Ultra Low Power Inference at the very Edge of the Network

Eric Flamand, CTO, GreenWaves Technologies

Large scale deployment and operation of smart data-sensors at the very edge of the network is feasible, for cost and scalability reasons, if we assume over the air (OTA) connectivity as well as battery operation. Deep-learning based approaches enable us to efficiently solve network bandwidth issues, since the huge amount of sampled raw data can be reduced at the edge to highly qualitative data requiring very limited bandwidth on the network side. This reduction comes with a computational cost. In order to limit, as much as possible, the amount of energy to be spent to support this computational complexity, we must look at every possible source of optimizations when designing a chip for mW class inference. This is the goal of Gap8: combining many hardware architectural and implementation optimizations, together with tool driven optimizations, in order to keep memory traffic and computing activity as low as possible. In this talk we will go through the various mechanism we have been using from core ISA extension, light weight vectorization, parallelization, agile and hierarchical power management as well as tool assisted explicit memory management. We will demonstrate that with this approach we can run small to medium complexity networks as well as pre and post processing steps under a couple of mW budget.