Research Problem
Sparse Matrix-Vector multiplication (SpMV) is an important kernel for many applications. However, due to lack of data locality and low FLOP to memory access ratio, SpMV's performance and efficiency are very poor on regular architectures. Overcoming this problem warrants re-thinking of the way we do SpMV, both in terms of software & hardware.

Proposed Solution
We propose an algorithm-hardware co-optimization approach to gain significant performance and energy efficiency for SpMV.

SpMV does not benefit from traditional memory hierarchy.

Background
For every two computes on average SpMV requires a random access to dense vector which is generally too large for the last level cache (LLC). This results in random access to DRAM which affects the system performance and efficiency in various ways.

An Example

<table>
<thead>
<tr>
<th>BDIM x BDIM matrix, 3 NNZ per row, 1KB DRAM page, 64B cache block, double precision data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Volume: 15GB</td>
</tr>
<tr>
<td>Number of Page Opened: 20M</td>
</tr>
<tr>
<td>Needed Redundant: 31%</td>
</tr>
<tr>
<td>Additional overhead: 8%</td>
</tr>
</tbody>
</table>

Solution: Convert random accesses to streaming access.

Will HBM with COTS work?
Probable not.

The key problem with SpMV on commercial off-the-shelf (COTS) architectures is the traditional memory hierarchy. Adding 3D stacked DRAM, such as High Bandwidth Memory (HBM), will only provide more DRAM bandwidth. Current SpMV implementations on COTS platforms will still have the same poor bandwidth utilization and other pertinent issues. Therefore, any significant improvement of performance or efficiency is unlikely as long as SpMV is implemented on any architecture which is built on traditional memory hierarchy.

SpMV does not benefit from traditional memory hierarchy.

3D DRAM Based Application Specific Hardware Accelerator for SpMV
Fazle Sadi, Larry Pileggi, Franz Franchetti

Contact: fsadi@cmu.edu

Carnegie Mellon University

Experimental Results

Data Volume vs LLC

Energy Efficiency Comparison

- CPU (Kxeon Phi) + MKL: 22nm, 10MB LLC, 102 GB/s + Peak BW
- GPU (GTX 980): 28nm, HBM 76GB/s + Peak BW

Proposed Accelerator + 2-Step: 28nm, HBM 64GB/s + Peak BW

Large sparse matrices with high sparsity and various sizes from University of Florida's collection have been used to compare the energy efficiency of the proposed system against state of the art architectures. It can be seen that the proposed accelerator can achieve up to 100x more energy efficiency than the best performing COTS architecture.

Acknowledgements
This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) under agreement number HR0011-13-2-0007. The work was also sponsored by Intelligence Advanced Research Program Agency and Space and Naval Warfare Systems Center Pacific under Contract No. N66001-12-C-2008. The content, views and conclusions presented in this document do not necessarily reflect the position or the policy of DARPA or IARPA or the U.S. Government. No official endorsement should be inferred.