Data caches are highly inefficient for streamed data, but they simplify programming models. With neural network and signal processing tasks, we can generally know the location and size of all data moved during an operation at compile time. The GAP AutoTiler is a C library that generates C code for all data movement using DMAs, optimizing dataflow and keeping GAP’s multiple cluster cores busy with data reducing the energy cost of inference using significant neural networks.

Tunable memory configuration options allow for different trade-offs between performance and L2 memory usage.