AS501: Notice about Final Project

KAIST Graduate School of AI Semiconductor

18 May 2024

Outline

KAIST EE

1. Introduction

2. Baseline HW + SW

3. Guidelines

Outline

KAIST EE

1. Introduction
2. Baseline HW + SW
3. Guidelines

Final Project

KAIST EE

- The final project's outline is to achieve better performance on your hardware than on baseline hardware when running baseline (or improved) software

Evaluation Standards

KAIST EE

1. Proposal (10 %)

2. Presentation (60 %)
   - Peer review (30 %)
     - Each group's final report will be evaluated by other groups
   - Professor review (30 %)

   - Standard: Presentation quality (50 %), Innovation of Idea (25 %), Performance improvement (25 %)

3. Final report (30 %)

Outline

1. Introduction

2. Baseline HW + SW

3. Guidelines

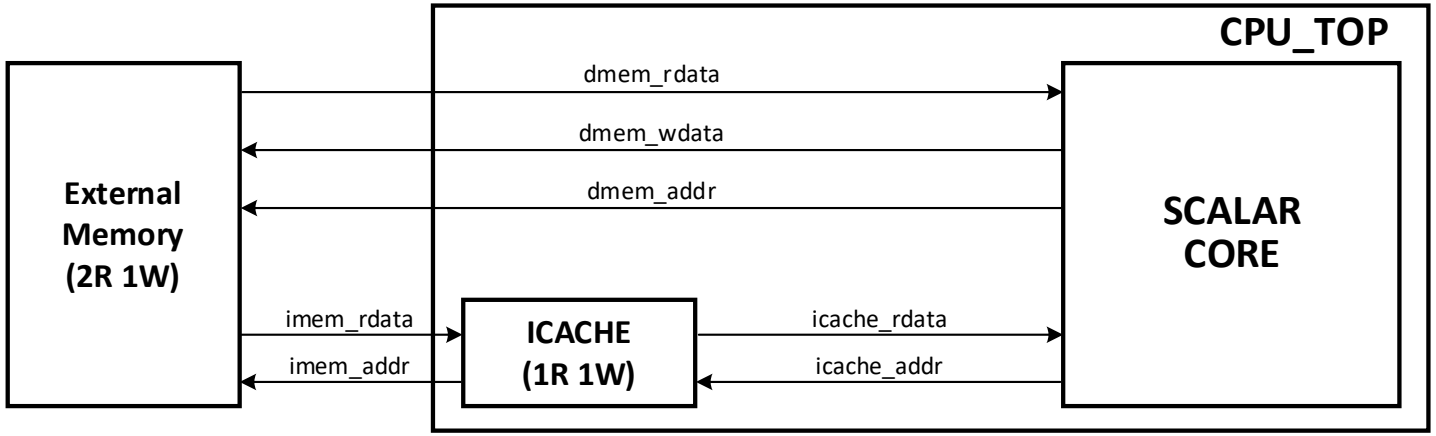
Baseline HW + SW File

- 1. Download **final.tar.gz** to /home/ASxxxxxxxx
- 2. tar –xzvf final.tar.gz

Baseline HW Specifications

KAIST EE

- Scalar core + Instruction cache + External memory



- Scalar core: Non-pipelined
- Instruction cache: Direct-mapped, SAED32nm Marco SRAM (512 B)
  - Hit: 1 cycle
  - Miss: 1 cycle + 40 ns
  - You can modify the clock period according to your hardware
- External Memory: Behavioral memory (32 MB)
  - **40 ns {4 x memory clock (10ns), fixed}** 메모리 읽기/쓰기 딜레이는 고정입니다
  - All sample parameters are already loaded
  - **You can't modify the external memory clock, # of ports, and word width**

    메모리 클럭, 포트 수, 메모리 가로 폭은 수정할 수 없습니다.

```
// Clock generator
// You can modify ClkPeriod according to your hardware
localparam ClkPeriod      = 10.0;
localparam ClkHalf        = ClkPeriod / 2;
initial begin
    clk = 1'b0;
    forever #ClkHalf clk = ~clk;
end
```

```
// Memory Clock generator
// You can't modify mClkPeriod
localparam mClkPeriod     = 10.0;
localparam mClkHalf       = mClkPeriod / 2;
initial begin
    mclk = 1'b0;
    forever #mClkHalf mclk = ~mclk;
end
```

Baseline HW - Synthesis Report

KAIST EE

- Clock Period: 10.0 ns

Timing

clock clk_i (rise edge)	10.00	10.00
clock network delay (ideal)	0.00	10.00
clock uncertainty	-0.70	9.30
SCORE/DECODER/COUNTER_INSTRET/cnt_data_o_reg_63_/CLK (DFFARX1_RVT)	0.00	9.30
library setup time	-0.03	9.27
data required time		9.27
data required time		9.27
data arrival time		-9.25
slack (MET)		0.01

Area

Combinational area:	18894.843989
Buf/Inv area:	1067.404808
Noncombinational area:	9405.869731
Macro/Black Box area:	28363.267578
Net Interconnect area:	11931.264669
Total cell area:	56663.981299
Total area:	68595.245968

Power

```
  Cell Internal Power  =   1.4433 mW  (100%)
  Net Switching Power  =   6.1077 uW    (0%)
                         ---------
Total Dynamic Power    =   1.4494 mW  (100%)

Cell Leakage Power     =   2.7357 mW
```

Power Group	Internal Power	Switching Power	Leakage Power	Total Power	(%)	Attrs
io_pad	0.0000	0.0000	0.0000	0.0000	(0.00%)	
memory	92.1192	0.2450	0.0000	92.3642	(2.21%)	
black_box	0.0000	0.0000	0.0000	0.0000	(0.00%)	
clock_network	1.3462e+03	0.0000	0.0000	1.3462e+03	(32.17%)	i
register	1.0023	0.3151	1.4365e+09	1.4378e+03	(34.36%)	
sequential	0.0000	0.0000	0.0000	0.0000	(0.00%)	
combinational	3.9537	5.5477	1.2992e+09	1.3087e+03	(31.27%)	
Total	1.4432e+03 uW	6.1079 uW	2.7357e+09 pW	4.1850e+03 uW		

- **Use these values as a baseline HW performance**

Baseline HW - P&R Report

KAIST EE

- Clock Period: 10.0 ns (by synthesis)

Timing

```
clock clk_i (rise edge)                                 10.00     10.00
clock network delay (propagated)                         0.61     10.61
clock reconvergence pessimism                            0.04     10.65
SCORE/DECODER/COUNTER_INSTRET/cnt_data_o_reg_63_/CLK (DFFARX1_RVT)
                                                         0.00     10.65 r
clock uncertainty                                       -0.10     10.55
library setup time                                      -0.06     10.49
data required time                                                10.49
-----------------------------------------------------------------------
data required time                                                10.49
data arrival time                                                -10.48
-----------------------------------------------------------------------
slack (MET)                                                       0.02
```

Power

```
   Cell Internal Power    = 9.25e+08 pW ( 83.3%)
  Net Switching Power     = 1.86e+08 pW ( 16.7%)
Total Dynamic Power       = 1.11e+09 pW (100.0%)

Cell Leakage Power        = 1.79e+09 pW

Power Group         Internal Power       Switching Power         Leakage Power          Total Power    (   %  )  Attrs
-------------------------------------------------------------------------------------------------------------------
io_pad                 0.00e+00             0.00e+00             0.00e+00             0.00e+00    (  0.0%)
memory                 7.30e+07             1.29e+06             0.00e+00             7.43e+07    (  2.6%)
black_box              0.00e+00             0.00e+00             6.24e+06             6.24e+06    (  0.2%)
clock_network          8.46e+08             1.74e+08             2.07e+07             1.04e+09    ( 35.8%)       i
register               1.18e+06             5.16e+05             6.79e+08             6.81e+08    ( 23.4%)
sequential             0.00e+00             0.00e+00             0.00e+00             0.00e+00    (  0.0%)
combinational          5.23e+06             1.03e+07             1.09e+09             1.10e+09    ( 38.0%)
-------------------------------------------------------------------------------------------------------------------
Total                  9.25e+08 pW          1.86e+08 pW          1.79e+09 pW          2.91e+09 pW
```

Area

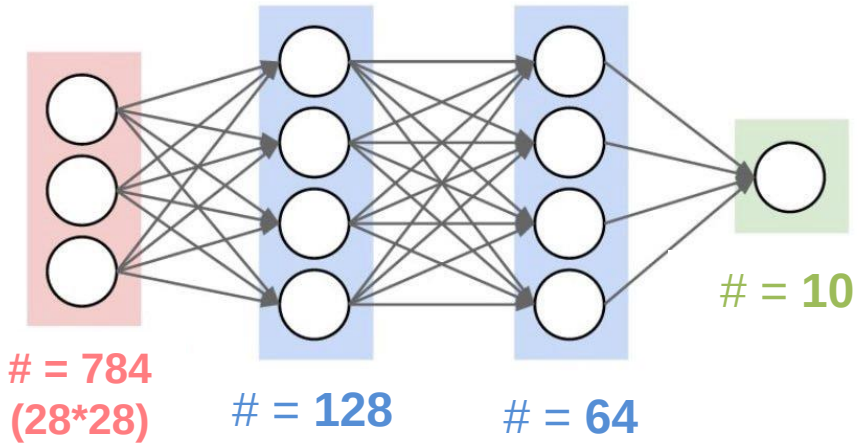
```
Cell Area (netlist):                     59273.03
Cell Area (netlist and physical only):   59727.94
```

- PnR result is slightly different between runs due to the varying division of tasks between threads
  PnR은 매번 돌릴 때마다, 결과가 미세하게 달라집니다. 값이 미세하게 다른 점에 대해 크게 신경 쓰지마세요
- **Use these values as a baseline HW performance**

Baseline SW Specifications

KAIST EE

- Multilayer perceptron



- program/test_code/mlp.c

```
for (image_idx = 0; image_idx < NUM_OF_TEST; ++image_idx){
    fully_connected(&input[INPUT_SIZE * image_idx], fc1_weight,
                    fc1_bias, fc1_output, INPUT_SIZE, HIDDEN1_SIZE);
    fully_connected(fc1_output, fc2_weight, fc2_bias, fc2_output,
                    HIDDEN1_SIZE, HIDDEN2_SIZE);
    fully_connected(fc2_output, fc3_weight, fc3_bias, output,
                    HIDDEN2_SIZE, OUTPUT_SIZE);

    if (max(output) == label[image_idx]){
        ++correct_count;
    }
}
```

- **All sample test codes have already been compiled in program/out according to # of test inputs**
- **You can retrain the model using Machine Learning techniques**
- **If you intentionally train with provided test set, you will be given a score of 0**

  의도적으로 test data를 이용해 학습을 수행하여, accuracy를 높이는 행위를 할 경우 0점 처리하겠습니다.

Baseline SW (continue)

KAIST EE

- Multilayer perceptron

```
----------------------------------------------------
Accuracy = 100.00
----------------------------------------------------
mcycleh = 00000000, mcycle = 08f09001
minstreth = 00000000, minstret = 0216888c
----------------------------------------------------
$finish at simulation time       14998393250000
           V C S   S i m u l a t i o n   R e p o r t
Time: 1499839325000000 fs
CPU Time:    536.030 seconds;       Data structure size:  64.9Mb
```

Accuracy, cycle, instruction count, and simulation time when # of test inputs = 10

```
----------------------------------------------------
Accuracy = 99.00
----------------------------------------------------
mcycleh = 00000000, mcycle = 596590d6
minstreth = 00000000, minstret = 14e155ae
----------------------------------------------------
$finish at simulation time      149982913750000
           V C S   S i m u l a t i o n   R e p o r t
Time: 14998291375000000 fs
CPU Time:   5297.910 seconds;       Data structure size:  64.9Mb
```

Accuracy, cycle, instruction count, and simulation time when # of test inputs = 100

```
----------------------------------------------------
Accuracy = 96.50
----------------------------------------------------
mcycleh = 00000003, mcycle = 7df7940f
minstreth = 00000000, minstret = d0cd54c6
----------------------------------------------------
$finish at simulation time     1499827987650000
           V C S   S i m u l a t i o n   R e p o r t
Time: 149982798765000000 fs
CPU Time:  53180.780 seconds;       Data structure size:  64.9Mb
```

Accuracy, cycle, instruction count, and simulation time when # of test inputs = 1000

Takes too long on baseline HW
(minimum 10 days)
Expected accuracy: 96.76%

when # of test inputs = 10000

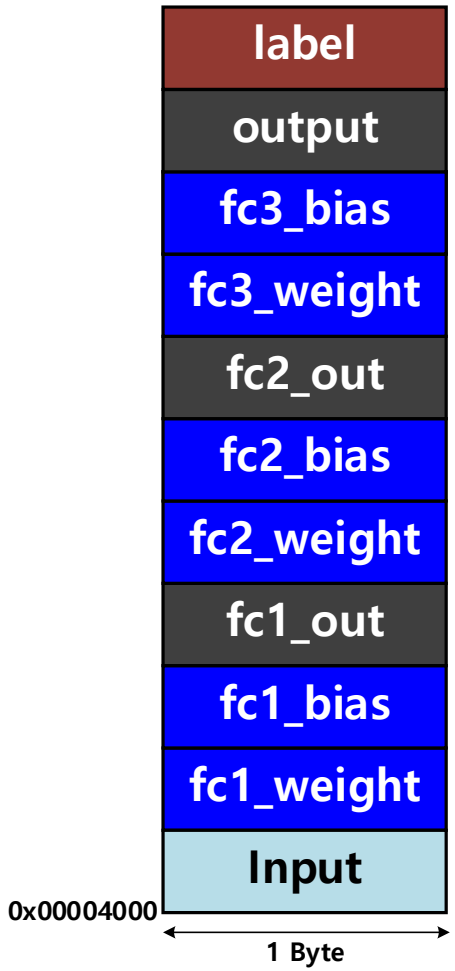
- **Use these values as a baseline SW performance**

Parameter Memory Map

```
localparam InputSize     = 'd784;
localparam Hidden1Size   = 'd128;
localparam Hidden2Size   = 'd64;
localparam OutputSize    = 'd10;
localparam NumOfTest     = 'd10;

localparam InputAddr     = DMemStart;
localparam Fc1WAddr      = (InputAddr  + InputSize   * NumOfTest   * 4);
localparam Fc1BAddr      = (Fc1WAddr   + InputSize   * Hidden1Size * 4);
localparam Fc1OAddr      = (Fc1BAddr   + Hidden1Size * 4);
localparam Fc2WAddr      = (Fc1OAddr   + Hidden1Size * 4);
localparam Fc2BAddr      = (Fc2WAddr   + Hidden1Size * Hidden2Size * 4);
localparam Fc2OAddr      = (Fc2BAddr   + Hidden2Size * 4);
localparam Fc3WAddr      = (Fc2OAddr   + Hidden2Size * 4);
localparam Fc3BAddr      = (Fc3WAddr   + Hidden2Size * OutputSize  * 4);
localparam OutputAddr    = (Fc3BAddr   + OutputSize  * 4);
localparam LabelAddr     = (OutputAddr + OutputSize  * 4);

localparam ImageInitFile = "../../program/test_code/image/image_10.txt";
localparam Fc1WInitFile  = "../../program/test_code/parameter/fc1_weight.txt";
localparam Fc1BInitFile  = "../../program/test_code/parameter/fc1_bias.txt";
localparam Fc2WInitFile  = "../../program/test_code/parameter/fc2_weight.txt";
localparam Fc2BInitFile  = "../../program/test_code/parameter/fc2_bias.txt";
localparam Fc3WInitFile  = "../../program/test_code/parameter/fc3_weight.txt";
localparam Fc3BInitFile  = "../../program/test_code/parameter/fc3_bias.txt";
localparam LabelInitFile = "../../program/test_code/label/label.txt";
```



- **All sample testbenches have already been written in sim/tb according to # of test inputs**

Parameter Memory Map (continue)

KAIST EE

- **Input and label must use the provided file (program/test_code/image)**
  - Pre-processing (zero-skipping, compression, etc) must be performed in your HW

  입력과 label은 제공된 파일을 사용하세요. 입력에 전처리 과정이 필요하다면, HW에서 수행해야 합니다.

- **You don't need to store intermediate output (fc1_out, and fc2_out)**

  최종 output이 아닌 중간 layer output은 외부 메모리에 꼭 쓸 필요는 없습니다. (HW에 따라 사이클 낭비가 될 수 있습니다.)

- **Please submit a file including all parameters and write the address in the report.**

  모든 parameter 파일을 제출하고 메모리 저장 주소를 report에 기재해주세요.

- **Performance result must include the action of storing the final output to external memory**
  - **Post-processing is also required**

  최종 output을 메모리에 쓰는 행동까지 성능에 포함되어야 합니다.
  만약 output의 데이터 형식이, baseline과 다르면 후처리를 HW에서 수행한 후 저장해야 합니다.

- We plan to use random test inputs to verify the authenticity of your report contents

  우리는 랜덤 test input을 사용해서, 제출한 parameter를 메모리에 적재하여, 보고서에 기재된 내용의 신뢰성을 검증할 예정입니다.

File Description

- program/testcode/mlp.h : Header file of multilayer perceptron
- program/testcode/mlp.c : Source code of multilayer perceptron

- program/testcode/parameter: weights and bias of each layer
- program/testcode/label: label

- program/testcode/image_10: 10 test image
- program/testcode/image_100: 100 test image
- program/testcode/image_1000: 1000 test image
- program/testcode/image_10000: 10000 test image

File Description (continue)

KAIST EE

- program/out/mlp_10.hex: Hex file when # of test inputs is 10
- program/out/mlp_100.hex: Hex file when # of test inputs is 100
- program/out/mlp_1000.hex: Hex file when # of test inputs is 1000
- program/out/mlp_10000.hex: Hex file when # of test inputs is 10000

- sim/run_vcs_10.sh: simulation script when # of test images is 10
- sim/run_vcs_100.sh: simulation script when # of test images is 100
- sim/run_vcs_1000.sh: simulation script when # of test images is 1000
- sim/run_vcs_10000.sh: simulation script when # of test images is 10000

- sim/tb/core_tb_10.sv: testbench when # of test images is 10
- sim/tb/core_tb_100.sv: testbench when # of test images is 100
- sim/tb/core_tb_1000.sv: testbench when # of test images is 1000
- sim/tb/core_tb_10000.sv: testbench when # of test images is 10000

File Description (continue)

KAIST EE

- syn/run_dc.sh: synthesis script

- pnr/run_icc2.sh: pnr script

Outline

1. Introduction

2. Baseline HW + SW

3. Guidelines

Presentation Guidelines

KAIST EE

- 16 teams are split into two days, respectively (**June 3rd and June 5th**).
  - **1 to 8 teams** will present **on June 3rd**, and **9 to 16 teams** will present **on June 5th**.

- Each group must present within **5 minutes** and answer the questions **for 3 minutes**.

- Freely describe what technique you used and the performance results

- Language
  - Slide: English
  - Presentation: Korean or English

Report Guidelines

KAIST EE

- Due (**June 12th**)
  - Upload to KLMS
- Each group must submit within **3 pages.**
- Please submit all files used in the final project, **including parameter.txt, RTL, TB**, etc, as a tar/zip/gz file.
- Language
  - Korean or English

Report Guidelines (Continue)

KAIST EE

- Criteria
  - Motivation
  - Technique
  - Performance evaluation.
    - Please show the performance improvement including reference.
    - For example, you need to prove timing improvement with timing report after synthesis.
  - Reference (ex: Synthesis report, P&R report)
    - If you explain the performance improvement based on not synthesis results but P&R results, it can be evaluated more favorably.

- Caution
  - If there are any issues regarding timing violations (setup time or hold time, etc), please include the explanation of these issues in the report.