A Configurable Simulator for Neural Network Processors

Jian Hu Jian Hu (✉ jian.hu@live.com)

Beihang University

Xianlong Zhang

Beihang University

Xiaohua Shi

Beihang University

Manuscript

Keywords: Neural networks, micro architecture, simulator, compiler

Posted Date: February 10th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-202912/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Deep learning has achieved competing results comparing with human beings in many fields. Traditionally, deep learning networks are executed on CPUs and GPUs. In recent years, more and more Neural Network accelerators have been introduced in both academia and industry to improve the performance and energy efficiency for deep learning networks. In this paper, we introduce a flexible and configurable functional NN accelerator simulator, which could be configured to simulate u-architectures for different NN accelerators. The extensible and configurable simulator is helpful for system-level exploration of u-architecture, as well as operator optimization algorithm developments. We also integrated the simulator into the TVM compilation stack as an optional back-end. Users can use TVM to write operators and execute them on the simulator. The simulator is going to be open sourced.

1. Introduction

Deep learning has been applied to image recognition, object detection, speech recognition and other fields. In some fields, deep learning has even achieved competing results compared to humans. Traditionally, CPUs and GPGPUs are widely used to execute neural networks (NNs), but more and more hardware accelerators have been introduced to improve the performance and energy efficiency of NNs computing.

Tensor Processing Unit (TPU) [11] is a custom ASIC deployed in Google's data center in 2015, which has been used to speed up the inference phase of NNs. TPU uses a 8-bit integer systolic matrix multiplier to do inference, it contains on-chip buffers with a total capacity of nearly 30 MiB. TPU supports MLP, RNN and CNN, and its inference performance is 15X-30X times faster than the NVIDIA K80 GPU and Intel E5-2699 CPU. The DianNao series, including DianNao [3], DaDianNao, ShiDianNao and PuDianNao, were introduced since 2014 by ICT, CAS. DaDianNao [6] is a multi-chip deep learning inference and training architecture, and supports both 16-bit and 32-bit fixed point computing. DaDianNao supports convolution, pooling, classifier and LRN layers, and when using a 64-node architecture, it achieves more than 2000x accelerations in convolution computation compared to GPU baselines. ShiDianNao [8] focuses on accelerating convolution operations in embedded applications, and supports pooling, classification, and normalization layers as well. ShiDianNao uses inter-PE data propagation to reduce memory access in convolution, which makes it high energy efficiency. PuDianNao [15] is a polyvalent ML accelerator, and focuses on accelerating ML techniques besides DNN (deep neural networks), such as SVM, k-means, classification tree, and etc. Cambricon [16] is the first instruction set architecture (ISA) for deep learning. Inspired by RISC ISA principle, rather than mapping each NN layer to a single instruction, Cambricon ISA decomposes NN layers into basic computations, and maps those computations to simple instructions. By so, users can further assemble new NN layers from those basic computations, which makes Cambricon ISA more flexible than its predecessors. They also implemented a prototype accelerator of Cambricon ISA, which achieved the same level of performance as DaDianNao in the experiments.

Some studies [2][17][18][28] have focused on designing ultra-low-power, high power-efficiency CNN accelerators for IOT devices. Thinker-II [28] is an energy efficiency reconfigurable DNN processor for IOT devices that uses binary/ternary weights to do calculations, it applies three techniques to improve energy efficiency and achieves 19.9TOPS/W power efficiency at a power consumption of 10mW. There are also DNN processors that use in-memory processing technologies, such as Neurocube [13], Prime [7], ISAAC [24], and PipeLayer [26]. They improve energy efficiency by reducing data movement between memory and processing units, the latter 3 use analog arithmetic for matrix calculations. Eyeriss [5] improves energy efficiency by reducing data movement through reconfigurable computation mapping, with a processing dataflow called row stationary, it also gates zero neurons to save power. Some NN processors improve performance and energy efficiency by taking advantage of the sparsity of NN models, including Cambricon-X [29], EIF [10], SCNN [21], etc. Cnvlutin [1] improves performance and energy efficiency by eliminating ineffectual operations in DNN, it's targeted at lower sparsity between 40–50\% zeroes. Stripes [12] is an NN accelerator that improves performance and energy efficiency by leveraging bit-serial computing units. There are studies which exploits 3D memory for NN accelerator, such as TETRIS [9], 3D memory allows it to use more area for processing units, and simplifies dataflow scheduling. Some studies focus on NN accelerator design and optimization on FPGAs, such as [22][25][30][14]. There are some studies focus on the method of designing NN accelerators, such as Minerva [23], which is a co-design approach to optimize NN accelerator across NN algorithms, architecture, and circuit.

TVM [4] is a deep learning compiler stack, it provides both graph-level and operator-level optimizations, and can target difference backends including CPU, GPU and hardware accelerators. DL Networks are first represented in computation graphs in TVM, on this level, TVM can do operator fusion, tensor layout transform and constant folding optimizations. After graph-level optimizations, operators are lowered to a form represented in TVM's tensor express language, then users can apply TVM's schedule primitives to create a schedule for operators. Schedules in TVM are mapping from tensor expression to actual low-level code to do the computation. For hardware accelerator backends, TVM provides tensorize schedule primitive, which can pattern match an unit of computation and replace it with accelerator's instruction.

VTA [19] is a programmable deep learning accelerator codesigned with TVM, and is intergrated into TVM. VTA has a two-level ISA, the first is task-level parallel memory transfer and compute ISA, the second is an microcoded-ISA which operates basic vector and matrix computations.

Hardware NN accelerator architecture is in a rapid development phase, so a parametric simulator that can be compatible with many NN accelerator architectures' characteristics is of great practical value to chip designers and compiler writers. It can be used to help with system-level exploration of chip architecture, or verify the effectiveness of optimizations of compilers, or evaluate the scheduling strategy of NN operators' implementations.

The contributions of this paper are the followings:

- a. Different NN accelerators have different ISA, memory hierarchy, and execution unit (types and sizes), so we designed and implemented a flexible and configurable NN accelerator simulator that is easy to

extend, and allows parameters of the simulated architecture to be modified by modifying configuration file.

- b. The simulator is a functional simulator which simulates the latencies of calculation and memory access, and the concurrent process between modules, and it gives the number of program execution cycles after the simulation is completed.
- c. We also integrated the simulator into the TVM compilation stack as a back-end, and the users can use TVM to write and generate operators and execute them on the simulator.

The rest of this paper is organized as follows. Section 2 introduces the architecture and ISA of the accelerator. Section 3 presents the software implementation of the simulator. Section 4 introduces the code generation system. Section 5 evaluates the simulator by different configurations.

2. Accelerator Architecture And Isa

In this section, we will present the architecture and instruction set architecture we designed and implemented in the simulator. The ISA is a two-level ISA: the first level is an RISC-like ISA, which contains instructions such as integer scalar calculation, branching, etc.. The second level is an ISA similar to the microcoded-ISA of VTA, which is responsible for tensor calculation and tensor data transfer, such as GEMM instructions. Multiple microcoded-ISA instructions form a microcoded kernel executed on the corresponding tensor pipeline. There is an extension instruction in the first-level ISA to launch a microcoded kernel on the corresponding tensor pipeline. In addition, the tensor pipelines employ the decoupled access-execute (DAE) architecture, so the first-level ISA contains two more instructions (dependency push/pop instructions) for explicit synchronization between tensor pipelines.

2.1 ACCELERATOR ARCHITECTURE

Figure 1. is the top-level block figure of this architecture. As illustrated in this figure, the architecture contains several stages to execute first-level ISA instructions: fetching, decoding, dispatching & register reading, execution & writing back. The integer scalars, which represent loop variables or addresses, are stored in register files and scalar memory. Similar to a typical RISC architecture, only the load/store instructions can access the scalar memory. A first-level ISA instruction will be fetched, decoded, and then injected into the in-order dispatch queue. When one instruction reaches the head of dispatch queue, and without dependency or anti-dependency to others, it will be dispatched to the corresponding execution unit. Branch instructions, integer scalar arithmetical and logical instructions will be sent to ALU, while scalar load/store instructions will be sent to load-store-unit (LSU). Other instructions for tensor pipeline synchronization or launching microcoded kernel will be sent to the corresponding tensor pipeline. There are FIFO queues between the dispatch queue and the execution unit or tensor pipeline, which are not shown on Figure 1. After execution, the results will be written to the register file. It should be noted that after one branch instruction has been dispatched, the dispatch process must be stalled until the branch result is known, because there is no re-ordering buffer in this architecture, the results will be committed directly and cannot be canceled.

In addition to the pipeline stages that execute the first-level ISA instructions, there are several tensor pipelines that execute the microcoded ISA instructions. Each tensor pipeline contains a pipeline controller, as well as an execution unit to perform actual works, and the execution unit may contain multiple pipeline stages inside. The pipeline controller is responsible for synchronizations between pipelines, as well as launching microcoded kernels (decoding and issuing microcoded ISA instructions in that kernel). Tasks running on different tensor pipelines are synchronized with the dependency pop/push instructions, to avoid data hazards. When a dependency push instruction is issued to a pipeline controller, the controller pushes a token to the corresponding dependency token queue, or waits for an empty slot to be available. In the case of the dependency pop instruction, the controller will either pop a token from the corresponding dependency token queue, or wait for a token to be available. A microcoded kernel consists of several microcoded-ISA instructions and is stored in the corresponding pipeline controllers internal storage. A launch kernel instruction contains two register operands, representing the extent of two nested loops. When a launch kernel instruction is issued to the pipeline controller, the controller uses the two operands as the extent to perform a two-level loop, and in the loop body, iteratively reads each microcoded ISA instructions of the kernel, decoding and then issue it into the execution unit. The procedure of launching a microcoded kernel and decoding and issuing each instruction is shown in algorithm 1.

Tensors, which represents data or weights, are stored in addressable on-chip scratchpad memories. In our simulator, we make the number of scratchpad memory configurable, and each tensor execution unit can access each memory, which makes the simulator more flexible. In our simulator, the capacity, bank width, bank number, read/write latencies, and type (one-port, or two-port) of each scratchpad memory are configurable. The simulator models the bank conflict between scratchpad memory accesses as well.

Current simulator implementation contains $3+N$ tensor pipelines, and can be easily extended.

The first pipeline is the matrix computing pipeline, which executes the microcode ISA instructions of GEMM. The execution unit (later referred as MAC) of matrix computing pipeline is divided into 4 pipeline stages: read, multiply, reduce and accumulate/write back. In the read stage, tensor operands are read from scratchpad memories. The multiply stage contains many multipliers that perform element-wise multiplication on broadcasted inputs. The reduce stage contains many adder trees that reduce sum the output of the multiply stage and get multiple dot products. In the last stage, the results are accumulated into partial sums stored in one scratchpad memory (later referred as accumulation buffer), or written back directly to one scratchpad memory. The simulator allows the input/output tensor size of the matrix computation instructions to be variable, with more flexible and configurable. These sizes are embedded into the microcoded ISA instructions.

Algorithm-1 The procedure of launching one microcoded kernel

Input: Two loop extents, ext0, ext1; the ID of the kernel to be launched, kld

```
1: kernel = get the kernel with ID kld
2: for var0 = 0; var0 < ext0; var0 ++ do
3:   for var1 = 0; var1 < ext1; var1 ++ do
4:   for all insn in instructions of kernel do
5:     while there is no free slot in output queue do
6:       end this cycle
7:     end while
8:     write opcode of insn to output
9:     for all arg in arguments of insn do
10:    if arg is immediate argument then
11:      argval = arg
12:    else if arg is composite argument then
13:      argval = arg.coef0 * var0 + arg.coef1 * var1 + arg.addend + localregs[arg.reg]
14:    end if
15:    write argval to output
16:  end for
17: end this cycle
18: end for
19: end for
20: end for
```

The second tensor pipeline is a vector computation pipeline, which performs vector activation functions, vector element-wise arithmetical and relational calculations, vector-immediate/scalar arithmetical and relational calculations, etc.. The execution unit of this pipeline is divided into three pipeline stages: read, calculate and write. As with matrix computing pipeline, we allow the input/output tensor sizes to be variable and embed the sizes into the instructions. Unlike matrix computing pipelines, vector computing pipelines cannot perform accumulate operation.

The third tensor pipeline is a memory transfer pipeline that is responsible for the data transfer between the scratchpad memory and the off-chip memory. Each memory transfer instruction can perform 2D memory transfer, making it easier to do data tiling. The remaining N tensor pipelines are also memory

transfer pipelines, which are responsible for data transfer between scratchpad memories. N is equal to the number of scratchpad memories, and the i -th one is responsible for copying data from other scratchpad memory to the i -th scratchpad memory. Although there are $N*N$ combinations of scratchpad memory pairs, we think N such pipelines are sufficient for modeling an actual architecture.

2.2 ACCELERATOR ISA

Table 1 is an overview of the instructions. As mentioned earlier, the ISA is a two-level ISA:

1. The first level is a RISC-like ISA, containing integer scalar calculation, branching and load/store instructions, as well as several extension instructions: instructions for synchronization between tensor pipelines, instructions for launching microcoded kernel, and instructions for assigning value to pipeline controller's local registers. The operands of these instructions are registers or integer immediates. The load/store instructions can be used to spill registers into a small scalar memory, or load parameters set by the host from the scalar memory. This level of ISA can be used to implement the control flow, which also makes it possible to influence the control flow with the value of the tensor data, which is necessary for operators such as ROI Pooling to be executed without copying data back to host.
2. The second level is microcoded ISA, which performs vector and matrix calculations as well as tensor memory transfers. Instructions in this ISA level cannot access the general register file. Their operands can be an integer immediate value, an immediate value of tensor data types, or a composite operand (consisting of two coefficients, an addend, and one pipeline controller local register, and is in the form of (C_1, C_2, A, R)). These operands are often used to represent the addresses of tensors, as well as strides or other parameters. Multiple microcoded ISA instructions form a microcoded kernel and will be stored in the pipeline controller. When launched by the pipeline controller, the instructions in the microcoded kernel are fetched and decoded by the controller, and sent to the corresponding tensor pipeline execution unit for execution.

Table 1 An overview of instructions

Type	Instructions		Operands
1 st level	control	branch, jump	immediate, register
	scalar	integer scalar arithmetical, relational, logical	
	load/store	integer scalar load and store	
2 nd level	vector	vector activation functions, vector binary arithmetical and relational, vector-immediate/scalar arithmetical and relational, etc.	integer immediate, tensor immediate, composite operand
	matrix	GEMM	
	data transfer	tensor DMA load/store, tensor copy, im2col	

Figure 3 presents the program fragment for fully-connected layer, and Figure 2 shows the source IR right before codegen. For brevity, we omit some instructions for register assignment and pipeline synchronization. Figure 3a shows instructions of the 1st-level ISA, which performs loops, pipeline synchronization and kernel launching operations, and Figure 3b shows the microcoded kernels composed of the 2nd-level ISA. As we can see, the 1st-level ISA instructions contain only one loop (namely, the loop on line 6 in original IR), and the rest of the loops are performed by the pipeline controllers when executing the LaunchKernel instructions, see algorithm 1. The 3rd and 4th parameters of each LaunchKernel instruction represent the two loop extents.

3. Simulator Implementations

The goal of our simulator is a functional simulator, that means, the simulator could simulate the concurrent processes between modules and pipelines, the latencies of memory accesses, instruction executions, and the bank conflicts of memory accesses. After completing the simulation of the program, it needs to give the number of cycles it takes to execute the program. In addition, the simulator must be sufficiently configurable. By modifying the configuration file or pre-defined configuration constants in the front-end code, users can modify the parameters of the simulated architecture, including:

1. The amount of scratchpad memories, the capacity, the number of banks, the bit-width, the read and write latency, and the type (one-port or two-port) of each scratchpad memory.
2. The execution latency for each instruction at each execution unit or pipeline stage.
3. Depth of various types of queues, such as queues from dispatch queues to execution units or pipeline controllers.
4. Data types of tensors. The simulator supports a variety of data types and mixed precision calculations. The data types of inputs and outputs of an instruction are encoded in the instruction.
5. The input and output tensor sizes of vector/matrix calculation instructions. These sizes are encoded in the instruction. This is equivalent to allowing the size of MAC and vector execution unit to be configurable.

Given the above requirements, we chose to develop simulator based on SystemC [20], which is a C++ class library providing a discrete event simulation interface. It contains a set of classes and macros, enabling users to simulate concurrent processes, which is important for hardware simulation. It also provides a notion of time into C++, enabling event sequencing. SystemC also provides data types for hardware modeling, such as 4 value logic vector, but users can still use C++ types. Figure 4 is the software architecture diagram of the simulator, which can be divided into three parts: modules, channels and common functions. Each one in the Modules part is a C++ class that inherits from `sc_module` (a base class in SystemC) and usually corresponds to a module/stage in the accelerator architecture. For example, the IFetch, IDecode, IDispatch and ALU in this figure correspond to fetch, decode, dispatch and ALU in Figure 1. A module can have several submodules, for example, Matrix Unit has four submodules, each representing one of the four pipeline stages. The read stage module further contains one Scratchpad Reader submodule for scratchpad memory read access. There are also inheritance hierarchies between module classes, for example, several different but similar pipeline controllers are required in the implementation, and inheritance mechanisms can help with code reuse. Each module has one or more processes (essentially member functions) which implement the function logic of the module. Modules interact with each other through channels, including the built-in channels of System C (such as `sc_signal`), and our custom channels, such as The Dependency Hub in the figure.

A module calls the methods of channels to write or read data, which transfer data between Modules or other functions. The Dependency Hub is a centralized dependency queue channel, through which each tensor pipeline sends dependency tokens to another pipeline and also gets tokens from others. The Scratchpad Memory Hub is a centralized channel that manages all scratchpad memories, and through it, each module can easily access any scratchpad memory. Using centralized channels makes it easier to extend new pipelines, as well as configure the memory hierarchy of simulated architecture, and etc. SRAM channels are a set of channels used to model scratchpad memory bank, all of which implement the same channel interface with different latencies and mutual-exclusive access modeling implementations. Classes in the Common part implement a variety of basic functions. The Memory class, for example, represents a memory. It implements methods of reading data, writing data and `memset`, but it has nothing to do with latency. `BitPacker` and `BitPackerImpl` are classes that represent the tensor data transferred between the modules. `BitPacker` is an abstract class that defines a set of virtual functions to perform type cast, element access, arithmetical and logical operations, and other operations. `BitPacker` manages the underlying data, as well as the size information. `BitPackerImpl` is a template class inherited from `BitPacker`, which represents tensors of a specific data type and implements the virtual functions defined by `BitPacker`. Further, we can extend data types to those not supported by C++, such as low bit and fixed-point numbers, by implementing the wrapper class of other data types, and using it as a template parameter for `BitPackerImpl`.

4. Codegen System

We integrate our backend into TVM by using TVM's `tensorize` primitive and some custom IR passes, and replacing the original code segment with `intrin` calls corresponding to microcoded instructions of our

architecture, e.g. GEMM. After this step, IRs become the form in Figure 2, and continue with the codegen operation. Our architecture uses a two-level ISA, so we first need to split the IR into 2 parts: the microcoded kernels consisting of the 2nd-level ISA instructions, and the IR consisting of the 1st-level ISA instructions. Microcoded kernels can be assembled as it is and loaded by the simulator, while the 1st-level ISA IR is further processed by code generator.

Since the first-level ISA is a register-based ISA, we need a code generator to do instruction selection, instruction scheduling, and register allocation, etc. We use the target-independent code generator of LLVM to do this work. The main job here is to write description files for our architecture's registers, register classes, instructions of the first-level ISA, as well as the selection patterns. We register LLVM intrinsic functions for instructions without corresponding LLVM instructions, including instructions launching the microcoded kernel and the dependency pop/push instructions. When LLVM IRs are generated from TVM IRs, these intrinsic calls will be converted to the appropriate LLVM intrinsic function calls.

We use an example to illustrate how to split the IR, for example, given the IR segment in Figure 2 for fully-connected layer, the code generator will convert IR codes from line 16 to 22 into a microcoded kernel. It first verifies whether the step lengths of the outermost two loops equal to 1, and extract the loop variables x_0 and y_0 . It also verifies whether inner loops have constant loop domains. Then it handles the GEMM call from line 19 to 21, and processes each argument according to the instruction template: for an immediate parameter, it verifies the type and value range of the arguments; for a composite parameter, it tries to represent the argument in the form of $\text{loop_vars}[\text{start}:\text{end}]$, where start and end are integers, contains only loop variables of inner loops, and contains only free variables that are defined outside the scope. then the parameter is converted to a composite parameter like: $\text{loop_vars}[\text{start}:\text{end}]$. After completing the previous steps, the code generator will unroll all inner loops, generate the microcoded kernels, and replace the original IR node with a launch kernel instruction along with instructions to set corresponding local registers. Figure 5 contains the IR codes after transformation, with only one loop in it. Figure 3a shows the assembly codes of the first-level IRs, which are generated from the transformed IRs in Figure 5. In this example, the 1st, 3rd and 5th parameters of the GEMM instruction are composite parameters, which represent the input/output addresses. The loop on line 18 has been unrolled, resulting in 4 GEMM microcoded instructions in the generated kernel. Figure 3b shows the extracted microcoded kernels, in which the one labeled as #4 is extracted from line 16 to 21 of the original IRs, which contain 4 GEMM instructions.

5. Experiments

5.1 FUNCTIONS OF THE SIMULATOR

In this subsection, we use the matrix multiplication example to demonstrate functions of the simulator, showing that it simulates the bank conflicts, memory access latency, and the concurrency between pipelines, etc. For example, we will compute $Y = XW$, where $X: (128, 1024)$, $W: (128, 1024)$. The configurations of test architecture are:

1. The shape of MAC is $8 \times 8 \times 8$. That means, each GEMM instruction reads two 8×8 tensors, and generate one 8×8 result. The throughput of the GEMM instruction is set to 1 per cycle (1/cycle).
2. Using 2 on-chip scratchpad memories to store input data and weight. The numbers of banks are both 8, and both have capacity of 32KB. The width of each bank is 8 bytes, the latencies are 1 cycle, and are all configured as two-port memory. The accumulation buffer has a capacity of 32KB. It is configured so that its latency does not become a bottleneck. It's two-port too.
3. The input data type of the GEMM instruction is int8, and the output data type is int16.

The simple method of completing the matrix multiplication is to decompose the whole calculation into multiple 8×8 matrix multiplications, loading two sub-matrixes of shape $8 \times (8 \times K)$ from both the X and W matrices at a time, where K should not be too small to avoid the overload of integer calculation, branch instructions, pipeline synchronization and launching microcoded kernel. Then we use the GEMM instruction to complete the partial calculation, repeat the operations until the calculation is completed, and save the accumulated final results back into the off-chip memory.

Based on the simple method, we apply several optimization methods to improve execution efficiency. The simulator will reflect the difference in the execution efficiency after applying optimizations. The optimization methods include:

1. Tensor tiling: Tiling matrices X and W so that the 8×8 sub-matrixes read by GEMM are stored continuously. This optimization helps to reduce or even avoid bank conflicts and improves the scratchpad memory read efficiency.
2. Data reuse: For many u-architectures, the cost of loading data from memory could be much higher than performing a single floating-point calculation. The same is true for our simulator, where the cost of loading data from the off-chip memory is much greater than computations. So we should reuse as much data already loaded into the on-chip scratchpad memory as we can, and do more calculations with the same amount of data loaded. We can do this by loading a few more rows of data at once, for example, we can load a submatrix of $32 \times (8 \times K)$ from both the X and W matrices at a time, then compute a total of $32 \times 32 \times 8 \times K$ MAC operations, with only $2 \times 32 \times 8 \times K$ data are read. It costs only 1/16 data read per MAC operation with the optimization.
3. Memory latency hiding: By making the processes of memory operations and calculations as overlapped as possible, we could increase the utilization of the memory and MAC. Since this architecture is an decoupled access-execute (DAE) architecture, we rely on TVM's cthread to do this optimization.

Because the simulator simulates the effects of bank conflicts, memory latencies and concurrency processes, these optimizations could obviously improve the computational efficiency. Based on the architecture configuration specified earlier, we set the value of K to 4, and the test results for applying these optimization methods are presented in Table 2.

Table 2 Results of optimizations

optimization	#cycles	MAC utilization (%)
original	564,749	5.8
tiling	466,445	7
data reuse	288,501	11.4
tiling & data reuse	57,237	57.2
all	36,926	88.7

5.2 CONFIGURABILITY OF THE SIMULATOR

In this subsection, we use an example to demonstrate the configurability of the simulator, and how those configurations affect the performance of the simulation. In this example, we use 3 operators to demonstrate the configurability of the MAC shape, scratchpad memory bandwidth and other parameters, and how these configurations affect the execution performance. The operators include:

1. A fully-connected operator: , where X: (128, 1024), W: (1024, 1024))
2. A conv2d operator: where input: (1, 16, 16, 256) in NHWC layout, kernel: (3, 3, 256, 256), strides: (1, 1)
3. A max pooling operator: where input: (40, 40, 256), ksize: (2, 2) strides:(2, 2)

The top-level block figure of the simulated architecture is shown in Figure 6, in which the scalar pipeline and other control units are ignored. 2 on-chip scratchpad memories (buf0 and buf1) are used to store data and the weight, which provides inputs to the MAC unit, which writes its output to the accumulation buffer. Another 2 smaller scratchpad memories (buf3 and buf4) are used to store the final results and the bias data, and are connected to the vector unit. The vector unit can also read data from the accumulation buffer for further calculations and add bias.

The five different configurations are as follows:

1. The shape of MAC is 8 x 8 x 8, the size of the vector unit is 64. The bandwidths and latencies of buf0 and buf1 match exactly what GEMM instructions needs, and both have the capacity of 32KB, the capacity of the accumulation buffer is also 32 KB, with all tree memories twoported. In the configuration, the bandwidth between the off-chip memory and the scratchpad memories is about 1/10 of the bandwidth required by the MAC unit.
2. The MAC shape is set to 16 x 16 x 16, the size of vector unit is 256, and with other configurations unchanged.
3. Increasing the bandwidths of the scratchpad memories, making them to match the requirement of the new MAC and the vector unit shape.
4. Quadrupling the capacity of the accumulation buffer, with better data reuse, since larger capacity of the accumulation buffer allows us to load more rows at once. We quadruple the capacity of the buf4 for loading more data at once for the vector unit as well.

5. Doubling the bandwidth from the off-chip memory to the on-chip scratchpad memories.

For each configuration, the total cycles and the utilization of the MAC/vector unit of the 3 operators are shown in Table 3. As shown in the table, for the fully-connected and the conv2d operators, simply increasing the shape of MAC can hardly improve performance, but leads to a low MAC utilization. Increasing the scratchpad memory bandwidth has also not resulted in significant performance improvements. It was not until the accumulation buffer capacity and the off-chip to on-chip bandwidth are increased that significant performance improvements were made. Even so, the MAC utilization is not objective, possibly because the computational load of the matrix multiplication in this example is too small relative to the new MAC shape for sufficient data reuse, or the memory bandwidth is still not large enough. As for the max pooling operator, because it's a memory bound operator (no data reuse possible), the vector unit utilization is always low, and the performance almost entirely depends on the memory bandwidth.

Table 3 Simulation results of different configurations

configuration	fully-connected	conv2d	max pooling
	#cycles	#cycles	#cycles
	MAC util (%)	MAC util (%)	MAC util (%)
original	285,379	315,290	42,779
	91.9	93.5	15
larger MAC	266,491	292,750	42,779
	12.3	12.6	3.7
wider scratchpad bandwidth	221,808	247,330	38,325
	14.8	14.9	4.2
larger accumulation buffer	115,187	127,809	36,171
	28.5	28.8	4.4
double bandwidth from off-chip to on-chip	61,923	68,417	20,771
	53.0	53.9	7.7

5.3 MIMICING OTHER U-ARCHITECTURES

Finally, we use an example to demonstrate the ability of the simulator to mimic a particular accelerator u-architecture, like the Da Vinci of Huawei [27]. The block figure of the mimicked architecture is shown in Figure 7. The configurations are:

1. There is a total of 5 on-chip scratchpad memories. The first (L1 buf) has a capacity of 1MB and is the only scratchpad memory that is connected to the off-chip memory. The second and third scratchpad memories (L0A and L0B) have a capacity of 64KB and are connected to the L1 buf, which provides inputs to the MAC unit. The accumulation buffer (L0C buf) has a capacity of 256KB, which is connected to the MAC unit and the vector unit. The capacity of the fourth scratchpad memory (uni buf) is 256KB and is connected to the L1 buf and vector unit. In the configuration, all memories except L1 buf have bandwidths that meet the requirements of the corresponding execution unit. The bandwidth of L1 buf to L0A or L0B is approximately 1/10 of the bandwidth required by the MAC unit (for both data and weight). And the bandwidth between the off-chip memory and the L1 buf is about 1/30 of the bandwidth required by the MAC unit.
2. The shape of MAC is 8 x 8 x 8, and the size of vector unit is 64.
3. The data type is configured as FP16+FP32, or int8.

With the above configurations, the memory hierarchy and MAC shape of the simulated uarchitecture are similar to the Da Vinci u-architecture. The tensor data type can also be configured to be the same as well. We still use the operators in subsection 5.2 as examples to do simulations. The results are shown in Table 4.

Table 4 Simulation results of the mimiced u-architecture

operator	#cycles	MAC util (%)	vector unit util (%)
fully-connected	377,611	69.4	NA
conv2d	495,322	59.5	NA
max pooling	106,980	NA	6

Declarations

Acknowledgements

This work was supported by Huawei Research Fund No.HF2018055044.

Funding

No Funding

Conflicts of interest/Competing interests

Not Applicable

Availability of data and material

Not Applicable

Code availability

Not Applicable

Authors' contributions

Not Applicable

References

- [1] Jorge Albericio, Patrick Judd, Tayler H Hetherington, Tor M Aamodt, Natalie D Enright Jerger, and Andreas Moshovos. 2016. Cnvlutin: ineffectual-neuron-free deep neural network computing. *international symposium on computer architecture* 44, 3 (2016), 1–13.
- [2] Renzo Andri, Lukas Cavigelli, Davide Rossi, and Luca Benini. 2017. YodaNN: An Architecture for Ultra-Low Power Binary-Weight CNN Acceleration. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* PP, 99 (2017), 1–1.
- [3] Tianshi Chen, Zidong Du, Ninghui Sun, Jia Wang, Chengyong Wu, Yunji Chen, and Olivier Temam. 2014. DianNao: a small-footprint high-throughput accelerator for ubiquitous machine-learning. *architectural support for programming languages and operating systems* 49, 4 (2014), 269–284.
- [4] Tianqi Chen, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Yan, Meghan Cowan, Haichen Shen, Leyuan Wang, Yuwei Hu, Luis Ceze, et al. 2018. TVM: an automated end-to-end optimizing compiler for deep learning. *operating systems design and implementation* (2018), 579–594.
- [5] Yuhsin Chen, Tushar Krishna, Joel S Emer, and Vivienne Sze. 2017. Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks. *IEEE Journal of Solid-state Circuits* 52, 1 (2017), 127–138.
- [6] Yunji Chen, Tao Luo, Shaoli Liu, Shijin Zhang, Liqiang He, Jia Wang, Ling Fei Li, Tianshi Chen, Zhiwei Xu, Ninghui Sun, et al. 2014. DaDianNao: A Machine-Learning Supercomputer. *international symposium on microarchitecture* (2014), 609–622.
- [7] Ping Chi, Shuangchen Li, Cong Xu, Tao Zhang, Jishen Zhao, Yongpan Liu, Yu Wang, and Yuan Xie. 2016. PRIME: a novel processing-in-memory architecture for neural network computation in ReRAM-based main memory. *international symposium on computer architecture* 44, 3 (2016), 27–39.
- [8] Zidong Du, Robert Fasthuber, Tianshi Chen, Paolo Ienne, Ling Fei Li, Tao Luo, Xiaobing Feng, Yunji Chen, and Olivier Temam. 2015. ShiDianNao: shifting vision processing closer to the sensor. *international symposium on computer architecture* 43, 3 (2015), 92–104.
- [9] Mingyu Gao, Jing Pu, Xuan Yang, Mark Horowitz, and Christoforos E Kozyrakis. 2017. TETRIS: Scalable and Efficient Neural Network Acceleration with 3D Memory. *architectural support for*

programming languages and operating systems 45, 1 (2017), 751–764.

- [10] Song Han, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram, Mark Horowitz, and William J Dally. 2016. EIE: efficient inference engine on compressed deep neural network. international symposium on computer architecture 44, 3 (2016), 243–254..
- [11] Norman P Jouppi, C S Young, Nishant Patil, David A Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh K Bhatia, Nan Boden, Al Borchers, et al. 2017. In-Datcenter Performance Analysis of a Tensor Processing Unit. international symposium on computer architecture 45, 2 (2017), 1–12.
- [12] Patrick Judd, Jorge Albericio, Tayler H Hetherington, Tor M Aamodt, and Andreas Moshovos. 2016. Stripes: bit-serial deep neural network computing. international symposium on microarchitecture (2016), 1–12.
- [13] Duckhwan Kim, Jaeha Kung, Sek Chai, Sudhakar Yalamanchili, and Saibal Mukhopadhyay. 2016. Neurocube: A Programmable Digital Neuromorphic Architecture with High-Density 3D Memory. In International Symposium on Computer Architecture.
- [14] Hongbin Li, Xitian Fan, Li Jiao, Wei Ping Cao, Xuegong Zhou, and Lingli Wang. 2016. A high performance FPGA-based accelerator for large-scale convolutional neural networks. (2016), 1–9.
- [15] Daofu Liu, Tianshi Chen, Shaoli Liu, Jinhong Zhou, Shengyuan Zhou, Olivier Teman, Xiaobing Feng, Xuehai Zhou, and Yunji Chen. 2015. PuDianNao: A Polyvalent Machine Learning Accelerator. architectural support for programming languages and operating systems 43, 1 (2015), 369–381.
- [16] Shaoli Liu, Zidong Du, Jinhua Tao, Dong Han, Tao Luo, Yuan Xie, Yunji Chen, and Tianshi Chen. 2016. Cambricon: an instruction set architecture for neural networks. international symposium on computer architecture 44, 3 (2016), 393–405.
- [17] Bert Moons, Roel Uytterhoeven, Wim Dehaene, and Marian Verhelst. 2017. 14.5 Envision: A 0.26-to-10TOPS/W subword-parallel dynamic-voltage-accuracy-frequency-scalable Convolutional Neural Network processor in 28nm FDSOI. In Solid-state Circuits Conference.
- [18] Bert Moons and Marian Verhelst. 2016. A 0.3-2.6 TOPS/W Precision-Scalable Processor for Real-Time Large-Scale ConvNets. In Vlsi Circuits.
- [19] Thierry Moreau, Tianqi Chen, Ziheng Jiang, Luis Ceze, Carlos Guestrin, and Arvind Krishnamurthy. 2018. VTA: An Open Hardware-Software Stack for Deep Learning. arXiv: Learning (2018).
- [20] P. R. Panda. 2001. SystemC - a modeling platform supporting multiple design abstractions. In International Symposium on System Synthesis.
- [21] Angshuman Parashar, Minsoo Rhu, Anurag Mukkara, Antonio Puglielli, Rangharajan Venkatesan, Brucek Khailany, Joel S Emer, Stephen W Keckler, and William J Dally. 2017. SCNN: An Accelerator for

Compressed-sparse Convolutional Neural Networks. international symposium on computer architecture 45, 2 (2017), 27–40.

[22] Jiantao Qiu, Jie Wang, Song Yao, Kaiyuan Guo, Boxun Li, Erjin Zhou, Jincheng Yu, Tianqi Tang, Ningyi Xu, Sen Song, et al. 2016. Going Deeper with Embedded FPGA Platform for Convolutional Neural Network. (2016), 26–35.

[23] Brandon Reagen, Paul N Whatmough, Robert Adolf, Saketh Rama, Hyunkwang Lee, Sae Kyu Lee, Jose Miguel Hernandezlobato, Guyeon Wei, and David M Brooks. 2016. Minerva: enabling low-power, highly-accurate deep neural network accelerators. international symposium on computer architecture 44, 3 (2016), 267–278.

[24] Ali Shafiee, Anirban Nag, Naveen Muralimanohar, Rajeev Balasubramonian, John Paul Strachan, Miao Hu, R Stanley Williams, and Vivek Srikumar. 2016. ISAAC: a convolutional neural network accelerator with in-situ analog arithmetic in crossbars. international symposium on computer architecture 44, 3 (2016), 14–26.

[25] Hardik Sharma, Jongse Park, Divya Mahajan, Emmanuel Amaro, Joon Kyung Kim, Chenkai Shao, Asit K Mishra, and Hadi Esmaeilzadeh. 2016. From high-level deep neural models to FPGAs. international symposium on microarchitecture (2016), 1–12.

[26] Linghao Song, Xuehai Qian, Hai Li, and Yiran Chen. 2017. PipeLayer: A Pipelined ReRAM-Based Accelerator for Deep Learning. (2017), 541–552.

[27] Liang Xiaoyao. 2019. Ascend AI Processor Architecture and Programming. TSINGHUA UNIVERSITY PRESS

[28] S. Yin, P. Ouyang, J. Yang, T. Lu, X. Li, L. Liu, and S. Wei. 2018. An Ultra-High Energy-Efficient Reconfigurable Processor for Deep Neural Networks with Binary/Ternary Weights in 28NM CMOS. In 2018 IEEE Symposium on VLSI Circuits. 37–38. <https://doi.org/10.1109/VLSIC.2018.8502388>

[29] Shijin Zhang, Zidong Du, Lei Zhang, Huiying Lan, Shaoli Liu, Ling Li, Qi Guo, Tianshi Chen, and Yunji Chen. 2016. Cambricon-x: an accelerator for sparse neural networks. international symposium on microarchitecture (2016), 1–12.

[30] Ritchie Zhao, Weinan Song, Wentao Zhang, Tianwei Xing, Jenghau Lin, Mani B Srivastava, Rajesh Gupta, and Zhiru Zhang. 2017. Accelerating Binarized Convolutional Neural Networks with Software-Programmable FPGAs. (2017), 15–24.

Figures

Figure 1

The top-level block figure of the u-architecture

```

NPU.DMABufLoad(bias, 0, 0, (0 + 0x50000000), 0, 128, 1)
for (xo, 0, 8) {
for (yo, 0, 8) {
NPU.Memset((((xo*8) + yo)*128) + 0x30000000), 8, 8, 16, 0.0, 1)
}}
for (ko, 0, 32) {
for (io, 0, 8) {
for (il, 0, 4) {
NPU.DMABufLoad(a, (((io*1024) + il) + (ko*4))*8), 1024,
((((io*4) + il)*64) + 0x10000000), 8, 8, 8)
}}
for (io, 0, 8) {
for (il, 0, 4) {
NPU.DMABufLoad(b, ..... )
}}
for (xo, 0, 8) {
for (yo, 0, 8) {
for (ki, 0, 4) {
NPU.GEMM((((xo*8) + yo)*128) + 0x30000000), 16,
((((xo*4) + ki)*64) + 0x10000000), 8,
((((yo*4) + ki)*64) + 0x20000000), 8, ..... )
}}
}
for (no, 0, 8) {
for (oco, 0, 8) {
NPU.MAddV((((no*8) + oco)*64) + 0x40000000), 8,
((((no*8) + oco)*128) + 0x30000000), 16,
((oco*16) + 0x50000000), 8, 8, 2)
}}
NPU.DMABufStore(out_host, 0, 0, (0 + 0x40000000), 0, 4096, 1)

```

Figure 2

IR for fully-connected layer

# \$g5: bias address #	#0:
\$g4: output address #	DMABufLoad {0:0:0:1}, 0,
\$g6: weight address #	{0:0:0x50000000:0}, 0, 128, 1
\$g7: data address	#1:
SetPipelineReg 1, 1,	Memset {1024:128:0x30000000:0},
\$g5 # \$g5=1, \$g8=8, \$g10=4	8, 8, 16, 0.0, 1
LaunchKernel 1, #0, \$g5, \$g5	#2:
LaunchKernel 7, #1, \$g8, \$g8	DMABufLoad {8192:8:0:1}, 1024,
# pipeline sync omitted	{256:64:0x10000000:0}, 8, 8, 8
DependPush 2, 1	#3:
% # \$g9 is loop var	DMABufLoad
AddIU \$g9, \$zero, 0	#4:
.LBB0_1:	GEMM {1024:128:0x30000000:0},
DependPop 2, 1	16, {256:0:0x10000000:0}, 8,
SHLI \$g11, \$g9, 5	{0:256:0x20000000:0}, 8,
AddU \$g12, \$g7, \$g11	GEMM {1024:128:0x30000000:0},
SetPipelineReg 1, 1, \$g12	16, {256:0:0x10000040:0}, 8,
LaunchKernel 1, #2, \$g8, \$g10	{0:256:0x20000040:0}, 8,
AddU \$g12, \$g6, \$g11	GEMM
SetPipelineReg 1, 1, \$g12	GEMM.....
LaunchKernel 1, #3, \$g8, \$g10	#5:
DependPush 1, 2	MAddV {512:64:0x40000000:0}, 8,
DependPop 1, 2	{1024:128:0x30000000:0}, 16,
LaunchKernel 2, #4, \$g8, \$g8	{0:16:0x50000000:0}, 8, 8, 2
DependPush 2, 1	#6:
AddIU \$g9, \$g9, 1	DMABufStore {0:0:0:1}, 0,
SLTIU \$g11, \$g9, 32	{0:0:0x40000000:0}, 0, 4096, 1)
BNEZ \$g11, .LBB0_1	
# pipeline sync omitted	
LaunchKernel 3, #5, \$g8, \$g8	
# pipeline sync omitted	
SetPipelineReg 1, 1, \$g4	
LaunchKernel 1, #6, \$g5, \$g5)	

(a) Assembly of 1st-level code. (b) Microcoded kernels.

Figure 3

Code fragment for fully-connected layer

Figure 4

The software architecture of the simulator

```

llvm.NPU.SetPipelineReg(1, 1, bias)
llvm.NPU.LaunchKernel(1, 0, 1, 1)
llvm.NPU.LaunchKernel(7, 1, 8, 8)
for (ko.outer, 0, 32) {
    llvm.NPU.SetPipelineReg(1, 1, a + (ko.outer*32))
6 llvm.NPU.LaunchKernel(1, 2, 8, 4)
    llvm.NPU.SetPipelineReg(1, 1, b + (ko.outer*32))
8 llvm.NPU.LaunchKernel(1, 3, 8, 4)
    llvm.NPU.LaunchKernel(2, 4, 8, 8)
}
llvm.NPU.LaunchKernel(3, 5, 8, 8)
llvm.NPU.SetPipelineReg(1, 1, out_host)
llvm.NPU.LaunchKernel(1, 6, 1, 1)

```

Figure 5

The IR after extracting microcoded kernel

Figure 6

Top-level block figure of simulated architecture

Figure 7

Top-level block figure of a mimicked architecture