The Edward S. Rogers Sr. Department of Electrical & Computer Engineering
UNIVERSITY OF TORONTO

Bit-Tactical: Exploiting Ineffectual Computations in Convolutional Neural Networks: Which, Why, and How

Alberto Delmas, Patrick Judd, Dylan Malone Stuart, Zissis Poulos, Mostafa Mahmoud, Sayeh Sharify, Milos Nikolic, Andreas Moshovos

Abstract

Bit-Tactical:
Deep neural network inference accelerator

- Targets primarily CNNs, can do any layer type
- Exploits *sparsity* and *effectual bit content*

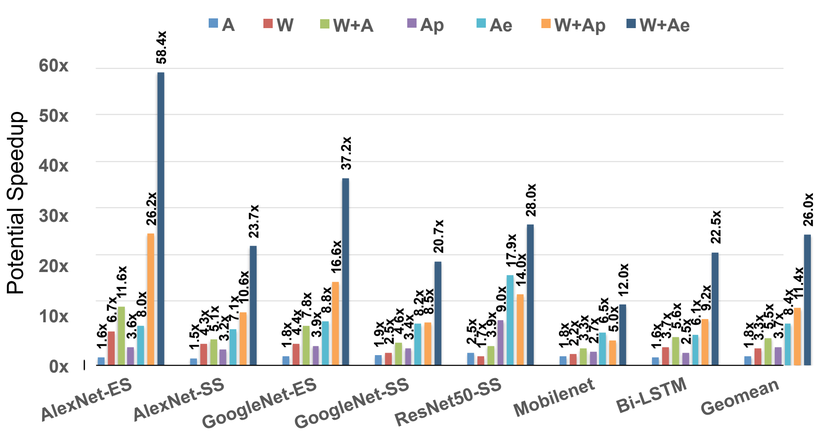
Over 10x faster than data-parallel accelerators

2x more energy efficient

32% area cost

Motivation

Previous accelerators exploit data sparsity in one dimension (zero-values, precision variability, zero-bits; on either Weights or Activations). Combining zero-value Weight sparsity along with zero-bit Activation sparsity exploitation provides the best performance potential for sparse networks, while still benefiting dense ones.

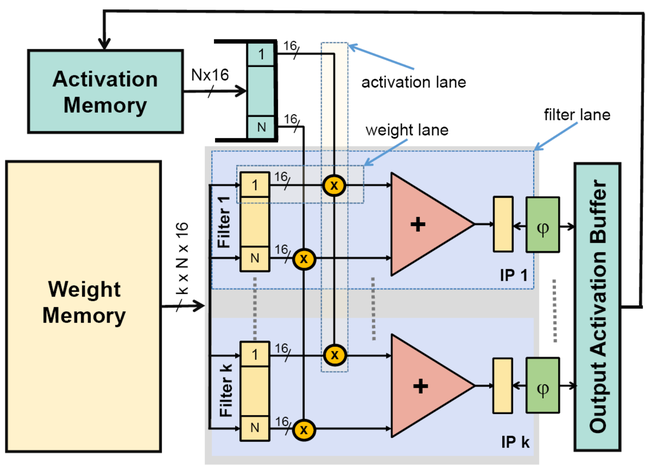


Most computation is ineffectual: over 31x potential

Baseline

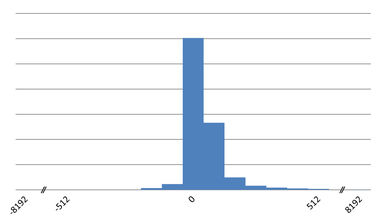
Straight-forward implementation

Synchronized parallel units that do not skip any work:

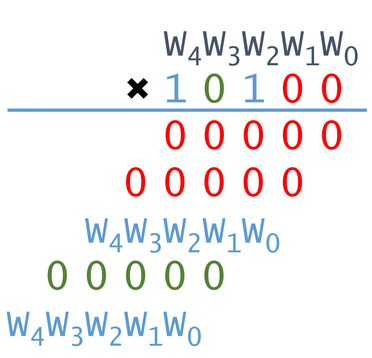


Backend: precision adaptability

Over **90%** of Activation bits are **ZERO**



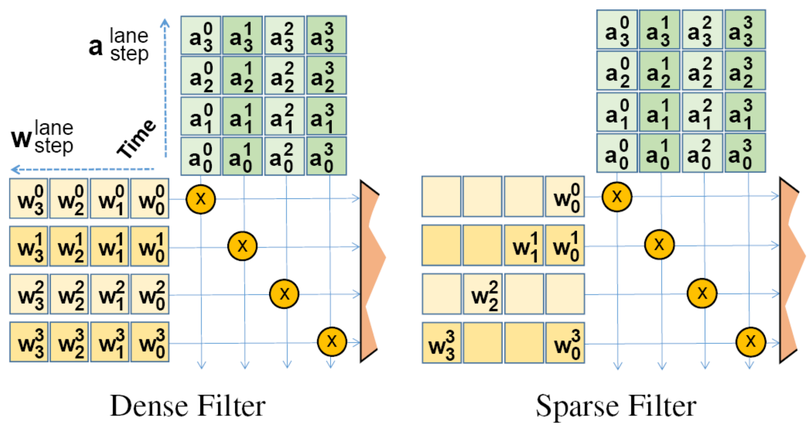
Distribution of activation values using a 16-bit fixed point representation.



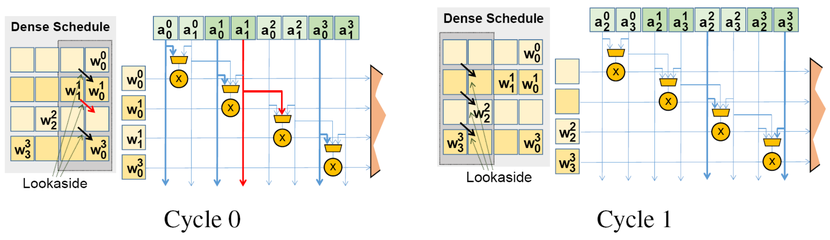
Bit-Tactical can work with multiple backends to take advantage of this: ***TCLp*** uses a bit-serial synchronized backend which skips zeros at both ends of the value; ***TCLe*** uses a term-serial backend that only processes *effectual terms*.

Frontend: Skipping 0-Weights

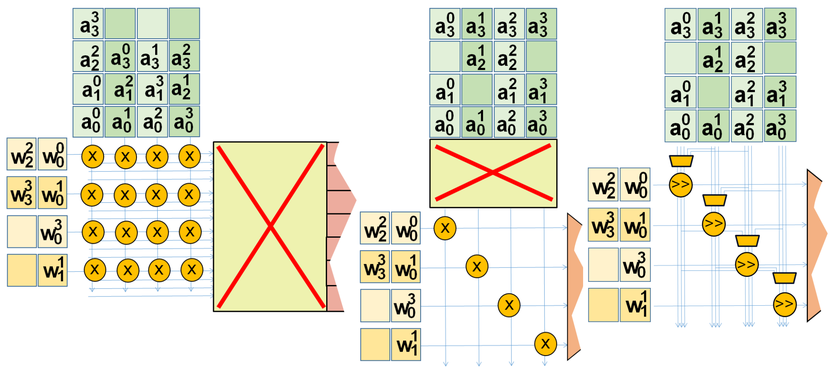
Sparsified networks present ineffectual computation in the form of zero-valued weights. A data-parallel accelerator will not take advantage of this, taking **4** cycles in this example:



Bit-Tactical borrows effectual weights temporally (from the future) and spatially (from different coordinates within the same filter). It only needs **2** cycles to process this example:

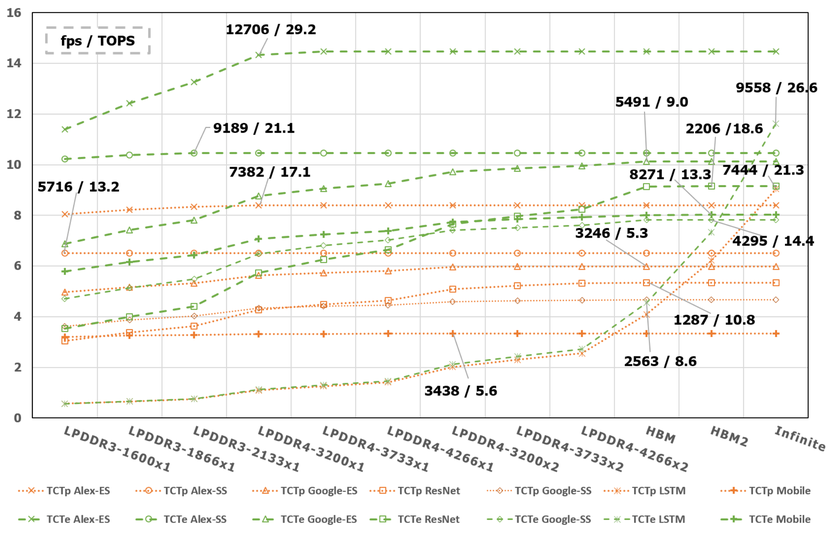


No-Crossbar design



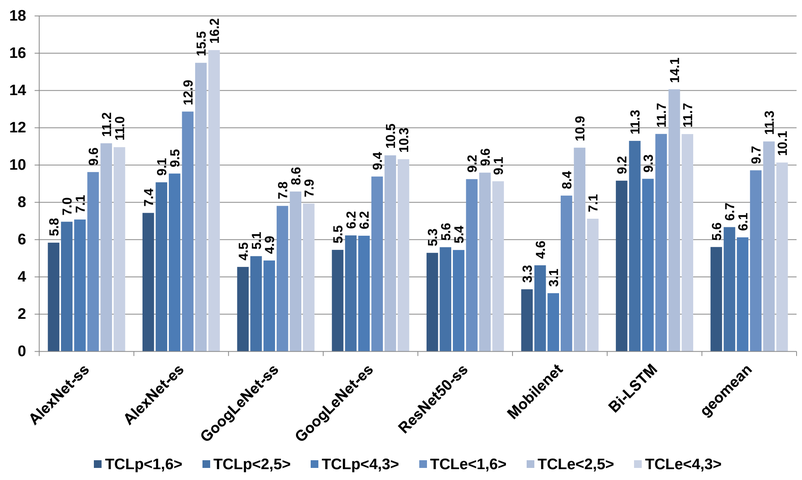
Instead of using crossbars to shuffle data around (be it the outputs – left; or one of the inputs – center), **Bit-Tactical** (right) utilizes small, local muxes to select effectual term pairs.

Off-Chip bandwidth characterization

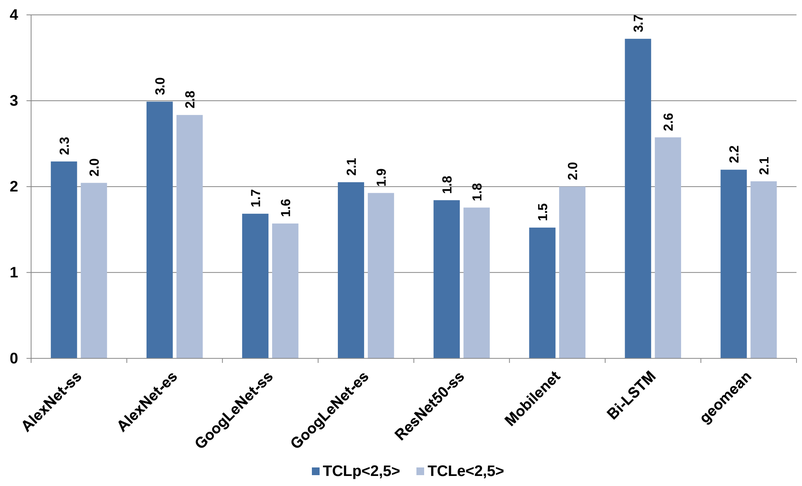


Results

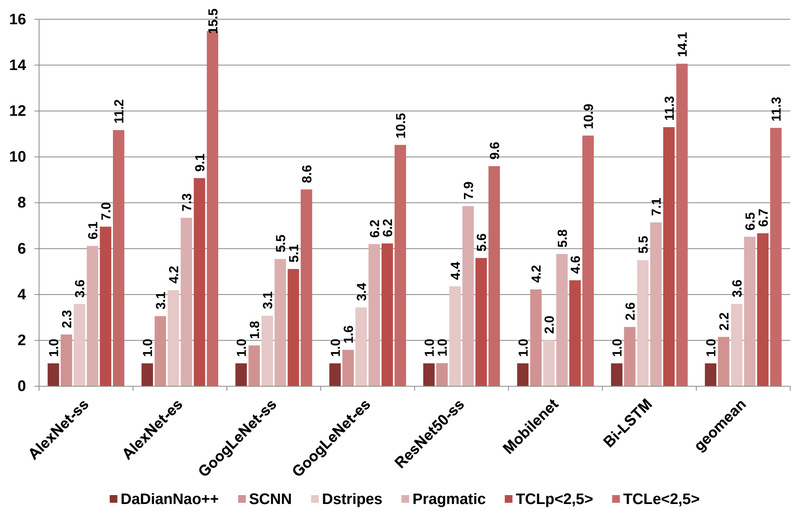
Throughput:



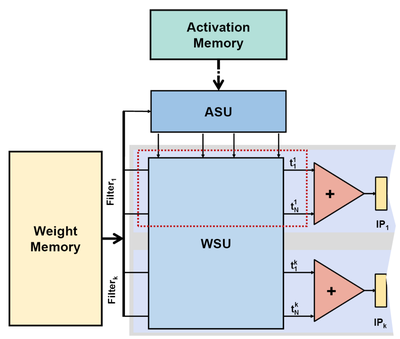
Energy efficiency:



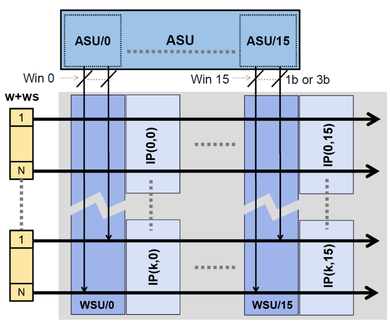
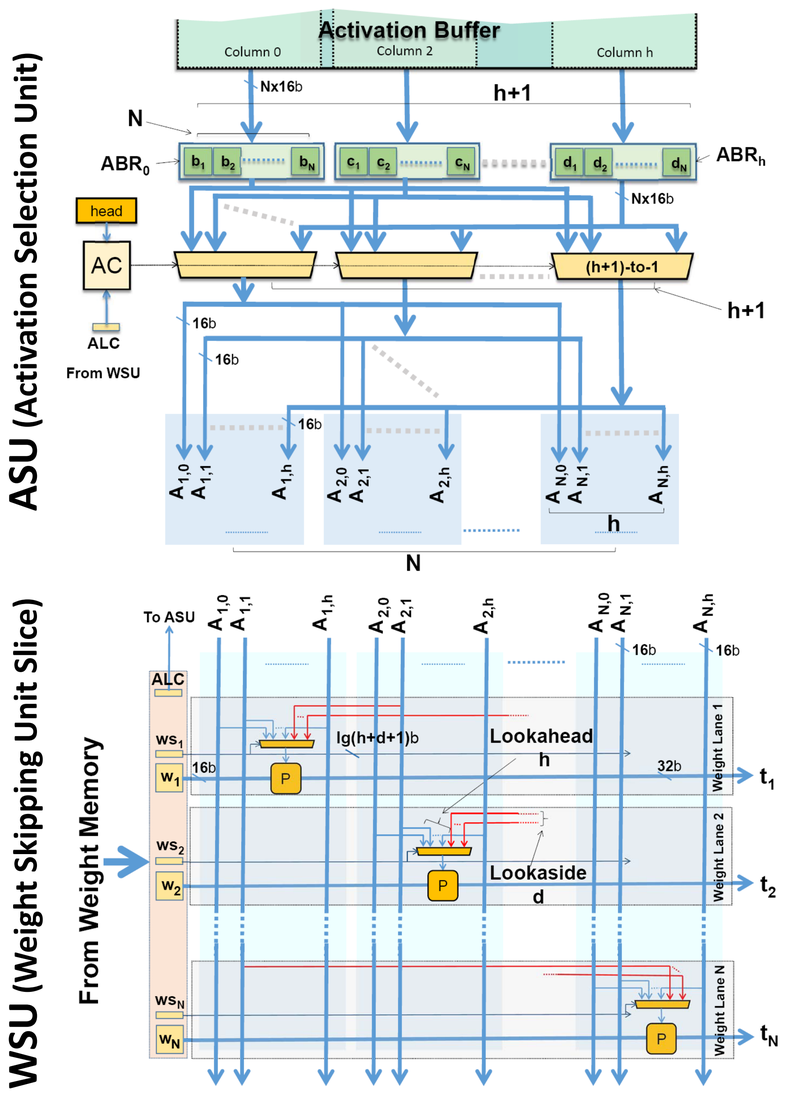
Comparison with other accelerators:



Organization



Tile: general view



TCLe/TCLp tile enabling serial activation processing