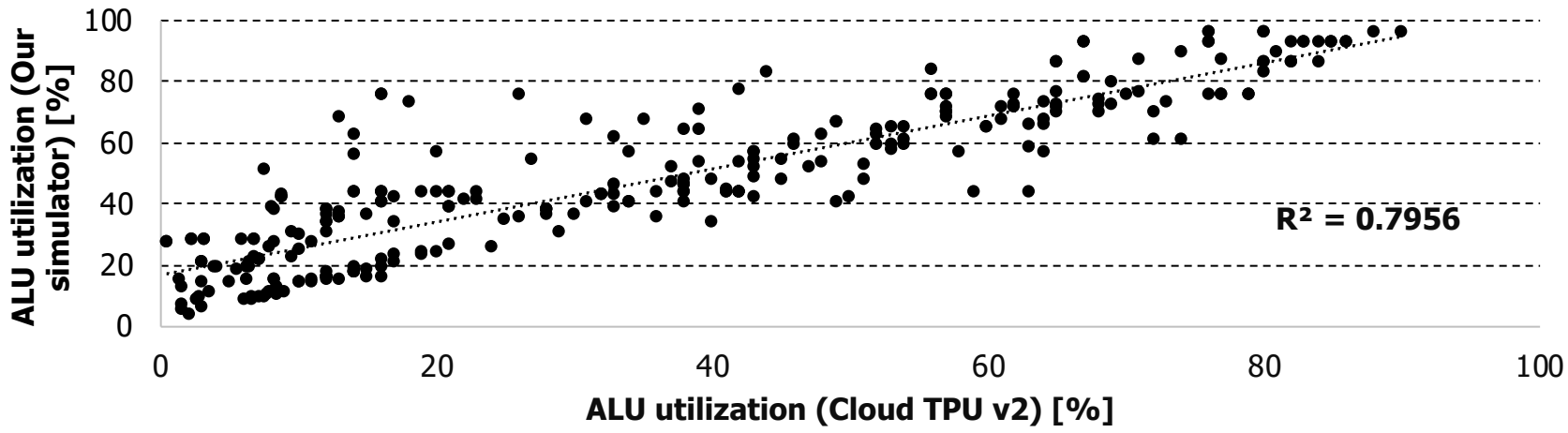
< TPU validation summary >

- Each datapoint corresponds to a single layer in our studied benchmarks in Table 2.

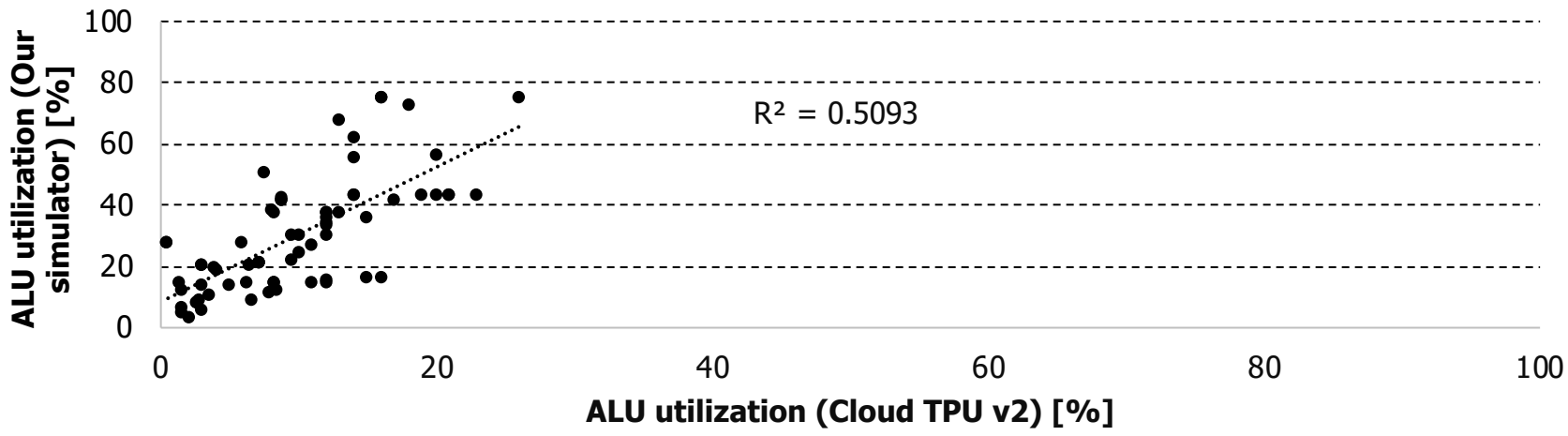
- In general, the biggest outliers are for DNNs executed with batch size '1':
  * Batch size 1: **51%** correlation
  * Batch size 4: **81%** correlation
  * Batch size 8: **86%** correlation
  * Batch size 16: **89%** correlation

- Publicly available documents from Google hints that **TensorFlow XLA compiler** heavily optimizes execution for layer configurations with low batch sizes which we are not able to reflect in our cycle-level simulator. Nonetheless, throughout the entire datapoints we validated, our model still **achieves an average 79.6% correlation** against Cloud TPUs, which we believe is reasonable given the black-box nature of TPUs.

[1] All batch size (1/4/8/16)



[2] Batch size of 1



[3] Batch size (4/8/16)