PIPULS: Predicting I/O Patterns Using LSTM in Storage Systems

Dr. Qing Yang
Distinguished Engineering Professor
University of Rhode Island
CSO, Shenzhen Dapu Microelectronics Ltd Co.
Storage I/O Patterns

◆ Vary greatly across different applications
◆ Accurate prediction can help
  ➢ Proper scheduling of SSD activities
  ➢ Garbage collection while idle
  ➢ Caching and prefetching etc.
  ➢ For Energy Efficiency:
    ✓ fully utilizing different power modes and fine tuning can result in energy saving
Supervised Learning Model

- **LSTM neural network:** recurrent neural network
  - Time series data: e.g. weather forecast, language modeling and speech recognition
Architecture of an LSTM Cell
PIPLULS Architecture

Collected I/O trace

Model parameter

FPGA implementation

IO features:
1. Address
2. Length
3. Frequency

Future I/O intensity

Offline Training

Online Testing

Host Computer

Flash Memory Controller

TensorFlow

Making Data Storage Smarter
Input & Output of the Model

- History window
- Sliding window
- Prediction window
- I/O intensive within 10ms in future?

I/O features:
1. Address
2. Length
3. RW ratio
4. RW interval

Current I/O

Future time slot

Input layer
4 cells:
1) I/O address
2) I/O length
3) RW ratio
4) RW interval

Hidden layer
128 cells

LSTM layer
128 cells

Output layer
1 cell:
I/O intensity
## Parameter Size of the Weight/Bias Matrix

<table>
<thead>
<tr>
<th>Layer</th>
<th>Weight matrix size</th>
<th>Numbers of weight matrix</th>
<th>Bias matrix size</th>
<th>Numbers of bias matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input layer</td>
<td>$128 \times 4$</td>
<td>1</td>
<td>$128 \times 4$</td>
<td>1</td>
</tr>
<tr>
<td>Hidden layer</td>
<td>$128 \times 12$</td>
<td>8</td>
<td>$128 \times 128$</td>
<td>8</td>
</tr>
<tr>
<td>Output layer</td>
<td>$128 \times 1$</td>
<td>1</td>
<td>$128 \times 1$</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total size</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>264K parameters</strong></td>
</tr>
</tbody>
</table>
HARDWARE ARCHITECTURE
Figure 9: The prediction result of three storage I/O traces (configuration: history=1K, predict period=1ms, precision=float 16 bit)
Making Data Storage Smarter

Flash Memory Summit 2019
Santa Clara, CA
## Hardware Resource Usage

<table>
<thead>
<tr>
<th></th>
<th>Xilinx VU9P</th>
<th>PIPULS FP32</th>
<th>PIPULS FP16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logic cells</td>
<td>2586K</td>
<td>2.1K</td>
<td>1.8K</td>
</tr>
<tr>
<td>LUT</td>
<td>1182K</td>
<td>4.1K</td>
<td>1.3K</td>
</tr>
<tr>
<td>Flip flop</td>
<td>2364K</td>
<td>9.5K</td>
<td>6.5K</td>
</tr>
<tr>
<td>DSP block</td>
<td>6,840</td>
<td>256</td>
<td>128</td>
</tr>
<tr>
<td>Block RAM</td>
<td>28MB</td>
<td>1MB</td>
<td>0.5MB</td>
</tr>
</tbody>
</table>
Summary and Conclusions

- **A New LSTM Model**
  - Predicting I/O Patterns
    - Highly Accurate
    - High Speed

- **Working Prototype**
  - Hardware inference
    - FPGA implementation
    - Low resource usage
  - Software Training