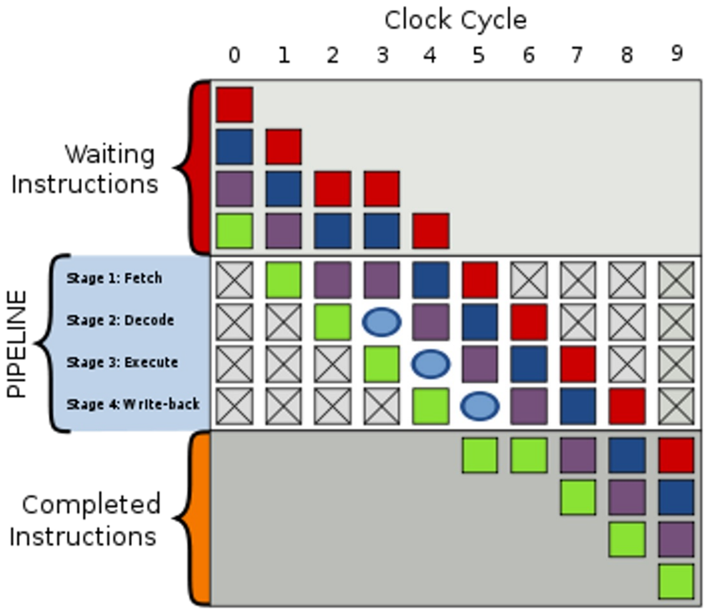
CS510 Project Presentation | Jun 4, 2019

Budget-aware Neural Network Branch Prediction

Jaeseok Huh

School of Computing

Motivation I. Deeper Pipelining and Popular Usage



HyperTransport™ CONSORTIUM

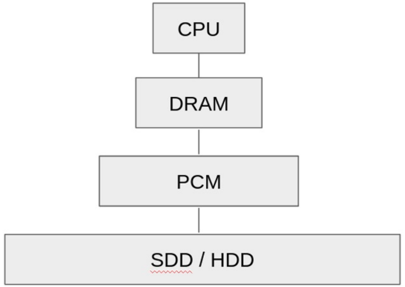
AMD

SAMSUNG

intel

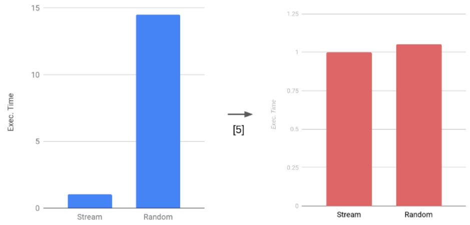
Motivation II.

- A well-defined problem with a publicly available dataset & simulator



Hierarchically [10]

Memory Performance Attack (or Fairness)



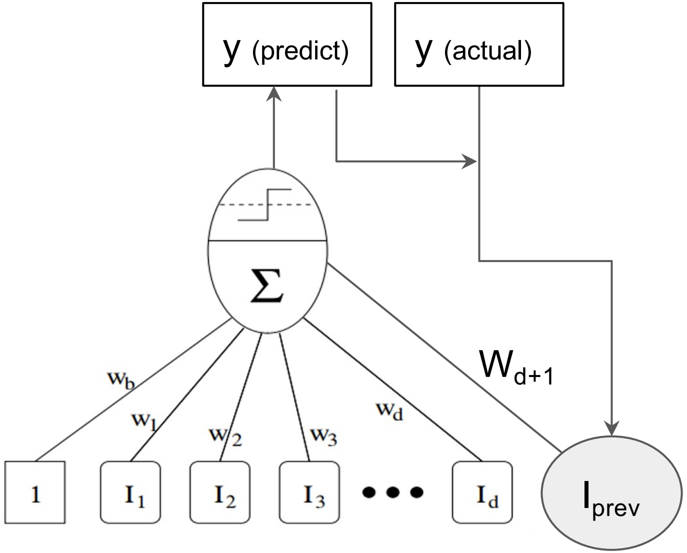
On 16-core System

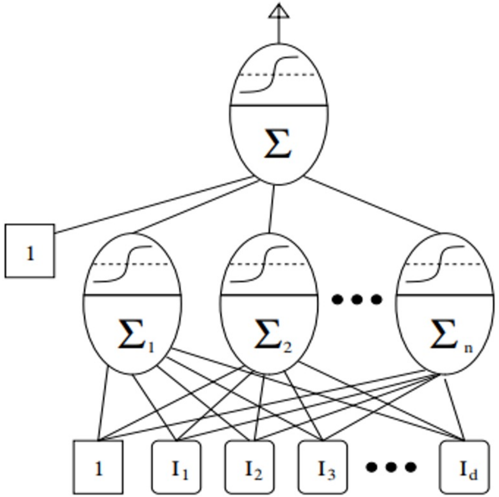
Original Proposal - *Guaranteeing Fairness in Hybrid-Memory System*

Approach

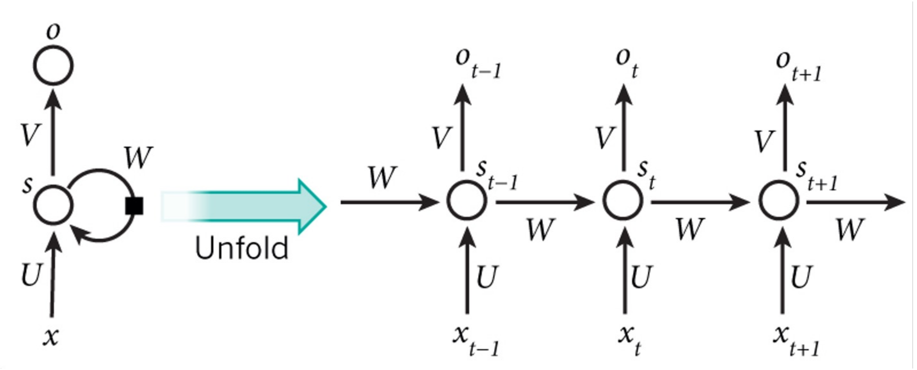
- Perceptron
- 2-layer Network
- RNN *(in progress)*
- TAGE-SC-L [2]

Perceptron & 2-layer Network





RNN - Real Time Recurrent Learning(RTRL)

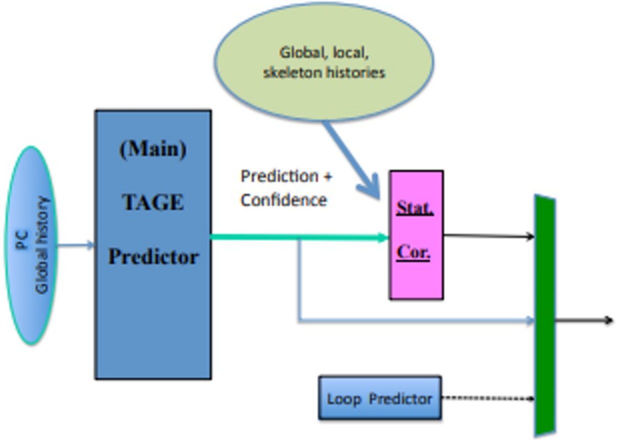


$$\frac{\partial u_k^{(T+1)}}{\partial w_{ij}} = f'(s_k^{(T)}) \left[\sum_{v\in U} w_{kv} \frac{\partial u_v^{(T)}}{\partial w_{ij}} + \delta_{ik} z_j^{(T)} \right]$$

Approximation for real-time training

TAGE-SC-L, 2016 [2]

- Find longest matching history
- 2-stage process with prediction as well as confidence for Statistical Corrector



Implementation

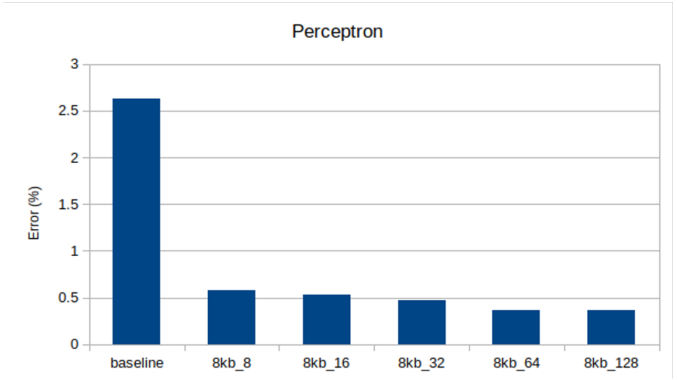
- Hardware budget (Storage) is limited to 8KB and 32KB
- The size of hash table is not less than that of input to the neural network
- # of bits for each neuron is also a design choice
- Learning rate must be tuned
- Circuitry-level implementation is left for future work
    - e.g. using SimpleScalar [3]

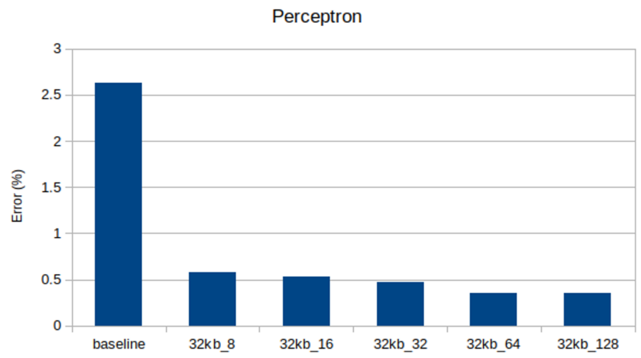
Experimental Methodology

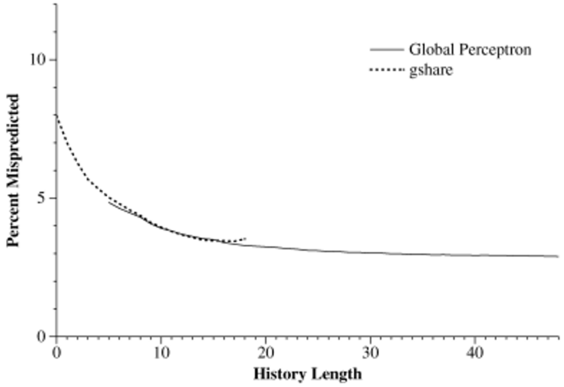
- Workloads & Simulator - Championship Branch Prediction(CBP-5), 2016
    - from 5th JILP Workshop on Computer Architecture Competitions (JWAC-5)
    - Workloads - **Random(4.9B; today)** / Small(today) / Large / **All (43B; final report)**
    - Simulator - Modified

- Criteria - Prediction Accuracy (%), Hardware Budget (KB)

Result: Perceptron

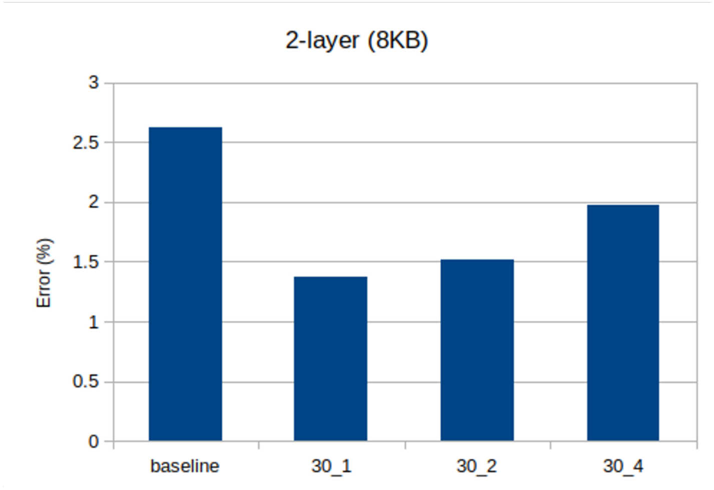


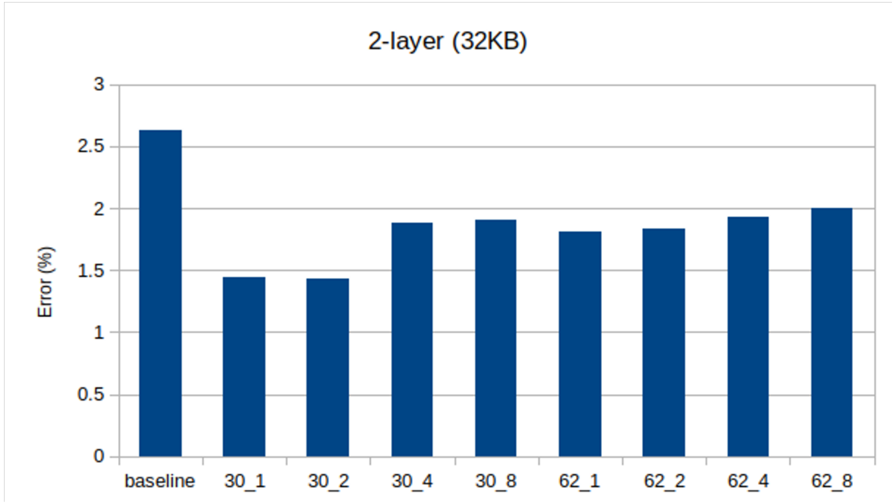




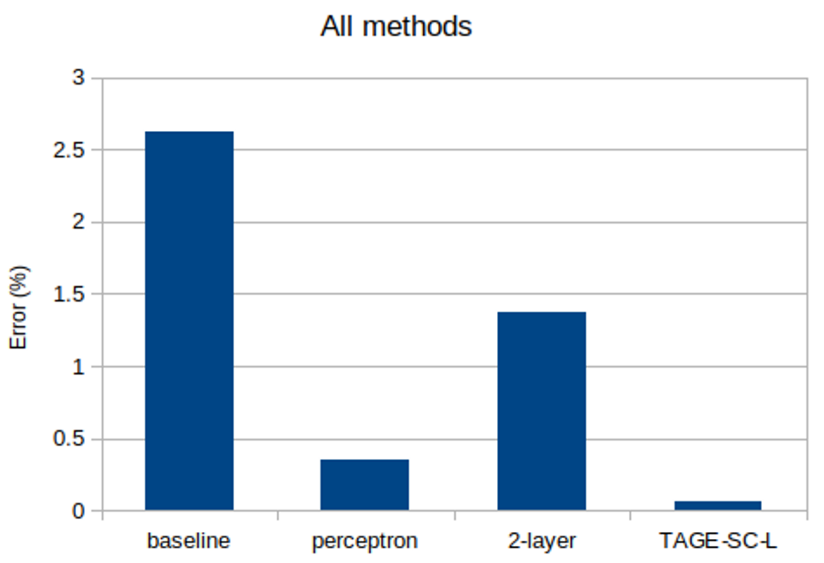
[3]

Result: 2-layer





Result: All Methods



Conclusion

- Perceptron performs better than 2-layer network despite its small size.
- TAGE-SC-L outperforms the others.
- 64-128 length of global branch history is needed for prediction.

Lessons

- In research, it is paramount to find dataset and evaluation method (or simulator).
- Testing with a small size of data saves loads of time.
- A quite large proportion of research related to HW is conducted in private businesses

References

[1] Smith, Andrew. "Branch Prediction with Neural Networks: Hidden layers and Recurrent Connections." Department of Computer Science University of California, San Diego La Jolla, CA 92307 (2004).

[2] Seznec, André. "Tage-sc-l branch predictors." JILP-Championship Branch Prediction. 2014.

[3] Jiménez, Daniel A., and Calvin Lin. "Neural methods for dynamic branch prediction." ACM Transactions on Computer Systems (TOCS) 20.4 (2002): 369-397.

- More references will be included in the final report. Only necessary ones are present here.