# Popular Science Summary

Nowadays, Artificial Intelligence (AI) is widely recognized for its ability to solve simple daily problems and even program complex algorithms. AI can be used for everything from chatbots that can give you a recipe with what you have in the pantry, determining the type of flower in a picture, to intricate problem-solving to figure out the optimal logistical setup in a store. When an AI makes a decision based on some input data is called inference.

Using AI to solve such problems takes a lot of energy. Using specialized hardware that is designed for a specific AI model to reduce the energy footprint shows great promise. One way to create energy optimized hardware is to use re-programmable integrated circuits called FPGAs, which allow for the creation of specific hardware that has a single goal and can achieve that goal efficiently with low power consumption. The cloud providers of today offer a wide range of scalable FPGA environments that can be easily used to deploy large clusters of FPGAs. It is in the interest of the cloud providers to reduce power consumption in the data centers, so we expect further expansion of the cloud FPGA market.

FPGAs have limited resources and when AI models become larger, multiple FPGAs need to be connected together in a chain. The data from one FPGA needs to be transferred to the next FPGA in the chain, introducing additional latency. One of the goals of this thesis is to explore how much this additional overhead influences the resulting performance.

Combining the scalability of the cloud and the flexibility of the FPGAs, we can show that connecting multiple FPGAs to solve inference adds no significant latency, and has a relatively low impact on overall bandwidth requirements. This work done in this thesis has created a solid foundation for future research to explore interconnecting clusters of FPGAs, to solve even larger inference problems such as chatbots, or even general compute problems.