

RISC-V R-Extension: Advancing Efficiency with Rented-Pipeline for Edge DNN Processing

Won Hyeok Kim
Department of Artificial Intelligence
Sungkyunkwan University
Suwon, Korea
2080fresh@skku.edu

Hyeong Jin Kim
Department of Semiconductor
Convergence Engineering
Sungkyunkwan University
Suwon, Korea
wsx0409@g.skku.edu

Tae Hee Han
Department of Semiconductor
Systems Engineering
Sungkyunkwan University
Suwon, Korea
than@skku.edu

Abstract—The proliferation of edge devices necessitates efficient computational architectures for lightweight tasks, particularly deep neural network (DNN) inference. Traditional NPUs, though effective for such operations, face challenges in power, cost, and area when integrated into lightweight edge devices. The RISC-V architecture, known for its modularity and open-source nature, offers a viable alternative. This paper introduces the RISC-V R-extension, a novel approach to enhancing DNN process efficiency on edge devices. The extension features rented-pipeline stages and architectural pipeline registers (APR), which optimize critical operation execution, thereby reducing latency and memory access frequency. Furthermore, this extension includes new custom instructions to support these architectural improvements. Through comprehensive analysis, this study demonstrates the boost of R-extension in edge device processing, setting the stage for more responsive and intelligent edge applications.

Index Terms—Architectural pipeline register, Custom instruction set architecture, DNN Acceleration, Lightweight edge devices, Rented-pipeline, RISC-V

I. INTRODUCTION

In the rapidly evolving landscape of computing, the efficiency of processing units is paramount, especially for small-edge devices like home appliances and IoT devices. These devices require deep neural network (DNN) inference workloads that adhere to strict power and space limitations. The RISC-V architecture, known for its open-source and modular nature, is a promising candidate for these challenges.

Central to edge device functionality is the processing of DNNs. Recently, NPUs and TPUs have been the mainstay for such tasks in smartphone-size devices. These specialized units offer optimized performance for neural network computations but are often impractical for integration into smaller, more resource-constrained edge devices due to their size, power, and cost. Consequently, there is a growing need for efficient CPU-based solutions, especially for devices that cannot accommodate separate neural processing hardware.

Previous advancements in RISC-V for DNN processing include F-extension for floating-point (FP) processing and multiply-accumulate (MAC) instructions. However, this research introduces R-extension, which improves upon these earlier developments. With its unique rented-pipeline mechanism and use of architectural pipeline registers (APR), the R-extension is tailored to enhance the processing of DNNs in

edge devices. The main contributions of R-extension are as follows:

- Rented-pipeline: Rent a memory (MEM) stage on the 5-stage pipeline when fetching proposed instructions. Renting a MEM stage allows the processor to consume two stages during execution without increasing the number of stages or critical path delay.
- APR (Architectural Pipeline Register): Located on the MEM/write-back (WB) pipeline register, accumulated results are constantly updated in the APR. By using APR, the processor does not need to use memory for read and write during MAC operations. When accumulation ends, the processor writes back APR data into the destination register and resets APR.
- Custom instructions: This paper represents two new instructions, *rflmac.s* and *rfsmac.s*, to support the previous two contributions. *rflmac.s* stands for single-precision FP MAC operation on R-extension, and *rfsmac.s* stands for store result of single-precision FP MAC operation working on R-extension.
- Versatility: The proposed microarchitecture facilitates the addition of further instructions for diverse accumulation operations, such as integer MAC operation and difference accumulator [1], enhancing its alterability. The potential integration into vector architectures also paves the way for increased parallelism, offering prospects for system speed enhancements.

The performance metrics of the R-extension are noteworthy. Compared to the F-extension's FP multiplication approach (RV64F), our R-extension (RV64R) achieves an increment of 29% in instructions-per-cycle (IPC) and a 34% decrement in the number of memory accesses. Furthermore, compared to the baseline that solely added a MAC instruction without altering the pipeline, RV64R exhibits 15% IPC and 22% memory access improvement with a negligible increase in area.

Advancements in R-extension can improve edge device processing for lightweight AI tasks, achieving remarkable efficiency and adaptability. This paper will delve into the details of the R-extension and its comparative advantages.

```

// H = Hin - Hfil + 1
// W = Win - Wfil + 1
for (i = 0; i < M; i++)
  for (j = 0; j < H; j += S)
    for (k = 0; k < W; k += S)
      for (l = 0; l < C; l++)
        for (m = 0; m < Hfil; m++)
          for (n = 0; n < Wfil; n++)
            Output[i][j/S][k/S] +=
              Input[l][j+m][k+n]*Filter[i][l][m][n];

```

(a-1)

```

000 <convolve>:
000: j 182      0ca: fmul.s fa5,fa3,fa5
006: j 16a     0e4: fadd.s fa5,fa4,fa5
00c: j 152     0fc: fsw fa5,0(a5)
012: j 13a     114: bge a5,a4,020
018: j 122     12c: bge a5,a4,01a
01e: j 10a     144: bge a5,a4,014
04a: flw fa4,0(a5) 15c: bge a5,a4,00e
08e: flw fa3,0(a5) 174: bge a5,a4,008
0c6: flw fa5,0(a5) 18c: bge a5,a4,002

```

(a-2)

```

for (i = 0; i < M; i++)
  for (j = 0; j < H; j += S)
    for (k = 0; k < W; k += S)
      for (l = 0; l < C; l++)
        for (m = 0; m < Hfil; m++)
          for (n = 0; n < Wfil; n++)
            fmac_s(Output[i][j/S][k/S],
              Input[l][j+m][k+n],Filter[i][l][m][n]);

```

(b-1)

```

000 <convolve>:
000: j 154      0ca: fmac.s fa5,fa4,fa3
006: j 13c     0ce: fsw fa5,0(a3)
00c: j 124     0e6: bge a5,a4,020
012: j 10c     0fe: bge a5,a4,01a
018: j 0f4     116: bge a5,a4,014
01e: j 0dc     12e: bge a5,a4,00e
08a: flw fa4,0(a5) 146: bge a5,a4,008
0c2: flw fa3,0(a5) 15e: bge a5,a4,002
0c6: flw fa5,0(a3)

```

(b-2)

```

for (i = 0; i < M; i++)
  for (j = 0; j < H; j += S)
    for (k = 0; k < W; k += S)
      for (l = 0; l < C; l++)
        for (m = 0; m < Hfil; m++)
          for (n = 0; n < Wfil; n++)
            rfmac_s(Input[l][j+m][k+n],
              Filter[i][l][m][n]);
            rfsmac_s(Output[i][j/S][k/S]);

```

(c-1)

```

000 <convolve>:
000: j 154      09c: rfmac.s fa5,fa4
006: j 13c     0b4: bge a5,a4,020
00c: j 124     0cc: bge a5,a4,01a
012: j 0da     0e4: bge a5,a4,014
018: j 0c2     0fe: rfmac.s fa5
01e: j 0aa     116: fsw fa5,0(a5)
060: flw fa5,0(a5) 12e: bge a5,a4,00e
098: flw fa4,0(a5) 146: bge a5,a4,008
15e: bge a5,a4,002

```

(c-2)

Fig. 1: (a) RV64F, (b) Baseline, (c) RV64R. Left: Convolution code, H_{in} : Input Height, W_{in} : Input Width, M : Number of Filter, C : Channel, H_{fil} : Filter Height, W_{fil} : Filter Width. Right: Assembly language after compile. Highlighted parts are the main instructions in the most inner of all loops.

II. BACKGROUND

A. Advanced Pipeline

The challenge in processing DNN inference on CPUs, particularly in lightweight devices, is managing frequent memory usage, which is essential to maintaining process capability within the power constraints of small-scale devices. In the context of RISC-V implementations for such devices, complex pipelining techniques like those used in superscalar processors, generally optimized for high performance CPUs, are unsuitable due to their high power consumption and area costs.

MAC operations, characterized by repetitive multiply and add sequences, are distinct from complex operations requiring diversified pipelines. These diversified pipelines, which are effective for handling extended latency operations, are not optimized for the continuous accumulation required in MAC

operations and also involve higher area costs and power consumption because of more pipeline stages, which are critical constraints in small devices [2]. In contrast, using rented-pipelining, the R-extension for RISC-V offers an accelerated solution for compact devices using the execution (EX) stage on multiplication and the MEM stage on accumulation. This approach is bespoke for MAC operations, which are fundamental in DNN inferences.

In MAC operations within assembly sequences shown at Figure 1(a-2), read-after-write (RAW) hazards occur when the processor attempts to accumulate new multiplication results with data not yet updated from a previous WB stage. Conventional data forwarding methods, which preempt RAW hazards by bypassing register write and read cycles, may not suffice for MAC scenarios where accumulated data is pending in the write-back stage and not immediately available for forwarding.

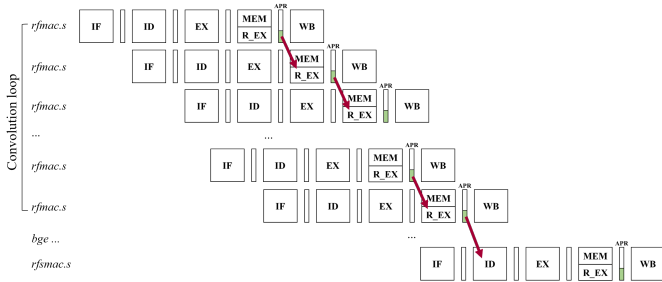


Fig. 2: Forwarding effect on MAC at R-extension.

TABLE I: Format field encoding.

fmt field	Mnemonic	Meaning
00	S	32-bit single-precision
01	D	64-bit double-precision
10	H	16-bit half-precision
11	Q	128-bit quad-precision

The use of APR in the R-extension enables efficient data forwarding for MAC operations described at Figure 2, reducing memory access. This approach contributes to DNN acceleration and power consumption reduction [3], making the R-extension suitable for energy-sensitive applications in lightweight devices.

B. RISC-V F Standard Extension

The compiler utilizes the F-extension when computing convolution in original RISC-V architectures with FP parameters [4]. This extension handles single-precision FP convolution operations, primarily through *fadd.s* (FP add) and *fmul.s* (FP multiply) instructions shown at Figure 1(a). These instructions are integral to the computational processes underlying DNNs.

As shown in Figure 3, in addition to requiring two types of source registers; a destination register and source register – which are standard for most RISC-V instructions – the FP operation instructions also necessitate fields to specify the format *fmt* and rounding mode *rm*. The format, encoded as a 2-bit field, indicates the precision of the FP. Table I shows the format field encoding. In the case of the F-extension, this field is set to ‘00’, corresponding to the mnemonic S for single-precision operations. Similarly, the D-extension, intended for double-precision, uses a different mnemonic, D.

This study introduces the R-extension, which also targets single-precision FP operations. It adopts the ‘00’ format field convention from the F-extension, ensuring compatibility and facilitating integration into the existing RISC-V infrastructure. The rounding mode, crucial for the precision of computations, is defined in the control and status register (CSR), which guarantees adherence to the IEEE 754 standard [5] and maintains the necessary precision for the given application.

Our study proposes the R-extension, a novel approach tailored for DNN workloads building upon the capabilities of the F-extension. The R-extension introduces custom instructions, specifically *rfmac.s* and *rfsmac.s*, designed to reduce the instruction count (IC) and memory usage during DNN operations.

C. Customized Instruction

Various efforts have focused on optimizing DNN processing through specialized instructions within the RISC-V architecture’s evolving landscape. One such example is the development of a Winograd-based convolution acceleration instruction [6], which performs a convolution calculation between a 4×4 input matrix and a 4×4 convolution kernel matrix and generates a 2×2 output matrix. While this custom instruction, *conv23*, enhances the efficiency of matrix convolutions for specific convolutional neural network (CNN) algorithms, it is restricted in its applicability. It does not tackle the core issue of MAC operations that are heavily utilized not only in convolution and fully connected layers of CNNs but also in other DNN architectures. This emphasizes the requirement for more adaptable solutions catering to a wide range of DNN architectures and operations.

Further research introduces *vmac* (MAC operation on RISC-V V-extension [4]) and *vload* (vector register load) instructions, mainly through enhancements in dot product computations and data transfer efficiency [7]. By supporting single instruction, multiple data (SIMD) architecture, these extensions substantially enhance the computational performance of DNN tasks. However, *vmac* is only for paralleling MAC operations; it does not improve memory usage. We transformed *vmac* to work in an FP scalar process for an experiment, and the instruction is called *fmac.s* with usage as Figure 1(b), our baseline that implements a MAC module at the EX stage and is compared with the R-extension.

The R-extension represents not merely an optimization for scalar processor environments but also an expandable entity within the RISC-V architecture. Harmonizing with the established architectural framework offers a viable extension path toward SIMD V-extension integration. This versatility underscores the extension’s potential to evolve with advancing vector processors, positioning the RISC-V architecture to meet the expanding computational demands.

III. METHODOLOGY

The specialized pipeline came out for MAC operation, and to support this architecture, there are two types of instructions.

A. Instruction Set Architecture (ISA)

To add custom instructions on RISC-V ISA to compile, we should first assign the instruction symbol, type, constituents, MASK, and MATCH. MASK is the 32-bit binary to filter out the opcode and functions(e.g., *funt3*, *funct5*, *funct7*, etc). MATCH is the 32-bit binary that contains actual opcodes and functions for each designated seat on instruction. After selecting the components for compilation, we need to add the required format to *riscv-gnu-toolchain* [8], then reconfigure and build the compiler.

Figure 3 presents two innovative instructions within the RISC-V R-extension. The instruction *rfmac.s* requires two source registers, *rs1* and *rs2*, and facilitates the multiplication of these sources, accumulating the result in the APR. This process eliminates the need for an immediate write-back to a

	31	27	26	25	24	20	19	15	14	12	11	7	6	0
	funct5	fmt	rs2	rs1	rm	rd	opcode							
<i>fmul.s</i>	FMUL (0x2)	S (0x0)	src2	src1	RM	dest	OP-FP (0x14)							
<i>fmac.s</i>	FMAC (0xC)	S (0x0)	src2	src1	RM	dest	OP-FP (0x14)							
<i>rfmac.s</i>	RFMAC (0xD)	S (0x0)	src2	src1	RM	-	OP-FP (0x14)							
<i>rfsmac.s</i>	RFSMAC (0xE)	S (0x0)	-	-	RM	dest	OP-FP (0x14)							

Fig. 3: Instruction format of F-extension (*fmul.s*), Baseline (*fmac.s*), and R-extension (*rfmac.s*, *rfsmac.s*).

	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0		
	funct5	fmt	rs2	rs1	rm	rd	opcode				quad																							
<i>fmul.s</i>	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
MASK_Fmul.S	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
MATCH_Fmul.S	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	1
<i>fmac.s</i>	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
MASK_Fmac.S	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
MATCH_Fmac.S	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	1
<i>rfmac.s</i>	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
MASK_RFmac.S	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
MATCH_RFmac.S	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	1
<i>rfsmac.s</i>	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
MASK_RFsmac.S	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
MATCH_RFsmac.S	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	1

Fig. 4: MASK and MATCH of F-extension, Baseline, and R-extension.

destination register *rd*, as the intermediate result is stored in the APR for further accumulation. As illustrated in Figure 1(c-1), the *rfsmac.s* instruction, employed at the end of the convolution loop where the final result is held in the APR, is tasked with writing this result from the APR to the *rd* during the instruction-decode (ID) stage and resetting the APR at the MEM stage, as illustrated in Figure 5.

The opcode for these instructions adheres to the F-extension to maintain consistency within FP operations. Unique functions mapped in Figure 4 are assigned to each instruction to ensure no overlap with existing instructions. Newly set MASK and MATCH leave only the essential parts, such as *rm*, *rs1*, *rs2*, and *rd*.

B. Microarchitecture

The letter R on R-extension stands for the rented-pipelining concept. At the original pipeline stages, instruction-fetch (IF), ID, EX, MEM, and WB, each stage has its exclusive role. However, *fmac.s* instruction of baseline in Figure 5 does not use the MEM stage and just waits for one cycle, and data instantly goes to the WB stage. It is a waste of pipeline stage and area cost on DNN operation.

The rented-pipeline maintains the performance and number of pipeline stages even though it optimizes accumulation operation into the RISC-V CPU. In Figure 5, an extra accumulator in the MEM stage named the rented execution stage (R_EX) is turned on instead of memory read and write when introduced instructions are fetched. Using EX and R_EX stages, two of the five pipelines are used for execution without area dissipation.

APR is focused on reducing memory usage by eliminating unnecessary *flw* (FP load) and *fsw* (FP store) instructions during DNN operation as compared between Figure 1(a-2), (b-2), and (c-2). In Figure 5, APR is located at the MEM/WB pipeline register. It only needs an extra 32-bit register to

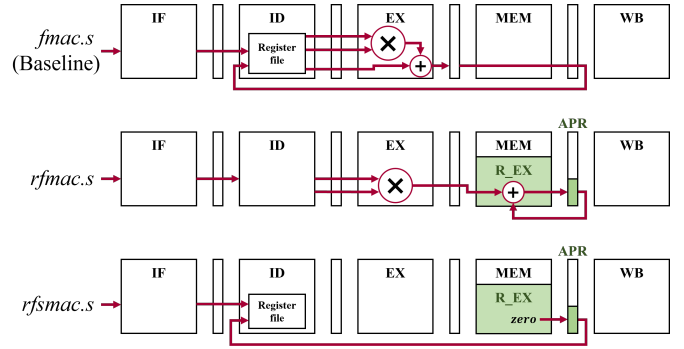


Fig. 5: Dataflow of Baseline and R-extension with R_EX and APR.

keep updating the partial sum from the accumulator at the R_EX stage. The input to the APR is determined by a MUX that selects between the accumulated data when the *rfmac.s* instruction is executed and *zero* for initialization when the *rfsmac.s* instruction is issued. Furthermore, the output of APR is interfaced with the R_EX stage to facilitate accumulation with newly multiplied data and with the ID stage, which is governed by the control bit determined by the presence of *rfsmac.s* instruction. With these two hardware contributions, rented-pipeline and APR of R-extension, the CPU can accelerate DNN inference with low power and area overhead avoidance.

IV. EXPERIMENTAL RESULTS

A. Simulation Result

To use R-extension instructions in C language, we should use inline assembly 'asm' and 'volatile' like Figure 6. After compiling with customized *riscv-gnu-toolchain*, the binary file containing R-extension instructions came out. Highlights on the right side of Figure 1 are the main instructions that are more repeated than others. F-extension has six instructions with four memory loads, one memory store, and two arithmetic operations. Baseline, with little advanced, has five instruc-

```
#define rfmac_s(rs1, rs2) \
    asm volatile ("rfmac.s %0, %1" : : "f"(rs1)
    , "f"(rs2))
#define rfsmac_s(rd) \
    asm volatile ("rfsmac.s %0" : "=f"(rd))
```

Fig. 6: Definitions of inline assembly of C language. These definitions are used in Figure 1.

TABLE II: Configuration of simulation on gem5. All experiments are done in the same configuration.

CPU	1 GHz, Single-core, In-Order
L1 Instruction Cache	512 KB, 2-way, 64 B line, 2 cycle latency
L1 Data Cache	512 KB, 2-way, 64 B line, 2 cycle latency
Memory	2 GB, DDR3, 1600 MT/s, 8 banks, 8-bit wide data

TABLE III: Comparison of performance factors by LeNet, Resnet-20, MobileNet-V1(Scaled).

		runtime (Second)	IC	IPC	memtype instructions	L1 cache overall access
LeNet	RV64F	0.066	44,310,154	0.666	19,288,578	23,071,838
	Baseline	0.048	35,792,547	0.740	16,043,778	19,841,884
	RV64R	0.032	27,010,675	0.847	12,045,594	15,449,482
	Enhancement Over RV64F	52.05 %	39.04 %	27.13 %	37.55 %	33.04 %
	Enhancement Over Baseline	34.05 %	24.54 %	14.43 %	24.92 %	22.14 %
ResNet20	RV64F	6.210	4,103,496,569	0.661	1,795,154,166	2,103,847,934
	Baseline	4.413	3,246,429,938	0.736	1,468,652,534	1,736,203,748
	RV64R	2.691	2,352,965,745	0.874	1,062,330,923	1,289,180,424
	Enhancement Over RV64F	56.66 %	42.66 %	32.30 %	40.82 %	38.72 %
	Enhancement Over Baseline	39.02 %	27.52 %	18.85 %	27.67 %	25.75 %
MobileNet V1	RV64F	7.035	4,923,965,486	0.700	2,130,037,330	2,599,414,994
	Baseline	5.255	4,122,177,959	0.784	1,824,588,370	2,222,467,107
	RV64R	3.720	3,307,689,859	0.889	1,453,124,800	1,813,851,904
	Enhancement Over RV64F	47.12 %	32.82 %	27.04 %	31.78 %	30.22 %
	Enhancement Over Baseline	29.21 %	19.76 %	13.34 %	20.36 %	18.39 %
Overall	Enhancement Over RV64F	51.94 %	38.18 %	28.82 %	36.72 %	33.99 %
	Enhancement Over Baseline	34.09 %	23.94 %	15.54 %	24.32 %	22.09 %

tions with three memory loads, one memory store, and one operation. R-extension reduces half of the memory-related instructions compared with the previous two architecture and arithmetic operation instructions. This condition leads to DNN acceleration with low area cost and power consumption described below.

Before using the gem5 simulator [9], we define the processor execution behavior in the gem5 instruction decoding algorithm. We make gem5 to accumulate data into the internal register APR when the opcode and function of *rfmac.s* are detected. In the case of *rfsmac.s*, the APR data is stored in the destination register and reset to *zero*.

As illustrated in Table II, we configured a system with separate L1 instruction and data caches, each with specific parameters such as size, associativity, and latency. The total cache size is set to 1 MB, operating under a 1 GHz clock domain. This setting was chosen to align with ARM Cortex-A55 specifications [10], which is usually used for home appliances SoC, such as smart TVs, representing the most demanding DNN inference scenarios. Such alignment ensures that our simulation environment realistically reflects the performance requirements of these applications. This setup utilizes a one-level cache architecture to precisely monitor the CPU’s memory access patterns via the L1 cache, providing insights for understanding overall system performance.

We benchmark DNN inference on edge devices using three models: LeNet for simple tasks, ResNet-20 for moderate complexity, and MobileNet-V1 for advanced applications, assessing a range from basic to complex AI capabilities [11]–[13]. RV64F is a RISC-V 64-bit F-extension that contains FP multiplication and addition, as well as FP load and storage. The baseline architecture is RV64F with a naïve MAC operation module integrated into the EX stage, and *fmac.s* instruction activates the module. We evaluate five major results: simulation runtime, IC, IPC, number of instructions for memory operations, and number of L1 cache accesses.

Table III shows that the enhancement rates are expressed in percentages, highlighting the performance gains of RV64R over RV64F and baseline. Across all models, RV64R demon-

TABLE IV: Implementation result of Baseline and RV64R with overhead.

	Baseline	RV32R	Overhead
LUT	1587	1559	-1.76 %
FF	1965	1997	1.63 %
I/O	357	357	0 %

strates substantial enhancements in all results compared to RV64F and baseline. This indicates that the RV64R architecture is more efficient and faster than RV64F and the baseline configuration.

The overall advancement rates compile the gains across all three neural network models, providing a comprehensive view of the performance boost. The RV64R architecture has advanced by about 50% over RV64F and about 32% over the baseline in terms of runtime, marking a significant increase in speed. The progress is also noteworthy for other metrics, especially regarding IPC and L1 cache accesses, which are vital for rapid and efficient processing.

B. Implementation Result

To explore the resource overhead of our R-extension, we use the Xilinx Vivado implementation 2023 tool to synthesize and implement the R-extension core and baseline core; F-extension with Naïve MAC operation into the xcvu095-ffva2104-2-e FPGA device. The Vivado tool recommends synthesizing FP IP (FP multiplier, FP adder) into DSP modules. However, we changed the options to synthesize them into LUT for comparative clarity. The resource utilization results are shown in Table IV. Compared with the baseline, the R-extension core overhead is 1.76% less LUT and 1.63% more FF resources. This result is primarily due to the R-extension’s design, which involves adding only a few multiplexers (MUXs) and altering the position of the accumulator, thus minimizing changes to the overall architecture. It is a tiny cost effect on changing the F-extension into the proposed R-extension, and it is worth considering the improvements in the upper result.

V. CONCLUSION

The research presented introduces advancements in RISC-V architecture through R-extension (RV64R). This extension, which incorporates novel elements like the rented-pipeline and APR, enhances the efficiency of MAC operations, a critical aspect for processing in neural network models. By integrating the new instructions *rfmac.s* and *rfsmac.s*, the RV64R architecture demonstrates performance improvements over the F-extension and the baseline architecture in neural network inference. This is evident in metrics such as benchmark runtime, IPC, and L1 cache accesses, where R-extension outperforms its counterparts, indicating a more efficient and faster processing capability. Notably, reducing cache accesses also leads to power consumption savings, a critical factor in designing energy-efficient computing systems [14].

Moreover, the research delves into the practical aspects of implementing the R-extension. Utilizing the Xilinx Vivado tool for RV64R synthesis and implementation, the study highlights the minimal resource overhead in transitioning from the baseline to the proposed R-extension. The slight increase in FF resources and the decrease in LUT usage underline the feasibility and efficiency of the R-extension. Considering the substantial performance gains in processing speed and efficiency, particularly in neural network models, and the added advantage of reduced power consumption through decreased cache access, the RV64R emerges as a valuable contribution to the RISC-V ecosystem, paving the way for efficient computing solutions in various applications.

ACKNOWLEDGMENT

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2019-0-00421, AI Graduate School Support Program(Sungkyunkwan University)); in part by the Competency Development Program for Industry Specialists of the Korean Ministry of Trade, Industry and Energy (MOTIE), operated by Korea Institute for Advancement of Technology (KIAT) (No. P0023704, Semiconductor-Track Graduate School(SKKU)); in part by the Ministry of Trade, Industry and Energy (MOTIE) under Grant 20011074; in part by the Technology Innovation Program (or Industrial Strategic Technology Development Program-Public-Private Joint Investment Advanced Semiconductor Talent Development Project) (RS-2023-00237136, Development of CXL/DDR5-based memory subsystem for AI accelerators) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea)(1415187686)

REFERENCES

- [1] RISC-V, "RISC-V 'P' Extension Proposal, Version 0.9.11-draft-20211209," 2021. [Online]. Available: <https://github.com/riscv/riscv-p-spec/blob/master/P-ext-proposal.pdf>.
- [2] N. Bhatt and A. Thakkar, "An efficient approach for low latency processing in stream data," *PeerJ Computer Science*, vol. 7, e426, 2021.
- [3] T. Vogelsang, "Understanding the Energy Consumption of Dynamic Random Access Memories," 2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture, Atlanta, GA, USA, 2010, pp. 363-374, doi: 10.1109/MICRO.2010.42.

- [4] A. Waterman, Y. Lee, D. A. Patterson, and K. Asanović, "The RISC-V Instruction Set Manual Volume I: Unprivileged ISA," Document Version 20191213, Dec. 2019, [Online]. Available: <https://riscv.org/specifications/>.
- [5] "IEEE Standard for Floating-Point Arithmetic," in *IEEE Std 754-2008*, vol., no., pp.1-70, 29 Aug. 2008, doi: 10.1109/IEEESTD.2008.4610935.
- [6] S. Wang, J. Zhu, Q. Wang, C. He and T. T. Ye, "Customized Instruction on RISC-V for Winograd-Based Convolution Acceleration," 2021 IEEE 32nd International Conference on Application-specific Systems, Architectures and Processors (ASAP), NJ, USA, 2021, pp. 65-68, doi: 10.1109/ASAP52443.2021.00018.
- [7] X. Yu, Z. Yang, L. Peng, B. Lin, W. Yang and L. Wang, "CNN Specific ISA Extensions Based on RISC-V Processors," 2022 5th International Conference on Circuits, Systems and Simulation (ICCSS), Nanjing, China, 2022, pp. 116-120, doi: 10.1109/ICCSS55260.2022.9802445.
- [8] "RISC-V GNU Toolchain", [online] Available: <https://github.com/riscv/riscv-gnu-toolchain>.
- [9] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M. D. Hill, and D. A. Wood, "The gem5 simulator," *SIGARCH Comput. Archit. News*, vol. 39, no. 2, pp. 1-7, May 2011, doi: 10.1145/2024716.2024718.
- [10] ARM, "Cortex-A55," ARM Developer. [Online]. Available: <https://developer.arm.com/Processors/Cortex-A55>.
- [11] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.
- [12] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.
- [13] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv preprint arXiv:1704.04861*, 2017.
- [14] V. Sze, Y. -H. Chen, T. -J. Yang and J. S. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," in *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295-2329, Dec. 2017, doi: 10.1109/JPROC.2017.2761740.