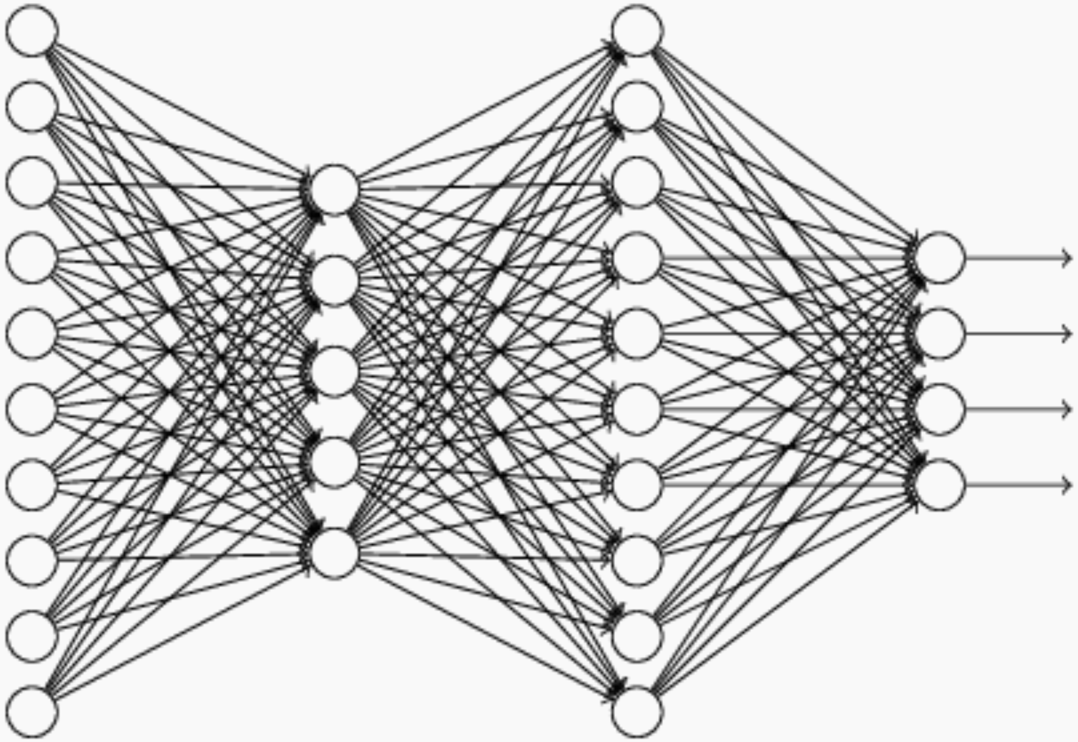
Mapping a Neural Network onto an FPGA



Project Demonstration

Author: Oliver Kugel

Supervisor: Dirk Koch

17th March 2017

Outline

1. What the Project is about

2. Inception-v3 Graph

3. Splitting and Running the Graph

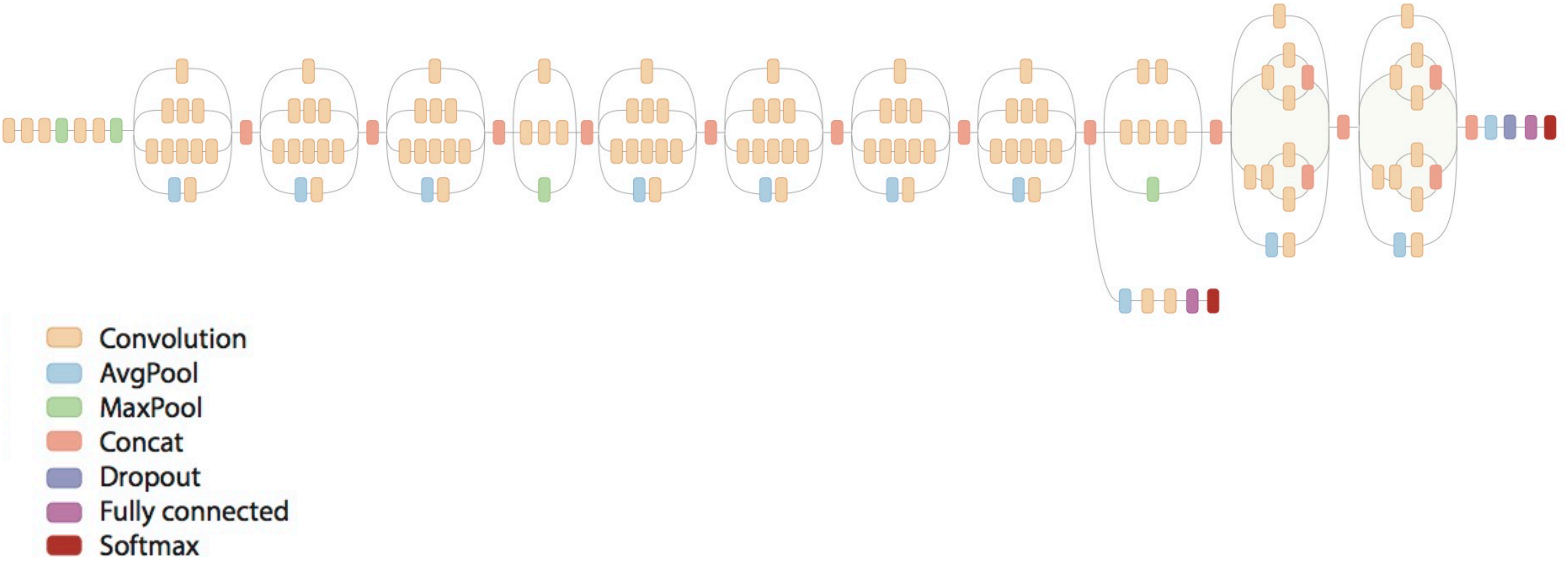
4. Convolution in Hardware

5. Questions

What is the Project about?

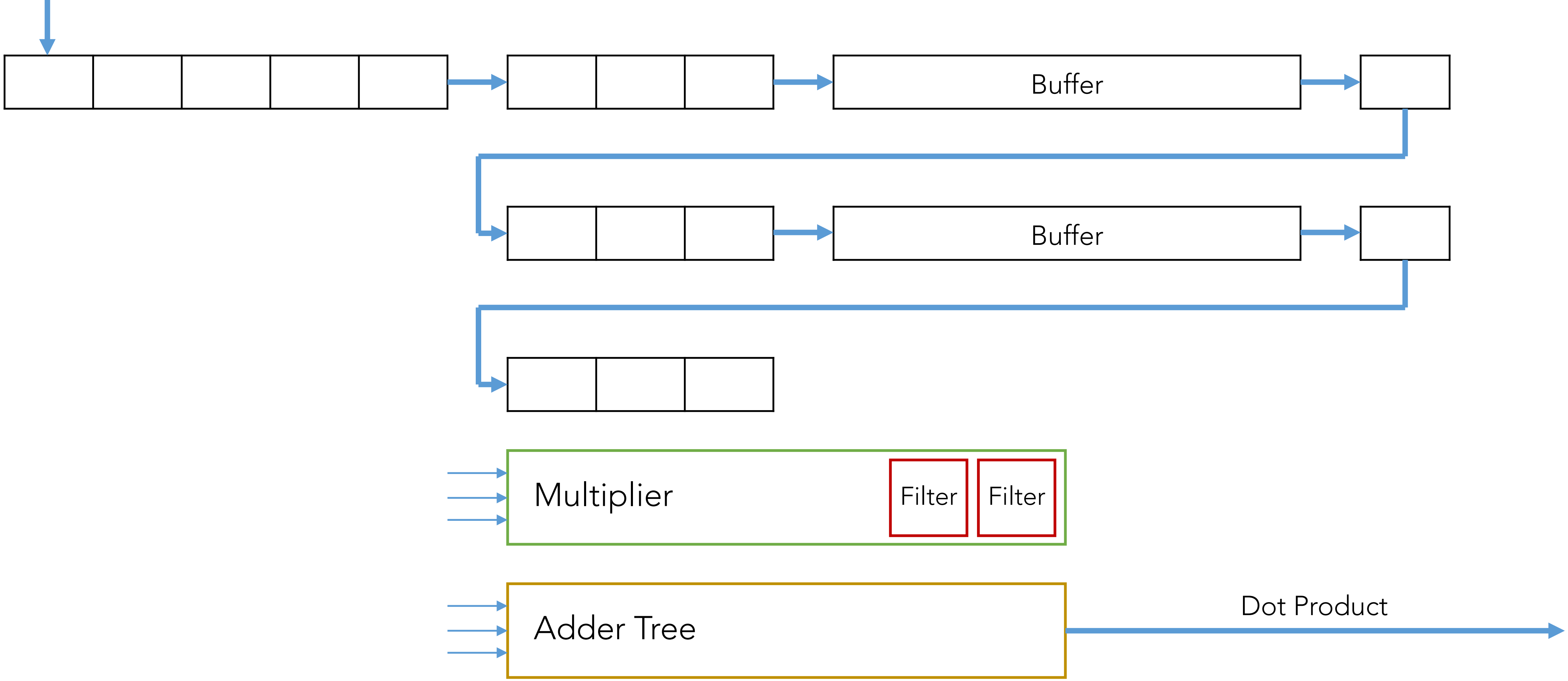
- Neural Networks seem complex but the operations performed are simple. Especially dot-product calculation.
- The weights of the Inception-v3 network sum up to 23 million, or 23MB → we can hold weights on-chip and run part of the graph.
- We can write hardware to speed up and parallelize the Convolutional Neural Network.
- Run a Neural Network on Hardware, interfacing with Software

Inception-v3



Splitting and Running the Graph

Convolution in Hardware



Buffer

Buffer

Multiplier

Filter

Filter

Adder Tree

Dot Product

Questions?