ABSTRACT
Sparse deep neural network (DNN) accelerators exploit the intrinsic redundancy in data representation to achieve high performance and energy efficiency. However, sparse weight and input activation arrays are unstructured, and their processing cannot take advantage of the regular data-access patterns offered by dense arrays, thus the processing incurs increased complexities in dataflow orchestration and resource management. In this work, we first present the importance of the data reduction mechanism, i.e., how partial sums are accumulated spatially or temporally, a perspective that has not been fully explored in current literature. Motivated by the reduction analysis, we propose Stitch-X, a novel DNN inference accelerator architecture that can stitch together sparse weights and input activations for parallel execution. Specifically, Stitch-X employs a novel dataflow that leverages both spatial and temporal reduction to balance energy efficiency and dataflow control complexity. Moreover, Stitch-X adopts a new runtime Parallelism Discovery Unit (PDU) to efficiently extract fine-grained parallelizable operations from irregular sparse data arrays to enable a higher performance over a wide range of input data densities and for a variety of DNN layers. Our evaluations show that Stitch-X consistently achieves a 3.8x speedup and improves energy-delay-squared-product (ED²P) by a factor of 10.3x over an efficient, dense DNN accelerator. Compared to a state-of-the-art sparse DNN accelerator, Stitch-X delivers 1.6x better performance. A silicon prototype of the Stitch-X architecture is scheduled for 2018.

1 INTRODUCTION
Deep learning [1] has emerged to be a key approach to solving complex cognition and learning problems. Deep neural networks (DNNs) in particular have become pervasive due to their successes across a variety of applications, including image recognition [2–6], object detection [7, 8], semantic segmentation [9–11], language translation [12], audio synthesis [13] and autonomous driving [14]. State-of-the-art DNNs [2–11] require up to billions of operations and hundreds of megabytes to store activations and weights, presenting substantial workloads and memory traffic.

The core computation behind DNN processing is the dot-product of input activations and weights. To obtain substantial performance and energy efficiency gains [15–20], many prior works, from research prototypes to industrial products, proposed specialized hardware to accelerate DNN processing [21–34]. Algorithmic techniques such as quantization and pruning [35, 36] can be used to sparsify data, which could be exploited to further enhance the performance of efficiency DNN processing [37–40]. A dot-product involves computing products followed by partial-sum reduction. To better understand the design complexities associated with partial-sum reduction for sparse data, we analyze the reduction mechanisms in existing DNN accelerator designs. Despite being overlooked in the well-known dataflow analysis [27], the reduction mechanism plays a key role in constructing efficient dataflows and can lead to more than 3x difference in the overall energy efficiency. Motivated by this insight, we propose Stitch-X, a sparse DNN accelerator architecture that leverages a hybrid, spatial-temporal reduction to balance energy efficiency and dataflow control complexity. To maximize the degree of parallelism for reduction, Stitch-X utilizes a novel and scalable Parallelism Discovery Unit (PDU) that dynamically stitches together input activation and weight pairs to produce readily reducible partial sums.

We evaluate Stitch-X over a suite of modern DNNs [2–4, 6] that have been pruned by methods demonstrated in [35, 36] with no accuracy loss when tested using ImageNet [41]. Stitch-X achieves up to 4.3x speedup (3.8x on average) and improves energy-delay-squared-product (ED²P) by up to 15.2x (10.3x on average) over a state-of-the-art dense DNN accelerator. To compare sparse DNN accelerators, we use the metric of Proximity to Oracle Speedup (PTOS), i.e., the achieved speedup benchmarked as a percentage of the oracle speedup. Stitch-X achieves a PTOS of 77.4%, surpassing the state-of-the-art sparse DNN accelerator by 1.6x.

2 DNN REDUCTION MECHANISM
We present a new perspective in DNN dataflow analysis: the method in which partial sums are accumulated, i.e., reduced. Spatial Reduction (SR) refers to spatial partial-sum accumulation without explicit storage during the reduction process. Given

1Oracle performance is defined as the total number of effectual multiplications. i.e., both operands are non-zero, divided by the number of available multipliers.
Temporal Reduction (TR) refers to partial-sum accumulation over time by using a single adder to accumulate one partial sum per time step over N steps. TR’s advantage is that there is less control dependence for partial-sum accumulation, since only one element is being accumulated at a time, instead of N in SR. However, from an energy perspective, TR incurs a register read and write cost each cycle, which can be significant especially when the register file is large. Examples of TR are found in TPU [21] and ShiDianNao [26].

Figure 2(b) shows a scalable PDU design that performs parallel search on IA and W arrays to find reducible IA and W pairs. The PDU operates on the channel index vectors of IA and W of size C at the same time. The PDU uses a C × C comparator matrix to search for matching IA and W channel indices in parallel. The comparator matrix produces a binary output at each junction, 1 for match and 0 for mismatch. Priority encoding is applied to each column of comparison outputs to obtain an ordered index sequence. Finally, iterative leading-one detection is used to locate the addresses for fetching W operands from the W register file (RF). A similar approach is used to locate the addresses for fetching IA operands from the IA RF. To target a higher degree of data parallelism, the decode width can be increased accordingly.

### 4 EVALUATION

A prototype Stitch-X accelerator architecture was synthesized at a 1.0 GHz clock frequency in a commercial 40 nm CMOS technology. The core area is 2.7 mm² and the overhead of having PDU and decoder to stitch input operands accounts for less than 8% of the area, significantly lower than the decoding overhead reported in the previous sparse DNN accelerator [38].

Figure 3 shows the overall performance of Stitch-X running a range of modern DNN workloads, including AlexNet [2], VGG-16 [3], ResNet-50 [4], and Inception-v3 [6]. For each network, we evaluated the speedup and CE utilization of Stitch-X over an input-stationary SR baseline DNN accelerator, denoted by D. Across all the networks evaluated, Stitch-X achieves a 3.8× speedup over the dense DNN accelerator while maintaining an average CE utilization of 74%. Compared to the oracle, Stitch-X achieves a PTOS of 77%, 1.6× better than the state-of-the-art sparse accelerator that operates on both sparse W and IA operands [40]. In addition, Stitch-X also achieves at least 70% CE utilization when processing fully-connected layers and 1 × 1 convolution layers, both of which are challenging to accelerate with existing DNN accelerator designs.
REFERENCES