Demo Title: **GrAI One – A Hybrid Neuromorphic and Dataflow Processor**

Author / Presenter: Jonathan Tapson, CSO, GrAI Matter Labs

Item to be demonstrated:

We will show a working system consisting of an FPGA host with a GrAI One accelerator chip. The system will be demonstrated using two applications, one of which is a voice-controlled game, and the other of which is a self-steering car application.

Relevance to TinyML:

Edge applications of ML differ from cloud or datacenter applications in that they are likely to process real-time data streams. They are often characterized as "Batch = 1", because they require processing of each input sample in real time. In this presentation, we will show that many edge ML applications are in fact "Batch << 1", in that the amount of information that changes from one data sample to the next is usually very small relative to the size of the sample. For example, video acquired from the forward-facing camera in a car changes very little from frame to frame; at 240fps, we find that more than 99% of pixels do not change value between any two frames.

In this type of processing, it becomes possible to exploit three forms of sparsity to reduce the computational load – sparsity of information in the data, sparsity of changes in time, and sparsity of activations in the network. In order to exploit this sparsity, we require algorithms and hardware which are entirely different than that which is optimal for cloud compute.

GrAI Matter Labs has developed a new hardware architecture, called NeuronFlow, for machine learning in edge applications. The architecture is based on a hybrid of neuromorphic and dataflow principles. This architecture is designed to work with hardware-specific algorithms to take advantage of the above-mentioned sparsity of changes in real-world data. The first chip from the NeuronFlow architecture, called GrAIOne, was produced in late 2019. This demonstration will show the GrAIOne chip in action in some representative applications.