

Feasibility and Performance Analysis of Scaled-YOLOv4 Object Detection on Graphcore IPU

Arjan Siddhpura¹, S.-Kazem Shekofteh¹, Holger Fröning¹

Hardware and Artificial Intelligence (HAWAII) Lab, Heidelberg University
`arjan.siddhpura@stud.uni-heidelberg.de`, `{kazem.shekofteh, holger.froening}@ziti.uni-heidelberg.de`

Abstract. The performance of deep learning models is heavily coupled to hardware, with SIMD-based GPUs being the de facto standard. Novel architectures like the Graphcore Intelligent Processing Unit (IPU), with its Multiple Instruction, Multiple Data (MIMD) design and distributed on-chip SRAM, offer a different paradigm. However, direct performance comparisons for complex, state-of-the-art computer vision models are sparse. This paper presents a comprehensive performance benchmark of the Scaled-YOLOv4-P5 object detection model on a Graphcore GC200 IPU against a comparable NVIDIA A30 GPU. We investigate the performance trade-offs by analyzing inference latency and throughput while varying image size, batch size, and floating-point precision. Our findings reveal a stark performance trade-off. The IPU excels in low-latency scenarios, delivering a 6.56 ms inference time at batch=1 (896 px), nearly 4x faster than the GPU's 26.17 ms. Conversely, the GPU's SIMD architecture scales near-linearly for high-throughput, while the IPU is severely memory-constrained. The IPU failed to compile at batch=2 for the native 896px resolution, limited by its ~900 MB on-chip SRAM. In contrast, the GPU's 24 GiB HBM2 memory handled batches of ≥ 64 at the same resolution. Furthermore, the IPU's Ahead-of-Time compilation incurs a major overhead: a full benchmark run at 896 px took 382.79 s on the IPU versus just 15.56 s on the GPU, with 75-88% of the IPU's time spent on compilation alone.

Keywords: Object Detection · Graphcore IPU · YOLOv4 · Performance Benchmarking · Hardware Accelerators