**Recreating DeePhi Benchmark Notes**

**Summary:**

Recreated numbers from Slide 7, i.e. average speedup on long audio files (>2 seconds) **~8X**; average power usage = **18 watts**, max power usage measured = **25 watts**. See output data.

**Questions:**

* Measuring error rate? Measuring power usage incrementally?
* Recreating P4, GPU benchmarks?

> The DeePhi AMI includes test audio files located in data/short\_audio/wav/, data/middle\_audio/wav/, and data/long\_audio/wav/. Used these in tests below.

> Should also take into consideration CPU usage when writing/parsing the outputs.

**Steps taken:**

1. Logged outputs from **aws\_test.py** and modified the numbers of tests it ran
   1. By default, aws\_test.py runs 1000 tests on one file using the single-test argument
2. Wrote **means.py** to parse outputs from aws\_test.py to find average use times
3. Calculated avg. speedup by dividing average CPU use time by average FPGA use time
4. Ran 25 tests per file and averaged the use times; avg. speedup = **7.629X**
5. Ran 100 tests per file and averaged the use times; avg. speedup = **8.106X**
6. Ran 10000 tests on a single file and measured power usage; avg. = **18w**, max = **25w**

**Contents:**

1. FPGA power consumption (1)
2. Command format/notes (2)
3. Example test outputs (3)
4. means.py code (4)
5. Output data (5-7)
6. Extra (8-)
7. **FPGA power consumption:** fpga-describe-local-image -S 0 -M

Average : 18 watts

Max measured : 25 watts

Last measured is almost always 18 watts, but sometimes jumps between 18 and 25; the average intermittently changes to 19 watts for a short period, perhaps denoting a cluster of power spikes.

\*\* Cloned the aws-fpga git repo located at <https://github.com/aws/aws-fpga> into /deepspeech2 for the power metrics because the version of aws-fpga on this AMI is deprecated.

1. **Command format:** (uses aws\_test.py - runs deepspeech2 multiple times on each audio file and outputs the average cpu and fpga runtimes for each file)

python aws\_test.py --fpga\_config deephi/config/fpga\_cnnblstm\_0.15.json --audio\_path data/long\_audio/wav/long\_audio3.wav --single\_test --multi\_loop\_format |& tee test\_output.txt && python means.py | tee -a means\_output.txt

**Explanations:**

python aws\_test.py --fpga\_config deephi/config/fpga\_cnnblstm\_0.15.json --audio\_path data/long\_audio/wav/long\_audio3.wav --single\_test

Runs the test itself on a single audio file, long\_audio3.wav; default number of runs is 1000.

\*\* The --fpga\_config argument is required to display both cpu and cpu+fpga use times.

--multi\_loop\_format

Outputs the results of test in comma-delimited format, which is parsed by means.py; see example test output section for more detail. Remove from command for detailed outputs.

\*\* Seems to only work with the --single\_test argument from above.

|& tee test\_output.txt

Outputs the results of all tests for one file to test\_output.txt; overwrites test\_output.txt with new test results when it reaches new file, after calling means.py.

Also outputs to console.

&& python means.py

Parses test\_output.txt, and finds the average use times for CPU and FPGA; code for means.py located below.

| tee -a means\_output.txt

Appends the average use times to means\_out.txt.

Also outputs to console.

1. **Example test output:**

Normal output from running aws\_test.py with --fpga\_config and --single\_test arguments:

===1256th test===

CNN use time : 0.037392377853393555 seconds

LSTM use time : 0.44629526138305664 seconds

FC use time : 0.00014209747314453125 seconds

SoftMax use time : 9.369850158691406e-05 seconds

decode use time : 0.00014472007751464844 seconds

Forward use time : 0.48421573638916016 seconds

[CPU] IT'LL BE NO DISAPPOINTMENT TO ME

[CPU] Decoded 2.18 seconds of audio in 0.518734 seconds

FPGA use time : 0.04261279106140137 seconds

FC use time : 0.0009109973907470703 seconds

SoftMax use time : 0.00010657310485839844 seconds

decode use time : 0.00013756752014160156 seconds

Forward use time : 0.043863534927368164 seconds

[FPGA] IT'LL BE NO DISAPPOINTMENT TO ME

[FPGA] Decoded 2.18 seconds of audio in 0.078382 seconds

Shortened output when adding the --multi\_loop\_format argument:

1256,0.037392377853393555,0.44629526138305664,0.00014209747314453125,9.369850158691406e-05,0.00014472007751464844,0.48421573638916016,0.518734,1.487

0.04261279106140137,0.0009109973907470703,0.00010657310485839844,0.00013756752014160156,0.043863534927368164,0.078382,0.166

\*\* This is the same as above, just without the formatting, and with the totals appended as the last value; **much easier to parse**. First line is CPU values, second line is FPGA values.

\*\* --multi\_loop\_format seems to only work with the --single\_test argument.

1. **means.py:** (used to calculate average CPU and FPGA runtime)

import numpy as np

total\_cpu = []

total\_fpga = []

odd = True          # used to parse each pair of lines as one pair; true = CPU, false = FPGA

tests = 0

total\_tests = 1000  # total number of tests ran

# parse output

with open('test\_output.txt') as f:

    # skip the first 9 lines of output file

    for \_ in range(9):

        next(f)

    # then iterate through the file; use short output format

    for line in f:

        if tests < total\_tests:

            # parse each pair of lines

            last = float(line.split(',')[-1])

            if odd:

                total\_cpu.append(last)

            else:

                total\_fpga.append(last)

                tests += 1

            odd = not odd

# calculate/output usetime averages and standard deviations

avg\_cpu = np.mean(total\_cpu)

std\_cpu = np.std(total\_cpu)

avg\_fpga = np.mean(total\_fpga)

std\_fpga = np.std(total\_fpga)

# output detailed info

print('Average CPU runtime : ' + '%.3f'%(avg\_cpu) + ' +- ' + '%.3f'%(std\_cpu))

print('Average FPGA runtime : ' + '%.3f'%(avg\_fpga) + ' +- ' + '%.3f'%(std\_fpga))

print('Average speedup: ' + '%.3f'%(avg\_cpu/avg\_fpga) + 'x\n')

# output only the speedup

#print('%.3f'%(avg\_cpu/avg\_fpga))

1. **Data:**

**25 tests per file:**

Average speedup for short audio files (<1 s): **3.983x**

Average speedup for middle audio files (1-2 s): **5.674x**

Average speedup for long audio files (>2 s): **7.629x**

**100 tests per file:**

Average speedup for short audio files (<1 s): **4.043x**

Average speedup for middle audio files (1-2 s): **6.324x**

Average speedup for long audio files (>2 s): **8.106x**

**Outputs for 25 tests per file:**

\*\* Each set of 3 outputs below, i.e. CPU use time, FPGA use time, and speedup, corresponds to the decode tests of one file in the directory listed.

Short audio directory: data/short\_audio/wav/ (5 files)

Average CPU runtime : 0.164 +- 0.002 seconds

Average FPGA runtime : 0.043 +- 0.001 seconds

Average speedup: 3.835x

Average CPU runtime : 0.205 +- 0.043 seconds

Average FPGA runtime : 0.045 +- 0.001 seconds

Average speedup: 4.593x

Average CPU runtime : 0.169 +- 0.013 seconds

Average FPGA runtime : 0.046 +- 0.001 seconds

Average speedup: 3.678x

Average CPU runtime : 0.190 +- 0.037 seconds

Average FPGA runtime : 0.045 +- 0.001 seconds

Average speedup: 4.256x

Average CPU runtime : 0.168 +- 0.001 seconds

Average FPGA runtime : 0.047 +- 0.001 seconds

Average speedup: 3.555x

Average speedup for short audio files: **3.983x**

Middle: data/middle\_audio/wav/ (20 files)

Average CPU runtime : 0.375 +- 0.001 seconds

Average FPGA runtime : 0.070 +- 0.001 seconds

Average speedup: 5.342x

Average CPU runtime : 0.592 +- 0.088 seconds

Average FPGA runtime : 0.077 +- 0.001 seconds

Average speedup: 7.690x

Average CPU runtime : 0.393 +- 0.001 seconds

Average FPGA runtime : 0.075 +- 0.001 seconds

Average speedup: 5.249x

Average CPU runtime : 0.548 +- 0.001 seconds

Average FPGA runtime : 0.213 +- 0.001 seconds

Average speedup: 2.578x

Average CPU runtime : 0.391 +- 0.001 seconds

Average FPGA runtime : 0.074 +- 0.001 seconds

Average speedup: 5.296x

Average CPU runtime : 0.735 +- 0.095 seconds

Average FPGA runtime : 0.204 +- 0.001 seconds

Average speedup: 3.602x

Average CPU runtime : 0.599 +- 0.088 seconds

Average FPGA runtime : 0.076 +- 0.001 seconds

Average speedup: 7.865x

Average CPU runtime : 0.412 +- 0.001 seconds

Average FPGA runtime : 0.076 +- 0.001 seconds

Average speedup: 5.396x

Average CPU runtime : 0.739 +- 0.045 seconds

Average FPGA runtime : 0.196 +- 0.000 seconds

Average speedup: 3.764x

Average CPU runtime : 0.636 +- 0.059 seconds

Average FPGA runtime : 0.078 +- 0.000 seconds

Average speedup: 8.170x

Average CPU runtime : 0.564 +- 0.094 seconds

Average FPGA runtime : 0.074 +- 0.001 seconds

Average speedup: 7.623x

Average CPU runtime : 0.487 +- 0.001 seconds

Average FPGA runtime : 0.179 +- 0.001 seconds

Average speedup: 2.726x

Average CPU runtime : 0.394 +- 0.001 seconds

Average FPGA runtime : 0.074 +- 0.001 seconds

Average speedup: 5.293x

Average CPU runtime : 0.469 +- 0.102 seconds

Average FPGA runtime : 0.077 +- 0.001 seconds

Average speedup: 6.081x

Average CPU runtime : 0.622 +- 0.075 seconds

Average FPGA runtime : 0.078 +- 0.000 seconds

Average speedup: 7.935x

Average CPU runtime : 0.405 +- 0.001 seconds

Average FPGA runtime : 0.077 +- 0.001 seconds

Average speedup: 5.277x

Average CPU runtime : 0.410 +- 0.062 seconds

Average FPGA runtime : 0.074 +- 0.001 seconds

Average speedup: 5.527x

Average CPU runtime : 0.551 +- 0.039 seconds

Average FPGA runtime : 0.078 +- 0.001 seconds

Average speedup: 7.107x

Average CPU runtime : 0.395 +- 0.001 seconds

Average FPGA runtime : 0.075 +- 0.001 seconds

Average speedup: 5.278x

Average speedup for middle audio files: **5.674x**

Long: data/long\_audio/wav/ (5 files)

Average CPU runtime : 1.575 +- 0.105 seconds

Average FPGA runtime : 0.293 +- 0.001 seconds

Average speedup: 5.384x

Average CPU runtime : 1.375 +- 0.157 seconds

Average FPGA runtime : 0.143 +- 0.000 seconds

Average speedup: 9.645x

Average CPU runtime : 1.180 +- 0.120 seconds

Average FPGA runtime : 0.141 +- 0.001 seconds

Average speedup: 8.355x

Average CPU runtime : 1.074 +- 0.231 seconds

Average FPGA runtime : 0.141 +- 0.001 seconds

Average speedup: 7.598x

Average CPU runtime : 1.007 +- 0.205 seconds

Average FPGA runtime : 0.141 +- 0.001 seconds

Average speedup: 7.165x

Average speedup for long audio files: **7.629x**

1. **Extra:**

Data (100 runs per file) // Speedup per file:

Short: 3.804,3.817,4.263,4.516,3.813 (avg speedup = 4.043x)

Middle: 5.340,7.178,6.081,7.265,6.155,6.162,5.476,6.347,7.313,6.475,7.302,6.198,6.633,7.104,5.437,5.469,7.316,5.463,5.438 (avg speedup = 6.324x)

Long: 9.613,6.525,9.101,8.828,6.462 (avg speedup = 8.106x)

Setup commands:

sudo bash

source /opt/Xilinx/SDx/2017.1.rte/setup.sh

cd ASR\_Accelerator/deepspeech2

source activate test\_py3