: To demonstrate the effectiveness of the proposed method on realistic scenarios, we use one example (USA_US101-20_1_T_1) shown in Figure 11 containing dynamic obstacles from the Commonroad [2] benchmarks, whose scenarios are partly recorded from real traffic. The execution time and solution path lengths are shown in Figure 12, which confirms effectiveness in reducing the execution time.

IV. CONCLUSION

In this paper, we proposed a software-hardware co-design approach to accelerate path planning for autonomous driving. In the algorithm/software level, we use Morton codes as the tags to index the approximate nearest neighbor and collision detection. In hardware, we use a content-addressable memory to efficiently search for nodes indexed by Morton codes. The experimental results show that the proposed acceleration can significantly improve the performance over the baseline, by 8X on average in terms of the execution time.
ACKNOWLEDGMENT

The authors would like to thank Helena Caminal, Dr. Jacopo Banfi and Dr. Mohamed Ismail for helpful discussions. This project is partially funded by NSF grant ECCS-1932501.

REFERENCES

[1] R. Alterovitz, S. Patil, and A. Derbakova. Rapidly-exploring roadmaps: Weighing exploration vs. refinement in optimal motion planning. In 2011 IEEE International Conference on Robotics and Automation, pages 3706–3712. IEEE, 2011.
[2] M. Althoff, M. Koschi, and S. Manzinger. Commonroad: Composable benchmarks for motion planning on roads. In 2017 IEEE Intelligent Vehicles Symposium (IV), pages 719–726. IEEE, 2017.
[3] J. Baert. https://github.com/forceflow/libmorton, 2018.
[4] O. G. Consortium et al. Discrete global grid systems abstract specification, 2017.
[5] N. B. et al. The gem5 simulator. SIGARCH Comput. Archit. News, 2011.
[6] J. Hardy and M. Campbell. Contingency planning over probabilistic obstacle predictions for autonomous road vehicles. IEEE Transactions on Robotics, 29(4):913–929, 2013.
[7] B. Ichter, B. Landry, E. Schmerling, and M. Pavone. Robust motion planning via perception-aware multiobjective search on gpus. 2017.
[8] Y. Kim, D. Shin, J. Lee, and H.-J. Yoo. Brain: A low-power deep search engine for autonomous robots. IEEE Micro, 37(5):11–19, 2017.
[9] S.-C. Lin, Y. Zhang, C.-H. Hsu, M. Skach, M. E. Haque, L. Tang, and J. Mars. The architectural implications of autonomous driving: Constraints and acceleration. In ASPLOS, pages 751–766, 2018.
[10] S. Murray, W. Floyd-Jones, G. Konidaris, and D. J. Sorin. A programmable architecture for robot motion planning acceleration. In ASAP, volume 2160, pages 185–188. IEEE, 2019.
[11] J. Pan and D. Manocha. Fast probabilistic collision checking for sampling-based motion planning using locality-sensitive hashing. The International Journal of Robotics Research, 35(12):1477–1496, 2016.
[12] J. Tsuiombikas. http://nuclear.mutantstargoat.com/sw/kdtree.
[13] Valgrind. https://valgrind.org.
[14] S. Xiao, N. Bergmann, and A. Postula. Parallel rrt-star architecture design for motion planning. In FPL, pages 1–4. IEEE, 2017.
[15] B. Yu, W. Hu, L. Xiu, J. Tang, S. Liu, and Y. Zhu. Building the computing system for autonomous micromobility vehicles: Design constraints and architectural optimizations. In MICRO, 2020.