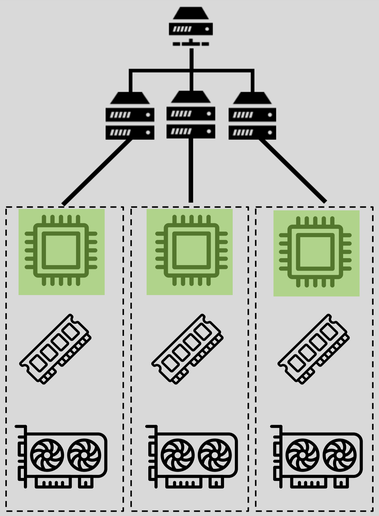
Presenter: Raghavendra Pradyumna Pothukuchi
Michael Wu
Ketaki Joshi
PI: Abhishek Bhattacharjee
Computer Science
Yale University
raghav.pothukuchi@yale.edu

Brain-Inspired Memory Page Prefetching for Accelerators in Edge Data Centers

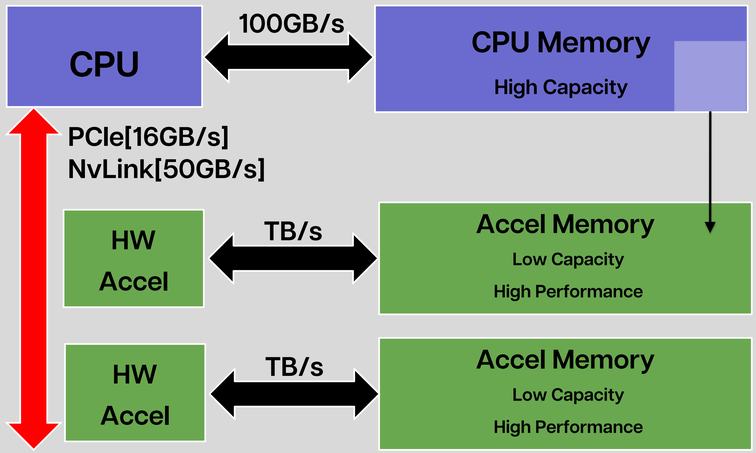
ATHENA

1. Accelerator Memory is Complicated

- Manual data orchestration is not scalable
- Unified address spaces can simplify memory management
- Transparent paging requires intelligent mechanisms for performance



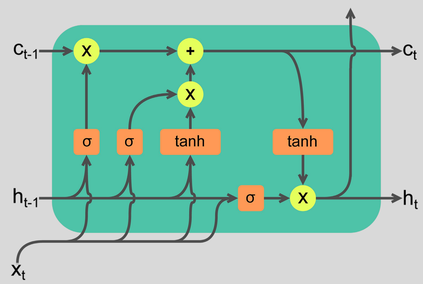
2. Transparent Prefetching



- Prefetching can automatically orchestrate data movement
- But it is challenging...

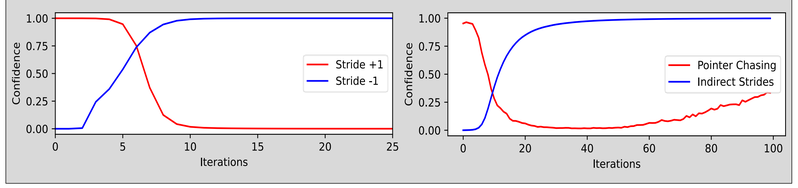
3. Machine Learning for Prefetching

- DNNs approaches are accurate but unrealistic
- High storage (>1GB) and compute (>60µs) overhead
  - Shi et al., *ASPLOS '21*



4. Online Learning is Difficult

- Complex deployments require online adaptation
- Neural networks encounter *catastrophic interference*

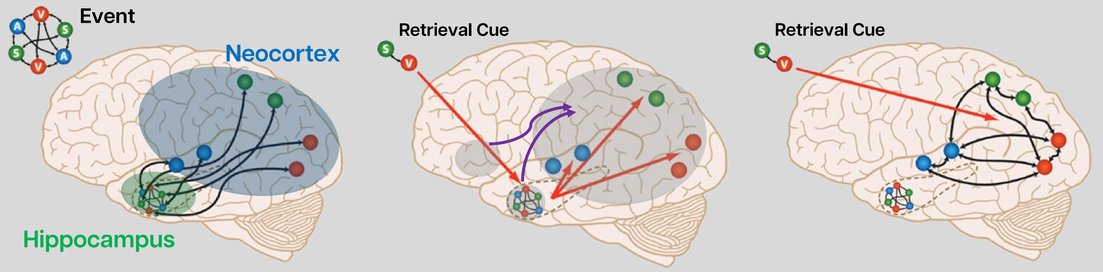


5. Brain-Inspired Prefetching

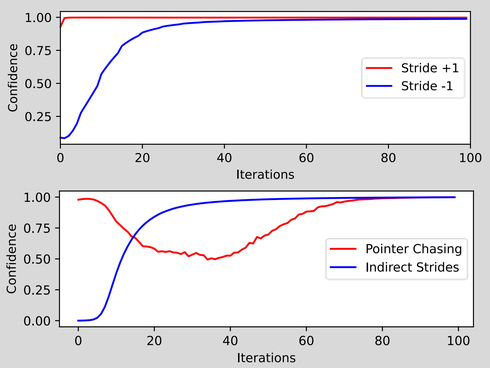
Wu et al., *HotOS '23*

Complementary Learning Systems

- Memory jointly handled between two systems



Step 1: Recording | Step 2: Replay | Step 3: Consolidation



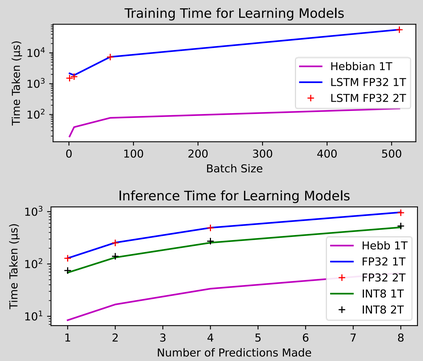
Replay remedies interference!

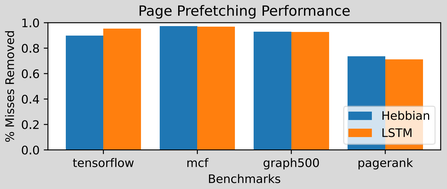
Hebbian Learning

- Simple pairwise update rule reduces compute and storage costs for learning
- For input neuron a_i and output b_j:

$$\Delta w_{ij} = (b_j \neq 0)[(a_i \neq 0) - (a_i == 0)]$$

- Sparsity further reduces storage
- Our network is <0.5MB!





Matched prefetching performance with $< \frac{1}{10}th$ training and inference cost!

Summary: Brain-inspired techniques can efficiently manage data movement.