Reviewer-One

Summary:

The paper presents a new NN-search algorithm, that reaches the peak performance on TPUs. The paper is based on observations of hardware architectural properties and the so called roofline performance model. The algorithm is implemented for TPUs. The evaluation is done on two TPU versions using two KNN datasets, and compared to several GPU algorithms.

Strengths:

1) Nice connection between theoretical observations and practical results, 2) Good performance, 3) Can have significant practical impact

Weaknesses:

1) Limited evaluation, only evaluated on two versions on TPUs, and two datasets, 2) Would have been interesting to see how general the algorithm is, i.e., would it reach the same performance limits in other hardware platforms also?

Questions:

The algorithm descriptions (Algorithm 1 and Algorithm 2 (suppl. mtrl)) looks relatively clear and straight-forward to implement on other platforms. Can you elaborate a bit on why it would be such a substantial effort to do?

Limitations:

I think the authors have adequately addressed the limitations of their work.

Ethics Flag: No

Soundness: 3 good

Presentation: 3 good

Contribution: 3 good

Rating: 6: Weak Accept

Confidence: 4

Code Of Conduct: Yes