HammerBlade: A Supercomputer for ML & Graphs

Michael Taylor[1], Luis Ceze[1], Mark Oskin[1], Adrian Sampson[2], Chris Batten[2], Zhiru Zheng[2]
[1]University of Washington, [2]Cornell University

Specialized Functions: Software Defined Hardware (SDH)

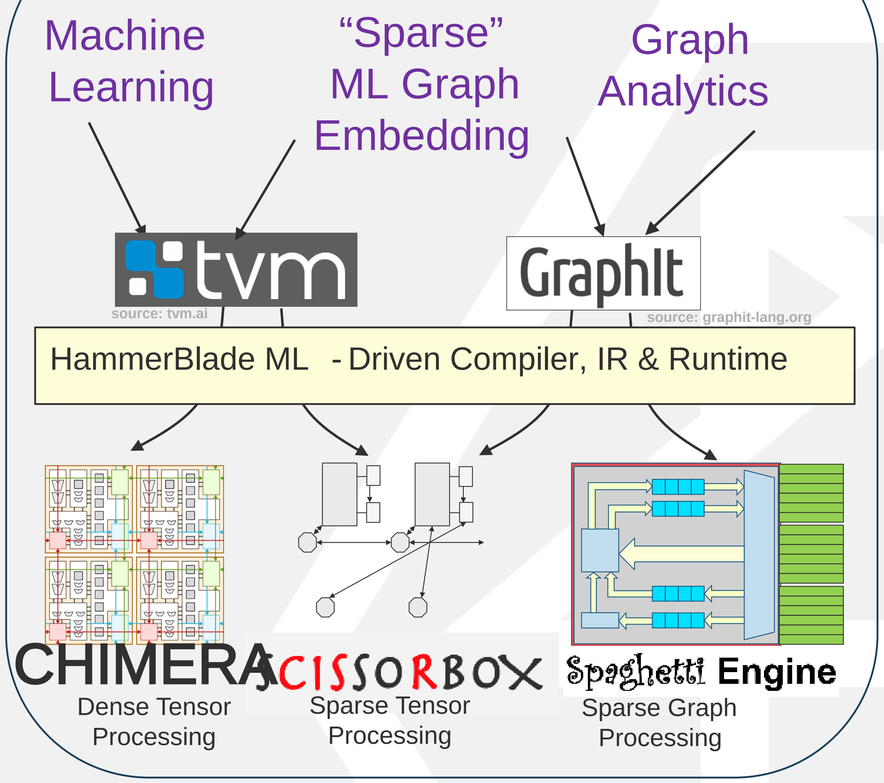
Defining Challenges

ML workloads are rapidly evolving but require extreme specialization to be feasible.

For example, ML is being combined with Graph computations to create new classes of revolutionary algorithms.

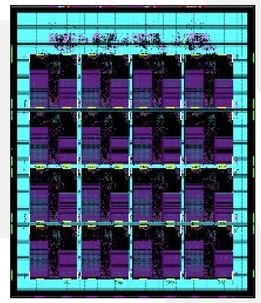
New architectures must attain this specialization while remaining sufficiently flexible.

Software Architecture



Accomplishments

Pytorch/TVM & Graphit Codes Running On Emulated HW



GF 14nm
Silicon Tapein

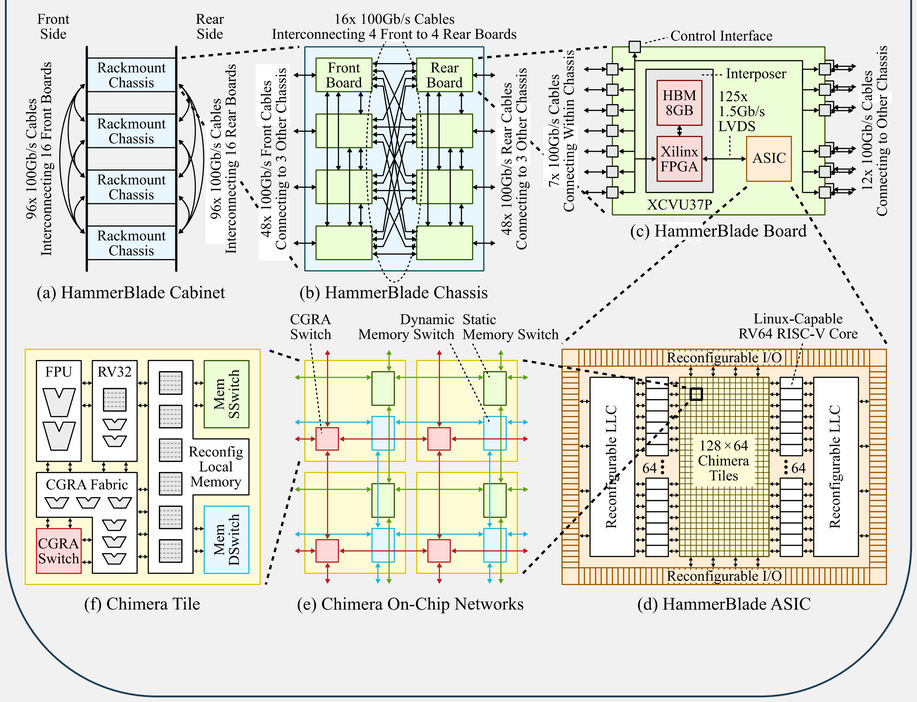
Anticipated Impact

Improves energy efficiency by 300X

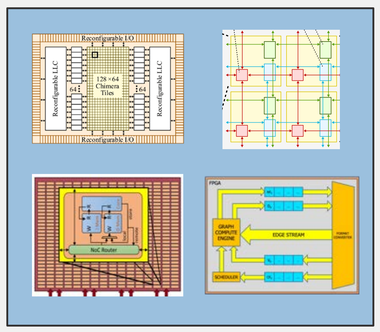
A HW Architecture that is future-proofed for rapidly evolving ML/Graph hybrid workloads.

Hardware Architecture

Overall Goal: Gate-to-Rack Scale HW & SW Specialization



HW/SW Transition Model Using Amazon Cloud FPGAs



Hardware Developers

Amazon EC2

HW, Tools & SW Distribution

aws FPGA Machine Instance

Meeting Place

GraphIt + tvm

Software Developers

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official values or policies of the Department of Defense or the U.S. Government.

THE ELECTRONICS RESURGENCE INITIATIVE