Exact Compute For Scalable AI

James Douglas Boyd
Founder and CEO/CTO, SciSci Research

The Decade Ahead: More Scale, Without Error

Over the past decade, the greatest contributor to scaling GPU FLOPS was lower-precision number representation. Over the next decade, this trick won't be repeated: precision can only be reduced so much before model performance is harmed. Thus, the industry needs a way to yield even greater speed from AI chips without severe rounding error, and we know that number representation is the most cost-effective way to scale. The solution is a new architecture with a novel number representation for fast exact arithmetic, delivering even better chip performance without rounding error.

The Age of Low-Precision

AI companies are scaling models and compute because of scaling laws, which suggest that model error decreases as one increases data, parameters, and compute (measured in floating-point operations per second, or FLOPS). AI compute has scaled by getting more FLOPS out of GPUs. Since 2012, the greatest source of GPU FLOPS growth has been number representation for lower-precision. Indeed, NVIDIA offers GPUs that do 8-bit or even 4-bit floating-point arithmetic. This is necessary due to the floating-point tradeoff between precision and speed.

Low-Precision Won't Continue to Scale

Looking forward, it seems unlikely that low-precision compute will continue to scale AI. Benefits are diminishing, and the risks are raised as precision is lowered. Epoch AI forecasts only a 2× FLOP increase by 2030 from 8-bit floating point adoption, a rather marginal advancement. On the other hand, further precision decreases invite more serious risks, requiring either workarounds or reversion to higher precision. These risks include fewer effective parameters, reduced weight accuracy, and unstable training.

Beyond Floating-Point

The way to increase operations per second is not to go after FLOPS at all; floating-point is a poor format for representing numbers and doing arithmetic. SciSci has a number representation scheme of its own, based not on floating-point arithmetic, but on exact p-adic arithmetic. SciSci is building the first fast and errorless AI accelerator chip, the exact processing unit (EPU), to perform exact arithmetic without any rounding error, and faster than floating-point, using this alternative number representation and its own original architecture. The EPU will give perfect-precision results and 15× speed gains *per operation per ALU*. SciSci will end the precision/speed tradeoff to scale AI.