Adjustable Multi-Stream Block-Wise Farthest Point Sampling Acceleration in Point Cloud Analysis

Journal:	*IEEE Transactions on Circuits and Systems II: Express Briefs*
Manuscript ID	TCAS-II-20241-2023
Manuscript Type:	Regular Paper - Letters
Date Submitted by the Author:	05-Dec-2023
Complete List of Authors:	Jiao, Hailong; Peking University Shenzhen Graduate School Zhou, Changchun; Peking University Shenzhen Graduate School Fu, Yuzhe; Peking University Shenzhen Graduate School Ma, Yanzhe; Peking University Shenzhen Graduate School Han, Eryi; Reconova Technologies Co Ltd He, Yifan; Reconova Technologies Co Ltd
EDICS:	DCS110 - Digital ASICs < Digital Circuits and Systems (and VLSI), DCS160A5 - Low power architectures < DCS160 - Low power digital systems < Digital Circuits and Systems (and VLSI), DCS230 - Digital VLSI < Digital Circuits and Systems (and VLSI), DCS230B0 - VLSI digital circuits, designs and implementations < DCS230 - Digital VLSI < Digital Circuits and Systems (and VLSI)
TCAS-II Subject Category Please select the subject category that most closely fits with the scope of your manuscript:	Digital Circuits and Systems and VLSI

Adjustable Multi-Stream Block-Wise Farthest Point Sampling Acceleration in Point Cloud Analysis

Changchun Zhou*, Yuzhe Fu*, Yanzhe Ma, Eryi Han, Yifan He, *Member, IEEE,* and Hailong Jiao, *Member, IEEE*

***Abstract*—Point cloud is increasingly used in a variety of applications. Farthest Point Sampling (FPS) is typically employed for down-sampling to reduce the size of point cloud and enhance the representational capability by preserving contour points in point cloud analysis. However, due to low parallelism and high computational complexity, high energy consumption and long latency are caused, which becomes a bottleneck of hardware acceleration. In this brief, we propose an adjustable multi-stream block-wise FPS algorithm, adjusted by four configurable parameters, according to hardware and accuracy requirements. A unified hardware architecture with one parameter is designed to implement the adjustable multi-stream block-wise FPS algorithm. Furthermore, we present a rapid searching algorithm to select the optimal configuration of the five parameters. Designed in an industrial 28-nm CMOS technology, the proposed hardware architecture achieves a latency of 0.005 (1.401) ms and a frame energy consumption of 0.09 (27.265) µJ/frame for 1 k (24 k) input points at 200 MHz and 0.9 V supply voltage. Compared to the state of the art, the proposed hardware architecture reduces the latency by up to 99.9%, saves the energy consumption by up to 99.5%, and improves the network accuracy by up to 9.34%.**

***Index Terms*—Point cloud neural networks, mapping, parallelism, FPS, acceleration framework, energy consumption, network accuracy, hardware architecture.**

I. Introduction

POINT cloud has recently gained popularity as a data source in a variety of fields, ranging from robotics [1], [2] and autonomous driving [3], [4] to augmented reality (AR) [5]. Farthest point sampling (FPS) algorithm is widely employed in point cloud neural networks (PNNs) [6]–[14], preserving shape and structural information, and significantly reducing the input point cloud size and computational complexity of PNNs. During FPS, the point which is farthest from the output point set is added to the output point set and then the process is iterated [16]. Therefore, FPS is inherently sequential with a long latency of $O(N^2)$ cycles and a high computational complexity of $O(N^2)$, where N represents the number of points in the point cloud [16]. Furthermore, FPS involves a substantial number of memory accesses of $O(N^2)$. Meanwhile, the grain of memory access in FPS is small, just the coordinates of one point. Therefore, FPS results in low memory access efficiency. These make FPS a bottleneck in hardware acceleration for point cloud analysis. Therefore, an FPS acceleration framework that supports parallel computing with low computational complexity, low latency, and low memory access overhead, is highly desirable.

To solve these issues, an adjustable multi-stream block-wise FPS algorithm is proposed to reduce computational complexity and maintain accuracy. A unified hardware architecture is developed to implement the proposed adjustable multi-stream block-wise FPS algorithm. A rapid searching algorithm is proposed to select the best algorithm and hardware configurations with specified priority between accuracy and energy consumption. The proposed FPS acceleration framework improves up to 9.34% network accuracy and reduces up to 99.5% energy consumption and 99.9% latency per frame, compared to the state of the art. Designed in the TSMC 28-nm CMOS technology, the proposed hardware architecture achieves a latency of 0.005 (1.401) ms and an energy consumption of 0.09 (27.265) µJ/frame for 1 k (24 k) input points at 200 MHz and 0.9 V supply voltage.

II. Background and Related Works

FPS acceleration has been investigated sporadically in recent years. An adjustable FPS algorithm is introduced in [16]. Point cloud is segmented into multiple slices, based on the spatial distribution. Similarly, the local FPS is then performed on each segment. However, if the data rearrangement mismatches the spatial distribution of point cloud, the network accuracy is significantly degraded. Furthermore, rearranging the point cloud data before processing leads to extra operations. [19] employs distribution-aware FPS. A spatial grid is used to group point clouds and accelerate FPS by skipping unnecessary computations based on a distance threshold, thereby maintaining accuracy. [20] employs a k-d tree to divide point clouds into buckets and performs intra-bucket FPS. [20] skips memory access of these buckets which are far from the last farthest point. However, both [19] and [20] suffer from serial FPS operations, thereby limiting computation parallelism and introducing long latency.

[15] introduces a block-wise FPS algorithm to reduce latency with two principal steps: prediction and sampling. Two key parameters, namely the sparsity coefficient (S) and the number of blocks (B), are employed. In Fig. 1, with S = 16 and B = 8, the prediction step begins by reducing the input points to a 16× sparser set. FPS is then performed on this sparse set, following a count of points per block, which is then scaled up by a factor

This work was supported in part by the National Natural Science Foundation of China 62074005. (**Changchun Zhou and Yuzhe Fu contributed equally to this work.*) (*Corresponding author: Hailong Jiao.*)

Changchun Zhou, Yuzhe Fu, Yanzhe Ma, and Hailong Jiao are with the School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, Shenzhen, 518055, China (e-mail: jiaohailong@pku.edu.cn). Eryi Han and Yifan He are with Reconova Technologies Co., Ltd., Xiamen, 361015, China.

of 16 (*S*) to estimate the number of points in each block. In the sampling phase, the input points are initially divided into 8 (*B*) blocks. A block-wise FPS, guided by the earlier predicted counts, is conducted on each block. The results of each block are aggregated to form the final sample output. An increase in the values of *S* and *B* tends to reduce the complexity and latency of FPS, yet at the cost of accuracy loss. In detail, a higher *S* leads to the dropout of critical contour points in the prediction step, thereby causing prediction distortion. Similarly, a larger *B* results in a smaller number of points in each block, which causes a substantial discrepancy between the predicted number and the actual number of points.

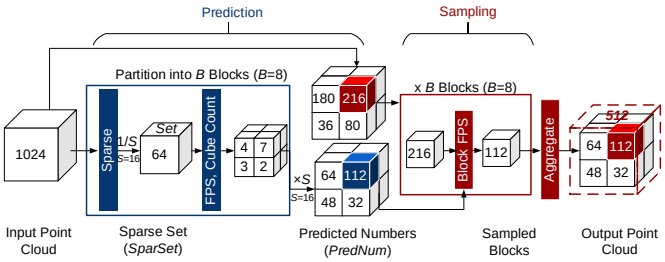

Fig. 1. The existing block-wise FPS algorithm in [15].

III. Adjustable Multi-Stream Block-Wise FPS Acceleration Framework

A. Adjustable Multi-Stream Block-Wise FPS (AMB-FPS)

The existing block-wise FPS in [15] suffers from large network accuracy loss and limited optimization space of computational complexity and parallelism. To address these challenges, an adjustable multi-stream block-wise FPS algorithm (*AMB-FPS*) is proposed to further improve parallelism, reduce complexity, and maintain accuracy. *AMB-FPS* is composed of a multi-stream prediction stage and a multi-stream sampling stage, as shown in Fig. 2.

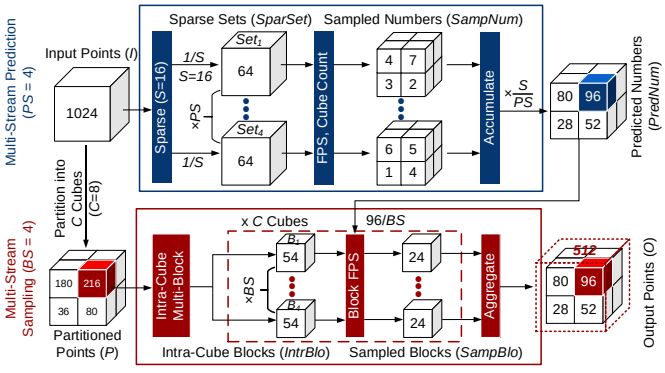

Fig. 2. Illustration of the mechanism of the proposed *AMB-FPS*.

Compared with [15], *AMB-FPS* employs three additional key parameters, namely the number of prediction streams (*PS*), the number of cubes (*C*), and the number of block streams (*BS*). The sparsity coefficient (*S*), *PS*, and *C* are used in the multi-stream prediction stage. *C* and *BS* are used in the multi-stream sampling stage.

Algorithm 1 Overall Calculation for *AMB-FPS*

Input: Input Points I, Sampled Number N_O, C, S, PS, BS
Output: Output Points O

// Multi-Stream Prediction
$PredNum = []$
for ps in PS **do**
  $SparSet = Sparse(I, S_{ps})$
  $SampNum = CubeCount(FPS(SparSet, N_O/S))$
  $PredNum.append(SampNum)$
$PredNum.aggregate()$
// Multi-Stream Sample
$SampBlo = []$
$P = CubePartition(I, C)$
for c in C **do**
  $IntrBlo = BlockPartition(P[c], BS)$
  for bs in BS **do**
    $temp = FPS(IntrBlo[bs], PredNum[c]/BS)$
    $SampBlo[c].append(temp)$
  $SampBlo[c].aggregate()$
$O = SampBlo.toPointCloud()$
return O

Algorithm 1 outlines the overall process of *AMB-FPS*. During the multi-stream prediction stage, *PS* distinct sparsity strategies are adopted to reduce the input points to *PS* small independent sparse point sets. FPS is performed on each set. Afterwards, the remained points of each set are partitioned into *C* cubes. The points are counted for each cube. The counted numbers in the same cube of *PS* sets are accumulated and then multiplied by *S/PS* to obtain the final predicted number of each cube in the original input point cloud. In the multi-stream sampling stage, the input point cloud is spatially partitioned into *C* cubes. Each cube is further uniformly sampled into *BS* blocks by adopting *BS* distinct sparsity strategies. The predicted number of each cube obtained in the first stage is also divided by *BS*. FPS is performed on each block. The remained points of all blocks are aggregated to obtain the final sampled points. An example of *AMB-FPS* configured with $C = 8$, $S = 16$, and $PS = BS = 4$ is illustrated in Fig. 2. The number of points before and after FPS is 1024 and 512, respectively. In the prediction stage, the introduction of *PS* aims to mitigate the imprecision caused by the loss of key points due to a larger *S*. *C* preserves the global features. *BS* improves the computational parallelism and reduces the load imbalance across *C* cubes. Therefore, reasonable *PS* and *BS* are beneficial for improving accuracy and reducing latency, respectively. The effect of the three new introduced parameters *PS*, *C*, and *BS* is also validated by experiments in Fig. 4 and discussed in detail in Section IV-A. However, a large *PS* causes excessive computation during the prediction stage, while an exaggerated *BS* makes FPS resemble random sampling. To address these challenges, in Section III-C, a rapid searching algorithm is proposed to select a reasonable configuration (*C*, *S*, *PS*, *BS*) for *AMB-FPS*.

B. Unified Hardware Architecture (UHA)

The unified hardware architecture of the proposed *AMB-FPS* acceleration framework is depicted in Fig. 3. *UHA* consists of a sampling module, a partition module, a memory module, a

control module, and an interface module. The sampling module includes M cores, each of which can be configured to support different numbers of input points and output points. Each core performs the FPS of each block in Fig. 2. The partition module mainly consists of 16 comparators, counters, and buffers. The partition module is responsible for partitioning coordinates of point cloud into multiple blocks and performing sparsity strategies, providing inputs for the sampling module. The memory module is composed of $M/2$ static random-access memory (SRAM) banks, each bank housing $N/M/2$ (N is the number of input points) 256-bit words. An instruction set architecture (ISA) is designed for the off-chip host to control *UHA*. During neural network inference, the off-chip host sends a set of instructions to the control module. The control module decodes the instruction and then controls the execution of the sampling module, the partition module, and the allocation of the memory module. In the hardware architecture, the number of cores M is configured based on the number of blocks of *AMB-FPS* in Fig. 2. Then, the numbers of SRAM banks and SRAM words are configured to $M/2$ and $N/M/2$, respectively. Therefore, a hardware implementation that aligns with the proposed *AMB-FPS* is generated.

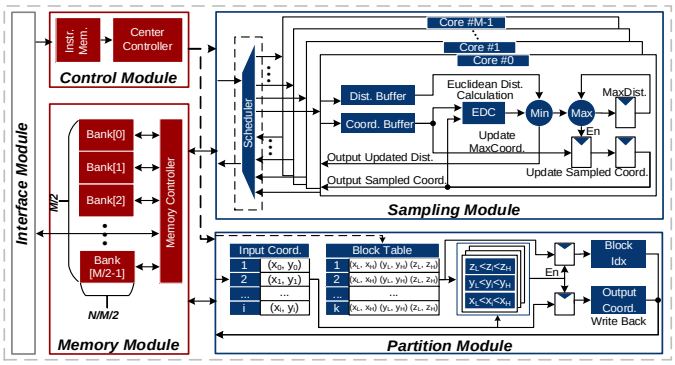

Fig. 3. The architecture of the proposed *UHA*.

C. Rapid Searching Algorithm (RSA)

All the configurations of *AMB-FPS* and *UHA* are designed for two goals: maintaining network accuracy and minimizing hardware energy consumption. However, to determine the optimal configuration, all the configurations need to be evaluated. For each configuration, replacing the original FPS by the configured *AMB-FPS*, retraining the network, and then evaluating the hardware are required. The entire process is significantly time-consuming and thus impractical. Therefore, a rapid searching algorithm (*RSA*) is proposed as shown in Algorithm 2, which rapidly generates the optimal configuration of *AMB-FPS* and hardware architecture for specified accuracy and hardware requirements.

In Algorithm 2, *Loss* is introduced as a comprehensive evaluation index of *AMB-FPS* and *UHA*, which combines the predicted algorithm accuracy (*Acc*) and hardware energy consumption (E) as shown in (5). A lower *Loss* value indicates superior *AMB-FPS* performance. The weight w is manually controlled to prioritize either accuracy or hardware energy consumption for *Loss*. In (4), *Acc* is the sum of mean absolute error (*MAE*) and an improved Mahalanobis distance (*IMD*). *MAE* and *IMD* are explained in Section IV-A. In (1), L is the latency of the two stages in *AMB-FPS*. Therefore, L is affected by the configurations of the proposed *AMB-FPS* and hardware architecture. In (2), *Power* is derived from the hardware post-synthesis simulation by linear fitting.

$$L = (R - \frac{R^2}{2}) \times (ceil\left(\frac{PS}{M}\right) \times \left(\frac{N}{S}\right)^2 + ceil(\frac{C \times BS}{M}) \times \left(\frac{N}{C \times BS}\right)^2). \quad (1)$$

$$Power = 2.9 + 1.1 \times M. \quad (2)$$

$$E = Power \times L. \quad (3)$$

$$Acc = Norm(IMD) + Norm(MAE). \quad (4)$$

$$Loss = w \times Acc + (1 - w) \times E. \quad (5)$$

In Algorithm 2, for each configuration (C, S, PS, BS, M), *Acc* and E are computed. *Loss* for each configuration is calculated according to (5) and is added to $Config_l$. Subsequently, the configuration in $Config_l$ with the smallest *Loss* is selected as the best configuration for *AMB-FPS* and hardware architecture.

Algorithm 2 Rapid Searching Algorithm

Input: w
Output: $[C, S, PS, BS, M]$
$C_l = [2, 4, 8, ...]$, $S_l = [2, 4, 8, 16, ...]$
$PS_l = BS_l = [1, 2, 3...]$, $M_l = [1, 2, 4, 8, 16, ...]$
$Config_l = []$
for C, S, PS, BS, M in $C_l, S_l, PS_l, BS_l, M_l$ **do**
  Compute *IMD*, *MAE*, *Acc*, *L*, *E*
  Compute $Loss(w)$
  $Config_l.append([C, S, PS, BS, M, Loss])$
$Config_l = Config_l.sort(Loss)$
$[C, S, PS, BS, M] = Config_l[0]$
return $[C, S, PS, BS, M]$

IV. Experimental Results

A. Analysis of Software and Hardware Implementations

To evaluate the accuracy of the proposed *AMB-FPS*, the PointNeXt-S network classifying ModelNet40 [17] dataset, PointNet++ [6] classifying ModelNet40, and PointNeXt-S segmenting S3DIS [21] dataset, are trained in PyTorch as the benchmarks. The open-source tool Distiller [18] is used for pruning, quantization to 8-bit data width, and retraining to recover the accuracy. *UHA* is designed in the TSMC 28-nm HPC+ CMOS technology. Cadence Genus is used for logic synthesis at 200 MHz and 0.9 V supply voltage. To evaluate the power consumption, Cadence NC-Sim is used to obtain the realistic switching activity of the hardware, in the form of TCF (toggle count format) files. The TCF files are then used for power evaluation with Cadence Genus.

The *MAE* and proposed *IMD* are used to evaluate the similarity of *AMB-FPS* and the original FPS. *MAE* represents the average of the absolute differences between two point clouds. *MAE* is

$$MAE = \frac{1}{n}\sum_1^n |A_i - B_i|, \quad (6)$$

where n is the total number of points in point clouds A and B. A smaller *MAE* indicates a higher similarity between the two point

clouds. Improved Mahalanobis Distance (*IMD*) is proposed to reflect the difference between two point clouds. The mean value of each dimension for the two point clouds (P_1 and P_2) is computed to derive two representative center points, which are u_1 and u_2. Subsequently, the covariance matrices of the two point clouds, namely S_1 and S_2, are calculated and accumulated to obtain an overall covariance matrix. The *IMD* between the point clouds is then computed based on

$$IMD(P_1, P_2) = \sqrt{(u_1 - u_2)^T (S_1 + S_2)^{-1} (u_1 - u_2)}\,. \quad (7)$$

A larger *IMD* value indicates a larger difference between the two point clouds. Compared to *MAE* focusing on local details, *IMD* prioritizes the global shape of point cloud. The impact of diverse configurations of *AMB-FPS* on *Acc* and *E* are depicted in Fig. 4. The output point cloud of the original FPS is employed as a baseline for computing *Acc*. A larger *C* causes a larger error in the prediction stage due to the linear decrease in the number of points in each cube. Consequently, *Acc* becomes worse, as shown in Fig. 4a. However, the reduction in points per cube induces a quadratic complexity reduction in the FPS of each cube, thus leading to a decline in *E*. As *S* increases, the sparse step is more aggressive in the prediction stage, which leads to the deterioration of *Acc*, as shown in Fig. 4b. In terms of *E*, an increase in *S* leads to fewer points performed by FPS for each set during the prediction stage in Fig. 2, thus contributing to a reduced prediction latency and resulting in a smaller *E*.

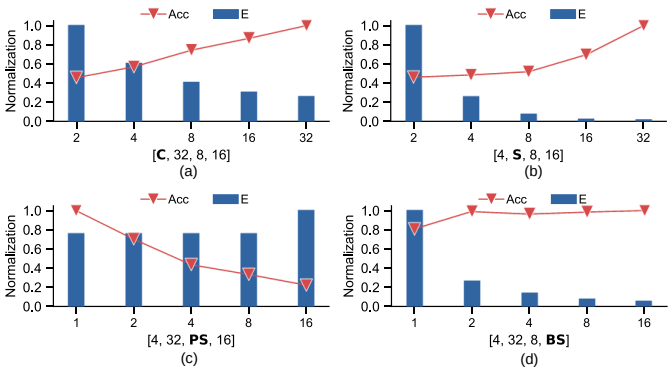

Fig. 4. The impact of different configurations [*C*, *S*, *PS*, *BS*] of *AMB-FPS* on *Acc* and *E*. (a) The numbers of cubes (*C*). (b) The sparse coefficient (*S*). (c) The numbers of prediction streams (*PS*). (d) The number of block streams (*BS*).

The parameter *PS* serves as a multi-stream prediction mechanism designed to compensate for the accuracy loss caused by aggressive *S*. A large *PS* causes excessive computation and worse *E* during the prediction stage while optimizing *Acc* in (4), as $PS \geq 8$ ($M = 8$) in this case, as shown in Fig. 4c. *BS* is used in the multi-stream sampling stage, which reduces hardware complexity by reducing the number of points in each block. Therefore, as shown in Fig. 4d, a larger *BS* leads to a smaller *E*. However, *BS* cannot be set to be excessively large, which reduces the number of points per block during the sampling stage and makes FPS resemble random sampling. Meanwhile, if only *BS* is employed (such as [*C*, *S*, *PS*, *BS*] = [1, 0, 0, 16] for *AMB-FPS*), the global features of point cloud are lost after uniformly sampling into *BS* blocks, leading to a significant network accuracy loss (>2%). Therefore, the introduction of other parameters [*C*, *S*, *PS*] is necessary to alleviate the impact of *BS*. The partition of input point cloud into *C* cubes preserves the global features. Meanwhile, *S* and *PS* enable the accurate prediction of the sampled number in each cube.

B. Comparisons with the State of the Art

To verify the effectiveness of the proposed *AMB-FPS* acceleration framework, we perform a comparative analysis with the original FPS and PNNPU [15]. For a comprehensive comparison, algorithm accuracy indicators (such as *MAE*, *IMD*, network accuracy, and mean Intersection-over-Union (mIoU)) and hardware comparative indicators (such as *L* and *E*) are employed. As benchmarks, PointNeXt-S and PointNet++ networks are used to classify ModeNet40 dataset, which is presented in Table I. To evaluate the efficiency of *AMB-FPS* on a large-scale dataset, PointNeXt-S is used to segment S3DIS dataset, which is presented in Table II.

In Tables I and II, AMB-FPS-v0 represents the hardware-prioritized version, which mainly optimizes *L* and *E* with less attention on *MAE* and *IMD*. AMB-FPS-v1 is the accuracy-prioritized version, focusing on improvements of *MAE* and *IMD* yet with hardware performance loss. In Table I, compared to the original FPS, AMB-FPS-v0 (AMB-FPS-v1) reduces *L* and *E* by 99.87% (99.8%) and 99.43% (98.5%), respectively, while ensuring negligible network accuracy loss (< 0.4%). Compared with PNNPU, AMB-FPS-v0 (AMB-FPS-v1) reduces *MAE*, *IMD*, *L*, and *E* by 40.09% (72.62%), 24.3% (68.05%), 83.87% (74.19%), and 76.06% (37.2%), respectively. In terms of network accuracy, AMB-FPS-v0 and AMB-FPS-v1 surpass PNNPU by 0.73% (0.08%) and 1.05% (0.24%) for PointNeXt-S (PointNet++), respectively. The dataset for both PointNet++ and PointNeXt-S is ModelNet40, and the sampling rate of both networks is 1/2. Therefore, the latency, power, and energy consumption are the same for running PointNet++ and PointNeXt-S. In Table II, compared to the original FPS, AMB-FPS-v0 (AMB-FPS-v1) reduces *L* and *E* by 99.9% (99.8%) and 99.5% (98.57%), respectively, with no accuracy loss in FP32 and < 1.9% accuracy loss in 8-bit quantization. Compared with PNNPU, AMB-FPS-v0 (AMB-FPS-v1) reduces *MAE*, *IMD*, *L*, and *E* by 34.92% (51.06%), 59.18% (68.45%), 86.72% (75%), and 79% (40.2%), respectively. In terms of mIoU of PointNeXt-S, AMB-FPS-v0 and AMB-FPS-v1 outperform PNNPU by 9.34% (2.76%) and 9.33% (2.91%) in FP32 (8-bit quantization), respectively. In [20], QuickFPS achieves a sampling time of 0.1 ms (3 ms) for 1 k (24 k) points at 1 GHz frequency, with estimated 100,000 (3,000,000) cycles and energy consumption of 18.6 µJ (4261.5 µJ). In Table I (Table II), AMB-FPS-v0 consumes a latency of 0.005 ms (1.401 ms) and energy of 0.09 µJ (27.265 µJ) at 200 MHz, indicating 960 (280,151) cycles. Compared to QuickFPS, the proposed AMB-FPS-v0 reduces

the number of cycles by 99.04% (90.66%) and saves energy by 99.52% (99.36%) for sampling 1 k (24 k) points.

TABLE I
PERFORMANCE COMPARISONS ON POINTNEXT-S AND POINTNET++ CLASSIFYING MODELNET40

	Original FPS	PNNPU [15] VLSI 2021	AMB-FPS-v0	AMB-FPS-v1
w : (1-w) ($/10^5$)	N/A	N/A	1:10	10:1
[*C*, *S*, *PS*, *BS*]	N/A	[16, 16, 1, 1]	[4, 32, 2, 16]	[2, 32, 16, 15]
Core (*M*)	1	16	64	30
Area (mm^2)	0.02	0.4*	1.63	0.76
SRAM (KB)	8	8	8	8
MAE	0	7.246	4.341	1.984
IMD	0	0.169	0.128	0.054
PointNeXt-S Accuracy (%) 8-bit	91.45	90.36	91.09	91.41
PointNet++ Accuracy (%) 8-bit	91.53	91.25	91.33	91.49
L (ms)	3.932	0.031*	0.005	0.008
Power (mW)	4	12.25*	18.74	28.7
E (µJ)	15.729	0.376*	0.090	0.236

*The simulation results of the re-implemented FPS hardware.

TABLE II
PERFORMANCE COMPARISONS ON POINTNEXT-S SEGMENTING S3DIS

	Original FPS	PNNPU [15] VLSI 2021	AMB-FPS-v0	AMB-FPS-v1
w : (1-w) ($/10^8$)	N/A	N/A	1:10	10:1
[*C*, *S*, *PS*, *BS*]	N/A	[16, 16, 1, 1]	[32, 32, 8, 4]	[2, 32, 16, 16]
Core (*M*)	1	16	128	32
Area (mm^2)	0.02	0.4*	3.27	0.81
SRAM (KB)	192	192	192	192
MAE	0	2.652	1.726	1.298
IMD	0	0.485	0.198	0.153
mIoU (%) FP32	62.51	53.79	63.13	63.12
mIoU (%) 8-bit	56.83	52.23	54.99	55.14
L (ms)	1350	10.547*	1.401	2.637
Power (mW)	4	12.25*	18.74	28.7
E (µJ)	5400	129.2*	27.265	77.256

*The simulation results of the re-implemented FPS hardware.

V. CONCLUSIONS

In this brief, an algorithm-hardware co-designed FPS acceleration framework is proposed to reduce energy consumption while maintaining accuracy. To improve parallelism and reduce computational complexity, an adjustable multi-stream block-wise FPS algorithm is proposed. A unified hardware architecture for the adjustable multi-stream block-wise FPS algorithm is designed. A rapid searching algorithm is developed to select the optimal configuration of the adjustable multi-stream block-wise FPS algorithm and the unified hardware architecture under specified priority between accuracy and energy savings. Benchmarked with PointNeXt-S and PointNet++ networks classifying ModelNet40 and PointNeXt-S network segmenting S3DIS, and designed in a 28-nm CMOS technology, the proposed FPS acceleration framework achieves a latency of 0.005 (1.401) ms and a frame energy consumption of 0.09 (27.265) µJ/frame for 1 k (24 k) input points at 200 MHz and 0.9 V supply voltage. Compared to the state of the art, the proposed FPS acceleration framework reduces up to 99.9% latency, saves up to 99.5% energy consumption, and improves up to 9.34% network accuracy.

REFERENCES

[1] M. Wen, Y. Dai, T. Chen, C. Zhao, J. Zhang, and D. Wang, "A robust sidewalk navigation method for mobile robots based on sparse semantic point cloud," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2022, pp. 7841-7846.

[2] Y. Cheng, J. Su, M. Jiang, and Y. Liu, "A novel radar point cloud generation method for robot environment perception," *IEEE Trans. Robot.*, vol. 38, no. 6, pp. 3754-3773, Dec. 2022.

[3] Y. Li *et al.*, "Deep learning for LiDAR point clouds in autonomous driving: a review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3412-3432, Aug. 2021.

[4] R. Abbasi, A. K. Bashir, H. J. Alyamani, F. Amin, J. Doh, and J. Chen, "Lidar point cloud compression, processing and learning for autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 1, pp. 962-979, Jan. 2023.

[5] L. Kästner, V. C. Frasineanu, and J. Lambrecht, "A 3d-deep-learning-based augmented reality calibration method for robotic environments using depth sensor data," in *Proc. Int. Conf. Robot. Automat. (ICRA)*, 2020, pp. 1135-1141.

[6] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5105-5114.

[7] X. Ma, C. Qin, H. You, H. Ran, and Y. Fu, "Rethinking network design and local geometry in point cloud: A simple residual MLP framework," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2022.

[8] G. Qian *et al.*, "PointNeXt: Revisiting PointNet++ with improved training and scaling strategies," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Nov. 2022, pp. 23192-23204.

[9] H. Zhao, L. Jiang, C. W. Fu, and J. Jia, "Pointweb: Enhancing local neighborhood features for point cloud processing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5565-5573.

[10] L. Li, L. He, J. Gao, and X. Han, "PSNet: Fast data structuring for hierarchical deep learning on point cloud," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6835-6849, Oct. 2022.

[11] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "Pointcnn: Convolution on x-transformed points," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, 2018.

[12] W. Wu, Z. Qi, and L. Fuxin, "Pointconv: Deep convolutional networks on 3d point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 9621–9630.

[13] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 16 259–16 268.

[14] Y. Zhang, Q. Hu, G. Xu, Y. Ma, J. Wan, and Y. Guo, "Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 18953–18962.

[15] S. Kim, J. Lee, D. Im, and H. -J. Yoo, "PNNPU: A 11.9 TOPS/W high-speed 3D point cloud-based neural network processor with block-based point processing for regular DRAM access," in *Proc. Symp. VLSI Circuits*, Jun. 2021, pp. 1-2.

[16] J. Li, J. Zhou, Y. Xiong, X. Chen, and C. Chakrabarti, "An adjustable farthest point sampling method for approximately-sorted point cloud data," in *Proc. IEEE Int. Workshop Signal Process. Syst. (SiPS)*, Oct. 2022, pp. 1-6.

[17] Z. Wu *et al.*, "3d shapenets: A deep representation for volumetric shapes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1912-1920.

[18] Distiller. Accessed: Sep. 14, 2023. [Online]. Available: https://github.com/IntelLabs/disiller.

[19] X. Yang, T. Fu, G. Dai, S. Zeng, K. Zhong, K. Hong, and Y. Wang, "An efficient accelerator for point-based and voxel-based point cloud neural networks," in *Proc. ACM/IEEE Design Automation Conf. (DAC)*, Jul. 2023, pp. 1-6.

[20] M. Han *et al.*, "QuickFPS: Architecture and algorithm co-design for farthest point sampling in large-scale point clouds," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 42, no. 11, pp. 4011-4024, Nov. 2023.

[21] I. Armeni *et al.*, "3D semantic parsing of large-scale indoor spaces," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1534-1543.