IBM Fast Fourier Transform

Accelerator type: Accelerator building block

The IBM Fast Fourier Transform (FFT) accelerator is most appropriate for applications that need to compute batches of small or medium-size FFTs. Examples include deep neural networks and IoT analytics.

Accelerator detail: The accelerator computes the one-dimensional discrete Fourier Transform on complex (float or fixed-point) data values. It is generic, and can use streaming FFT IP cores from Altera or SPIRAL.1 Using the Altera core, the AFU can support variable input sizes. The SPIRAL version can be customized to power-of-two input sizes of up to 32k.

IBM provides two options for using the accelerator: a simple low-level API, or an FFTW wrapper to use the accelerator transparently from the FFTW library.

The accelerator can be extended to support 2D FFTs with moderate additional effort.

Competitive advantage

The accelerator provides low-latency CAPI significantly faster than a CPU for small jobs. It outperforms one core for any batch size larger than 1. For larger batches, FFT is 2.2x faster and 16x more energy efficient than one core running optimized FFTW in software.

CAPI

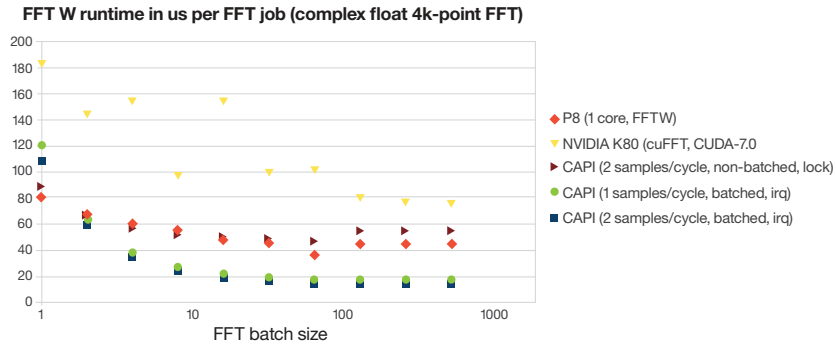

Figure 1: Runtime in microseconds for FFTW jobs on IBM® POWER8® using software FFTW, cuFFTW (on NVIDIA K80) and CAPI. The figure compares two different batched CAPI FFT versions: The two-samples/cycle FFT is based on the SPIRAL FFT IP, while the one-sample/cycle version uses the Altera core. The non-batched version has blocking semantics and signals completion after each FFT call. In the batched versions, the software labels the last FFT job in a batch with a flag, and the accelerator signals completion only after this job. In this case, multiple FFTs are computed simultaneously in pipelined fashion. The IRQ versions notify completion by an interrupt, while the lock version uses a spin-lock.

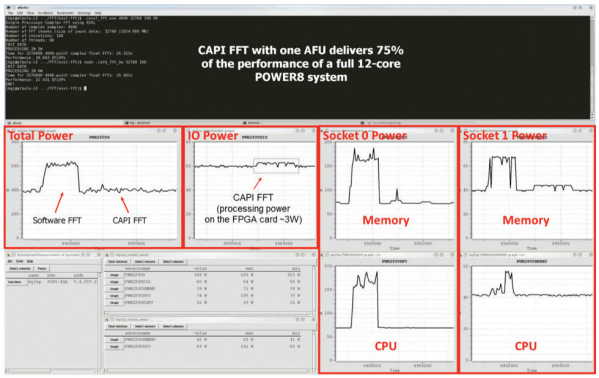

Figure 2: Power measurements for software and CAPI FFT on POWER8. The Amester tool was used to measure the power consumption of a POWER8 node when computing the FFT in software and hardware. The different windows show the power consumption of the entire node, the IO power (PCIe subsystem), and the CPU and memory power of the two processor sockets.

Experiments showed an efficiency of 0.1-0.3 GFLOP/W for the software FFT on POWER8, while the CAPI FFT delivers more than 3.3 GFLOP/W (Figure 2 shows a power trace for case 3 and 4 of the following list).

1. 1D-FFT on 1 core → 10.6 GFLOP @ 50W = 0.21 GFLOP/W
2. 1D-FFT on 12 cores (12 threads, SMT off) → 33.5 GFLOP @ 108W (DVFS off) = 0.31 GFLOP/W
3. 1D-FFT on 12 cores (96 threads, SMT8) → 30.6 GFLOP @ 193W (DVFS on) = 0.12 GFLOP/W
4. 1D-FFT on 1 AFU → 23.6 GFLOP @ 7W = 3.37 GFLOP/W

For more information, visit

www.cognitive.ptopenlab.com/accelerator/applications/2