FPGA-based Accelerator for Long Short-Term Memory Recurrent Neural Networks

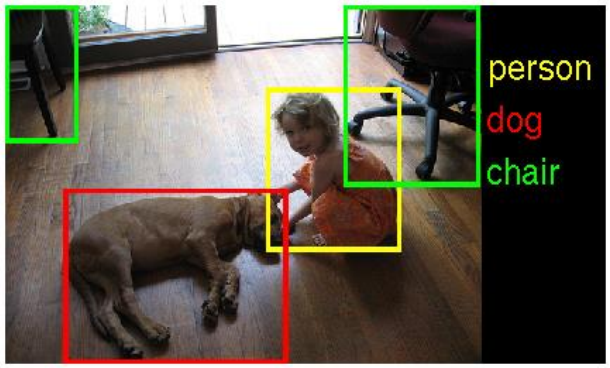
Yijin Guan[1], Zhihang Yuan[1], Guangyu Sun[1,3], Jason Cong[2,3,1]

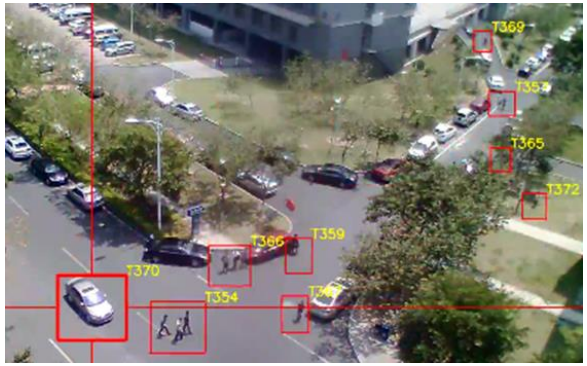
[1]Center for Energy-Efficient Computing and Applications, Peking University, China
[2]Computer Science Department, University of California, Los Angeles, USA
[3]PKU/UCLA Joint Research Institute in Science and Engineering

Deep Learning

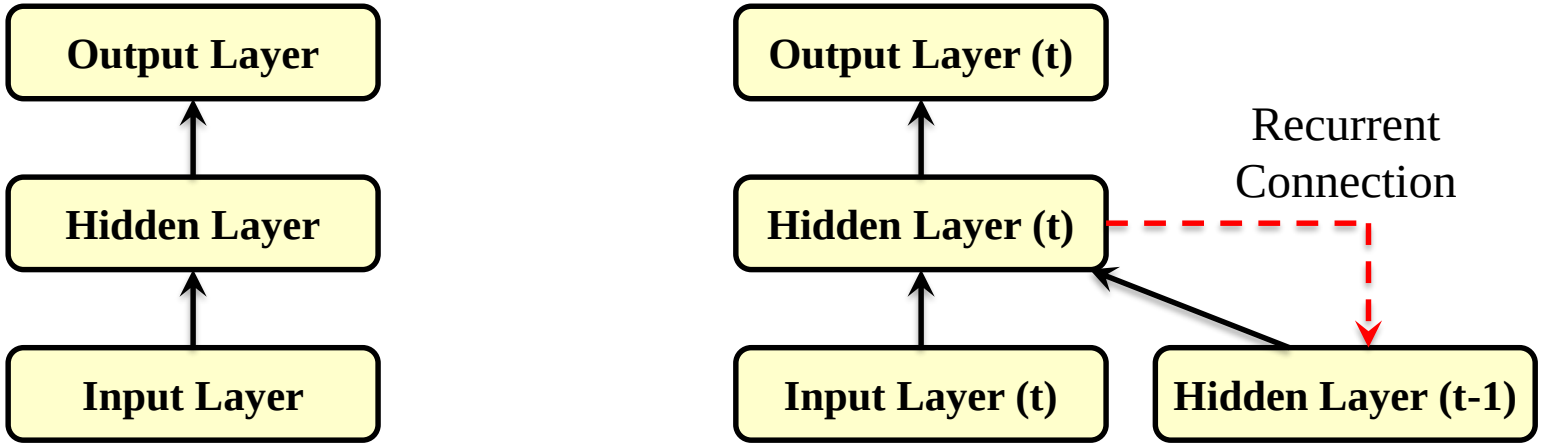
Scenarios

Applications





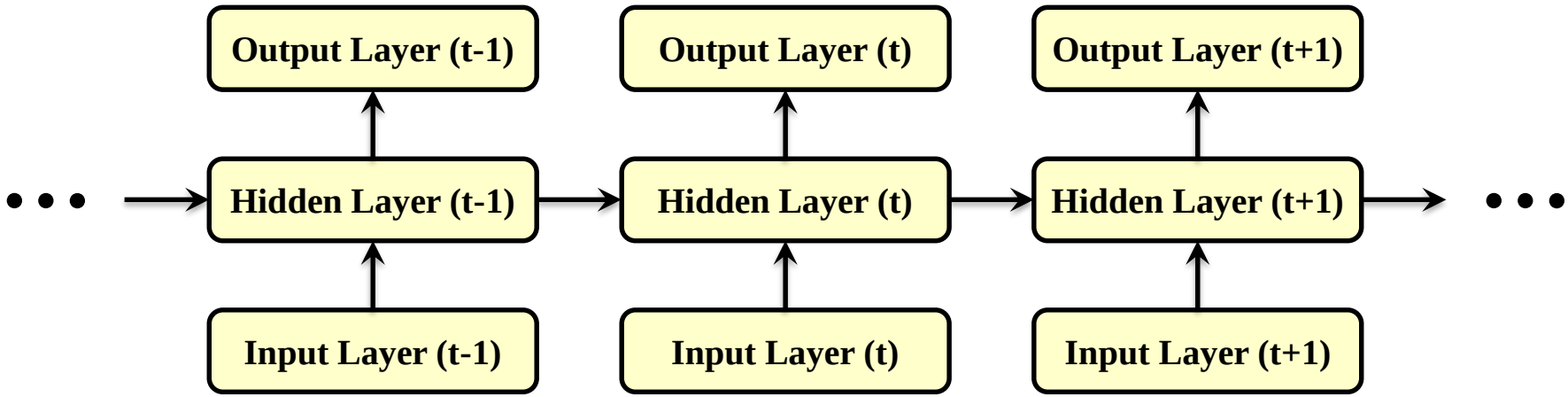
Recurrent Neural Network



Feed-forward NN

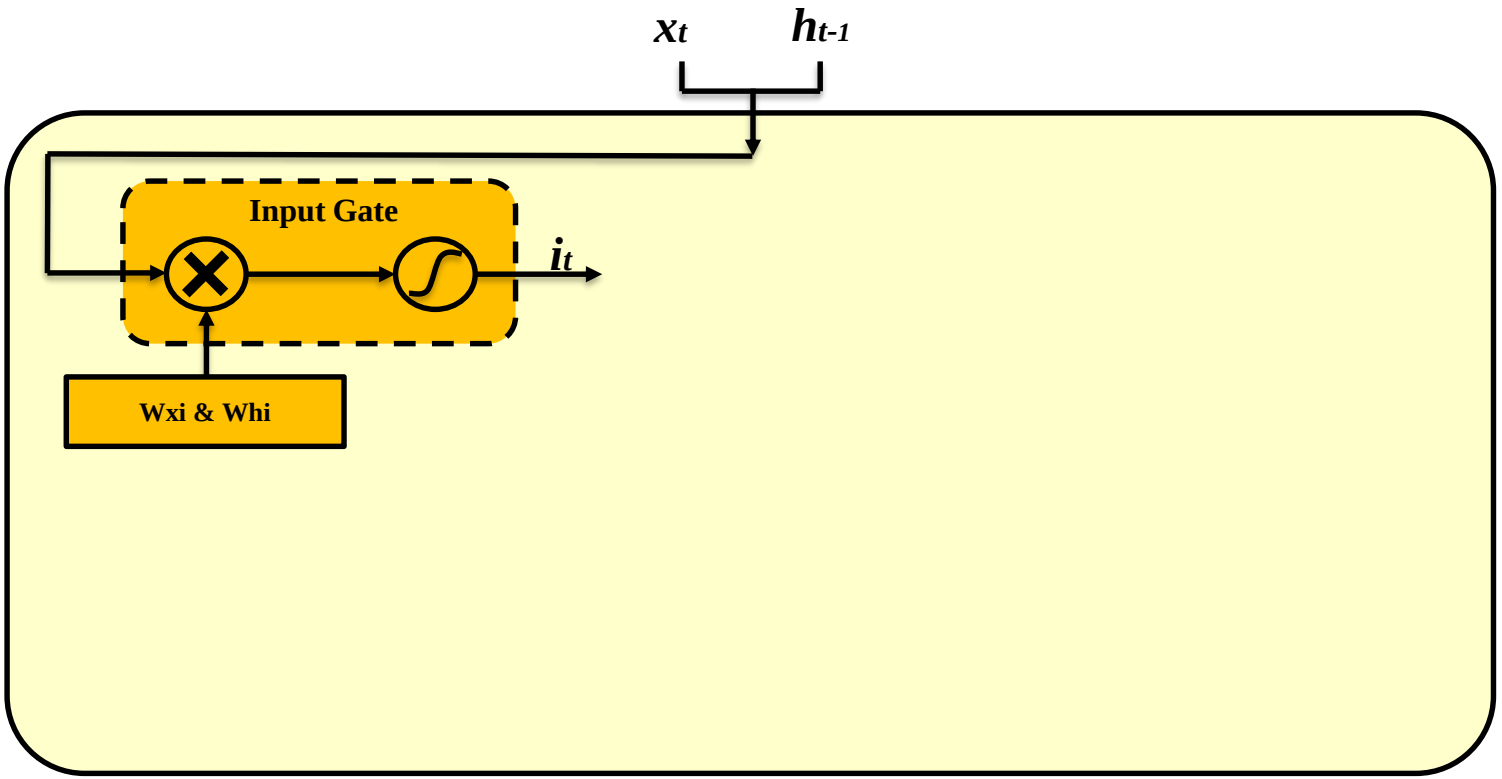
RNN

Recurrent Neural Network



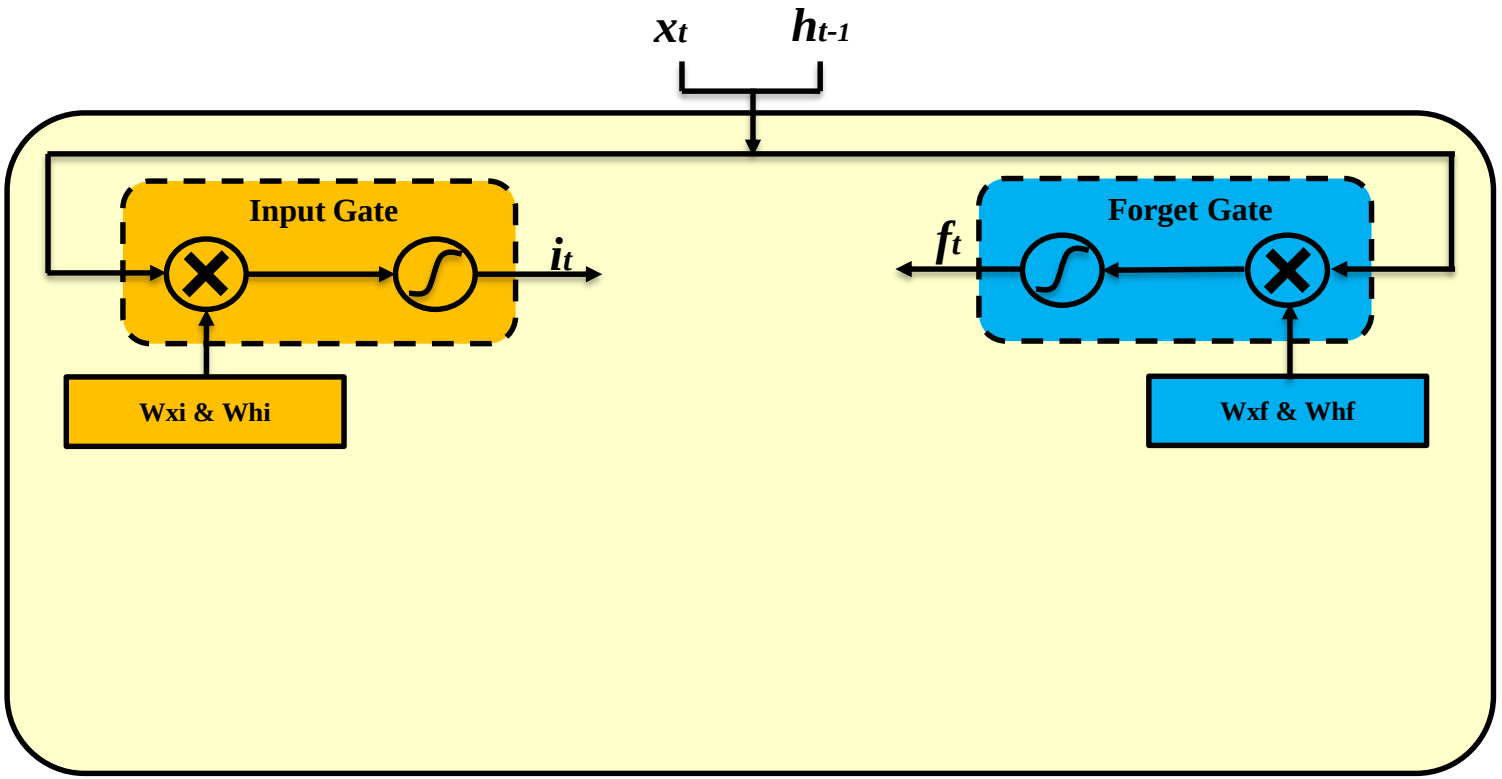
RNN unfolds into a DNN over time

Long Short-Term Memory



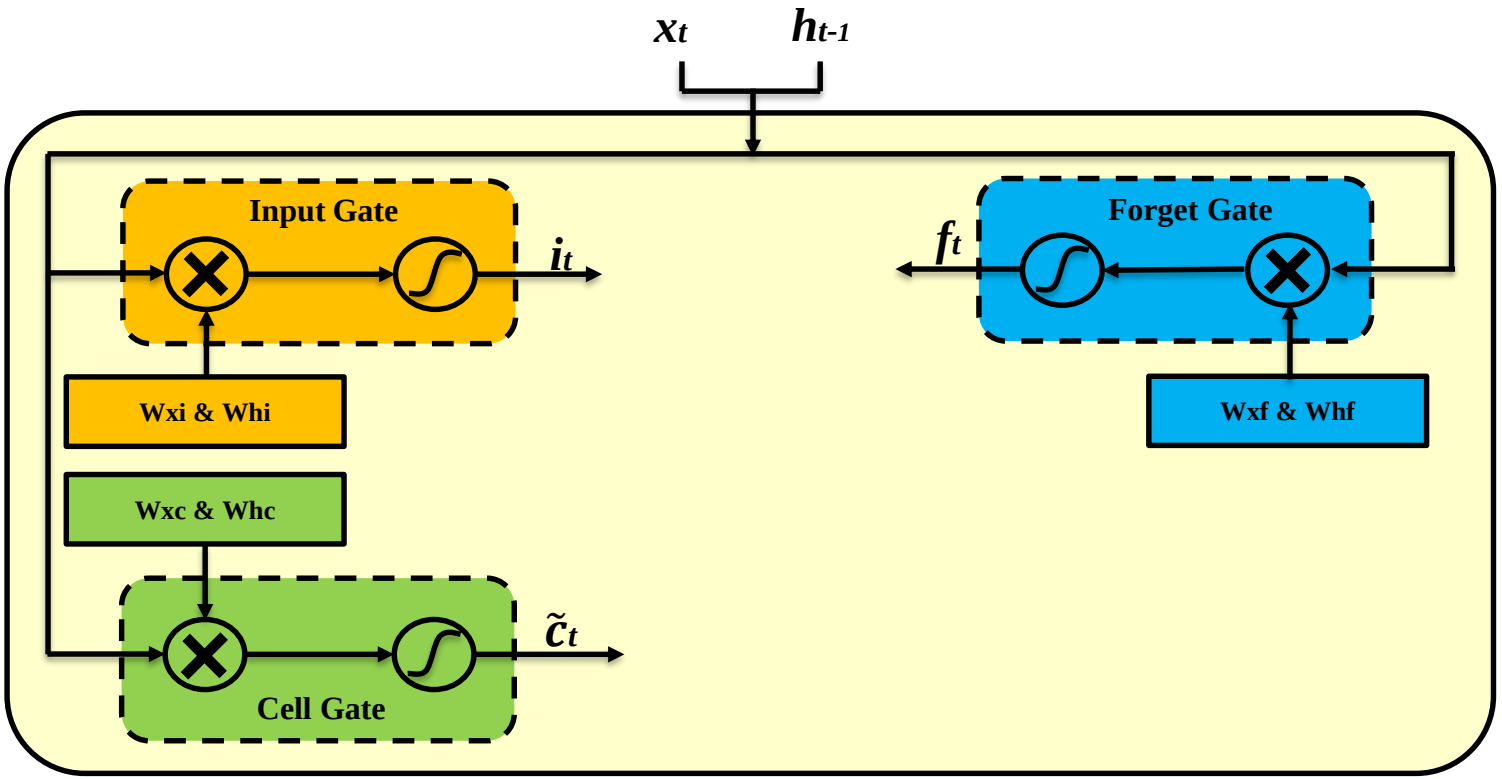
$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + b_i\right)$$

Long Short-Term Memory



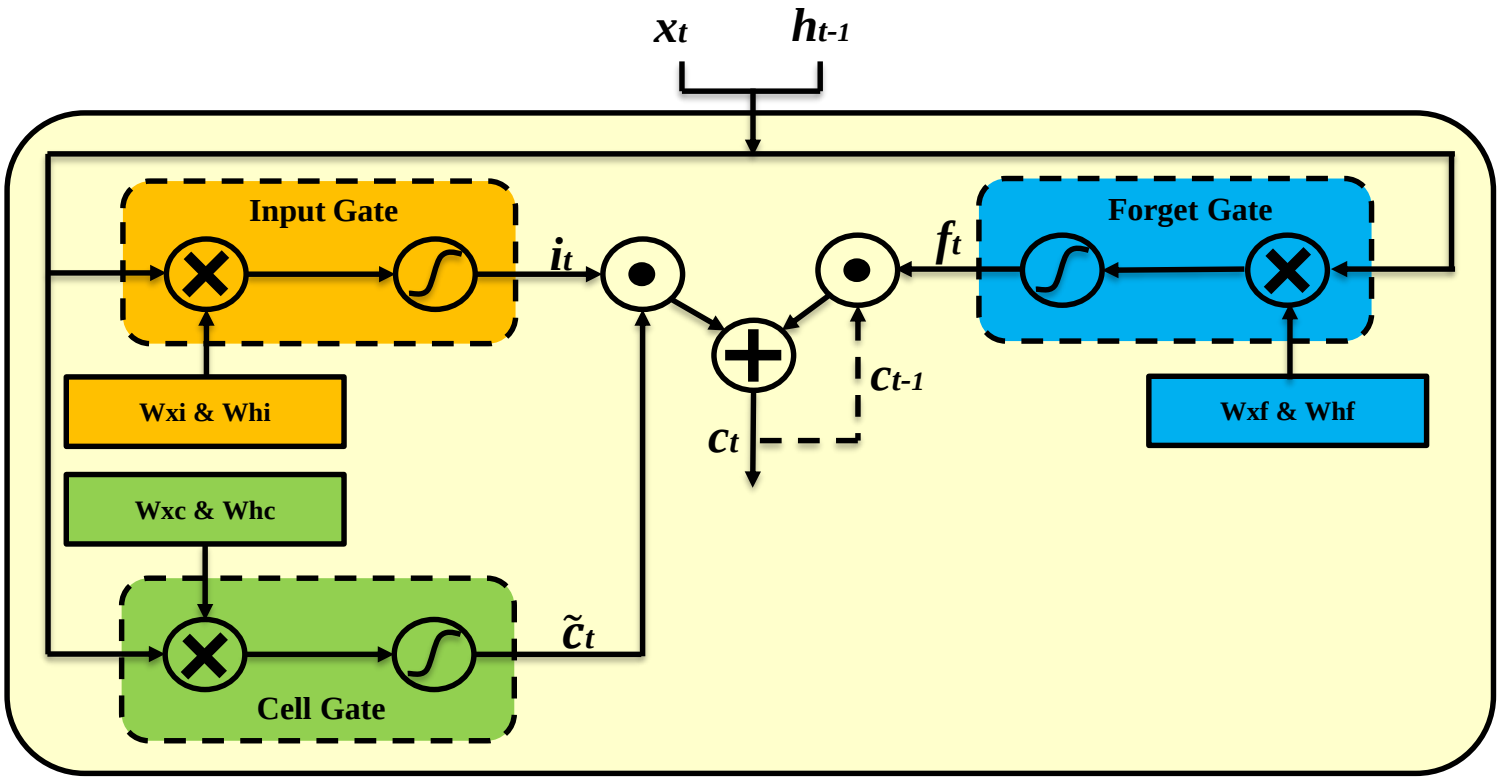
$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right)$$

Long Short-Term Memory



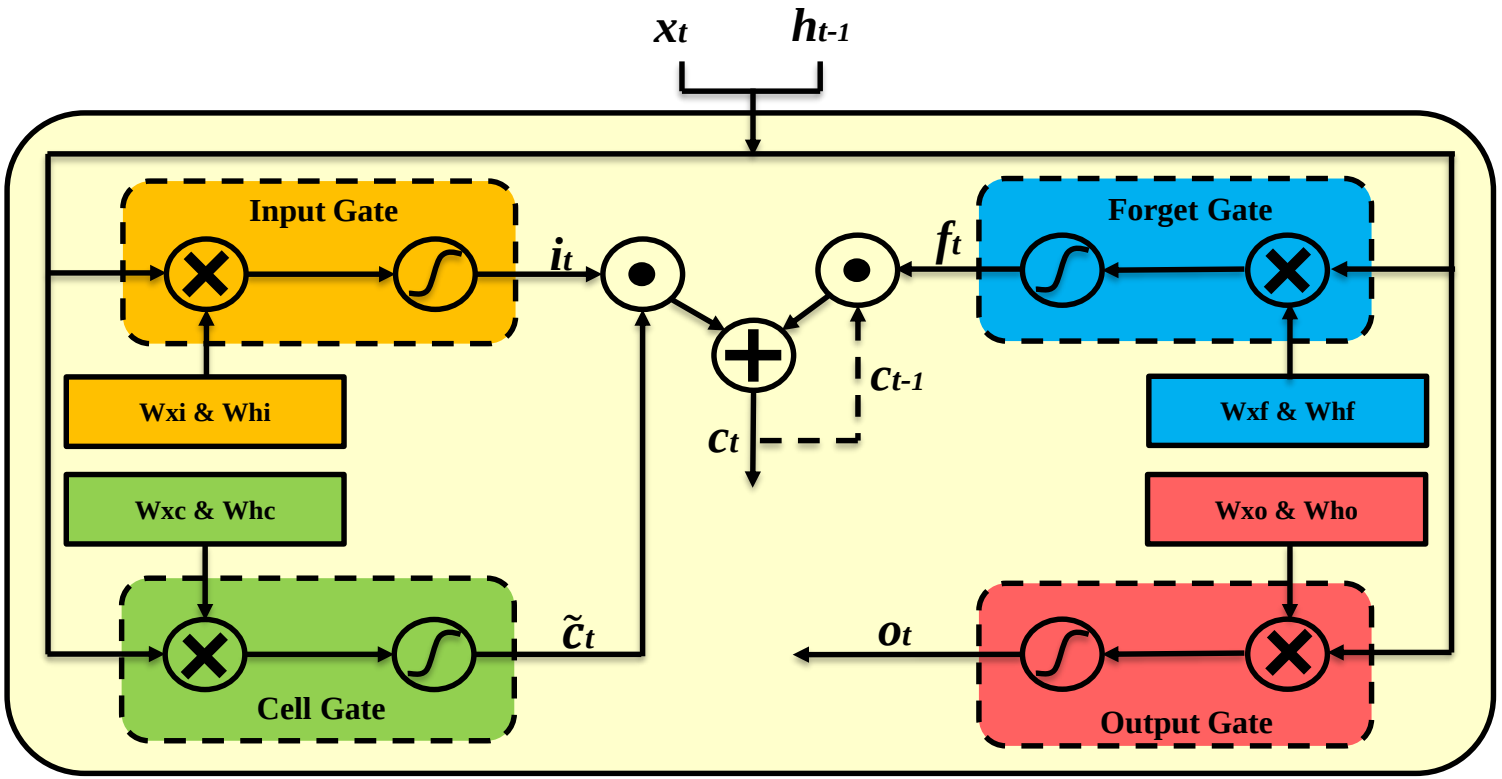
$$\tilde{c}_t = tanh\,(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$

Long Short-Term Memory



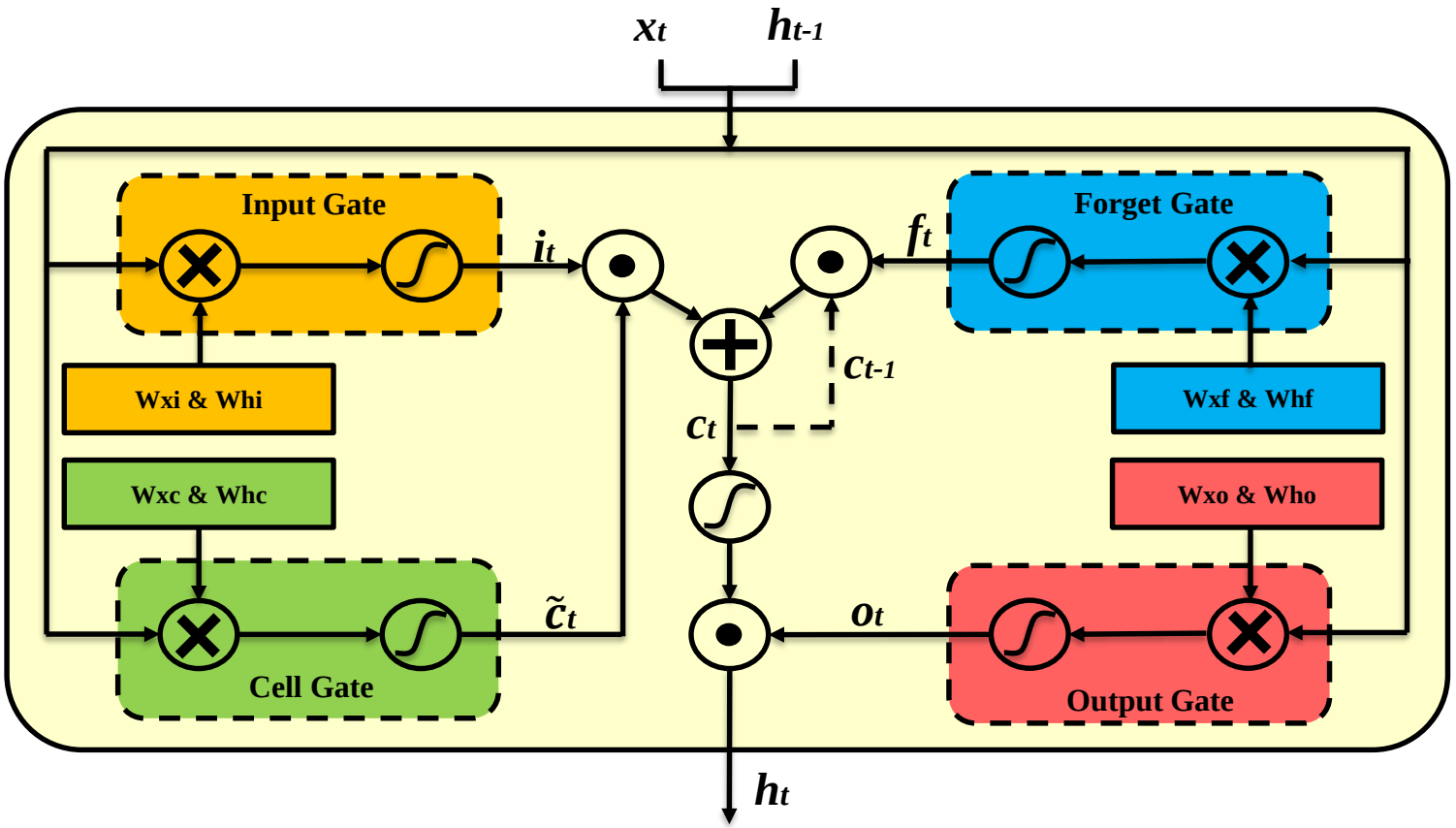
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

Long Short-Term Memory



$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + b_o\right)$$

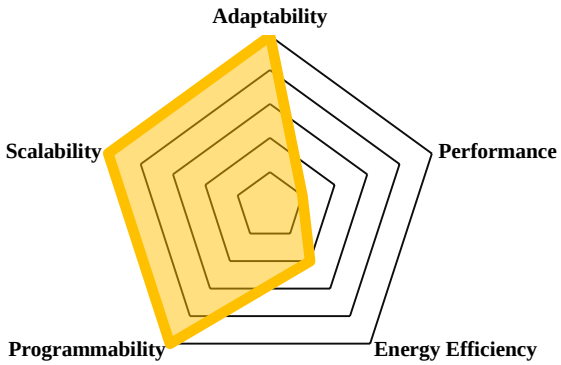
Long Short-Term Memory



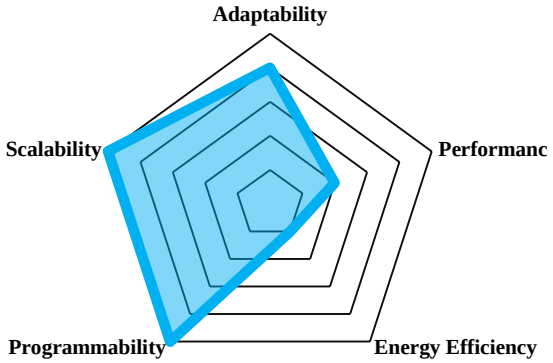
$$h_t = o_t \odot tanh(c_t)$$

Why FPGA

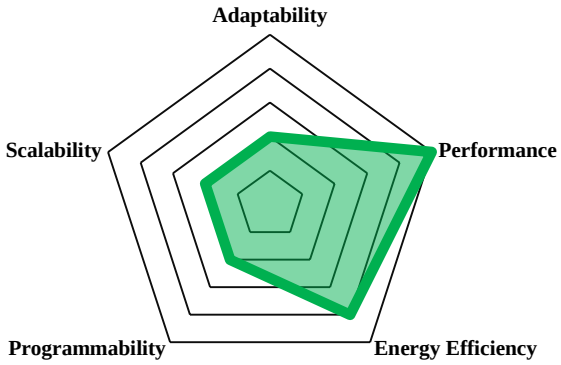
CPU



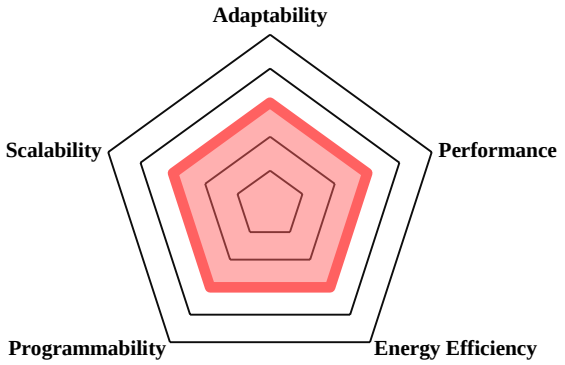
GPU



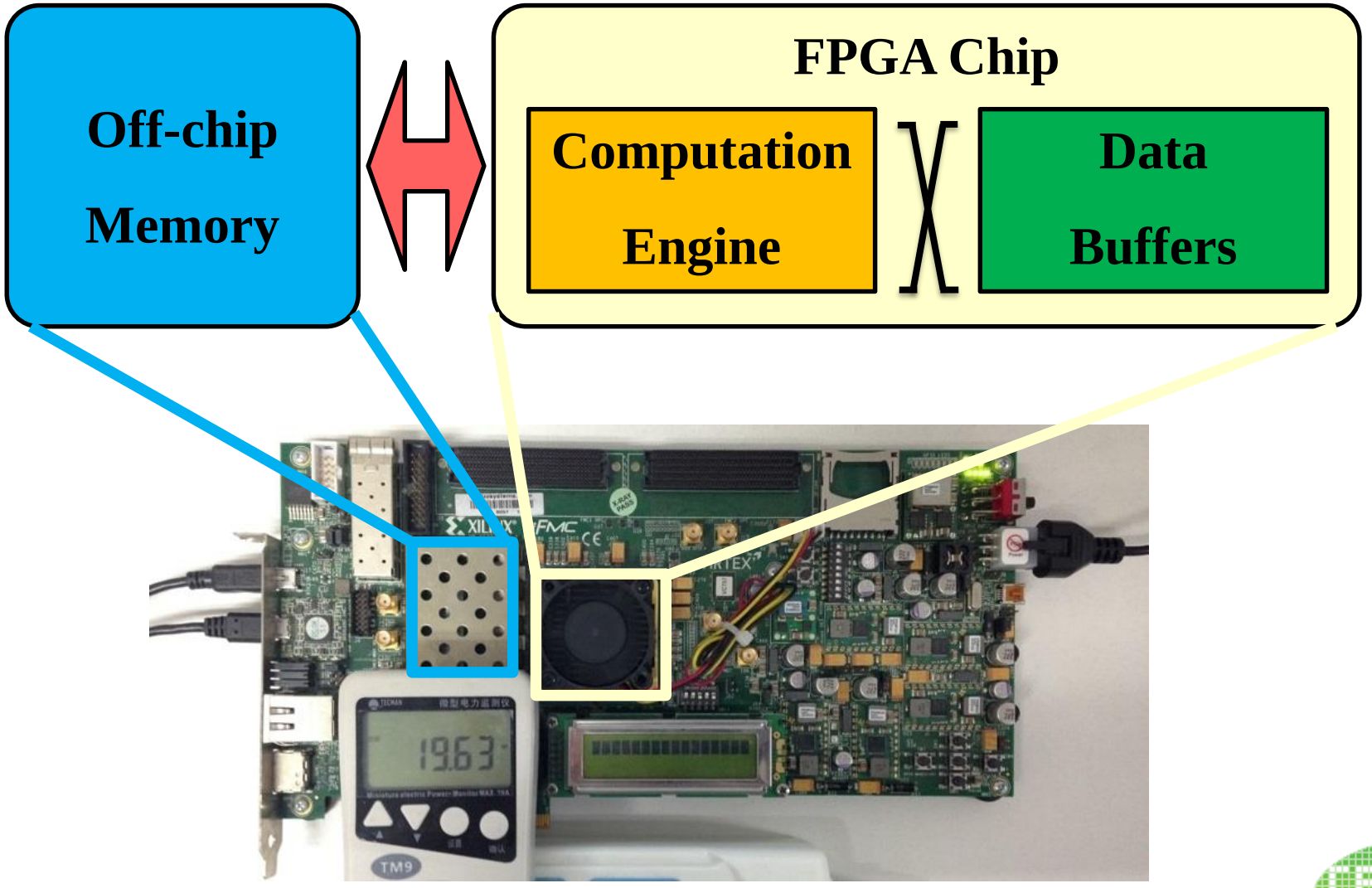
ASIC



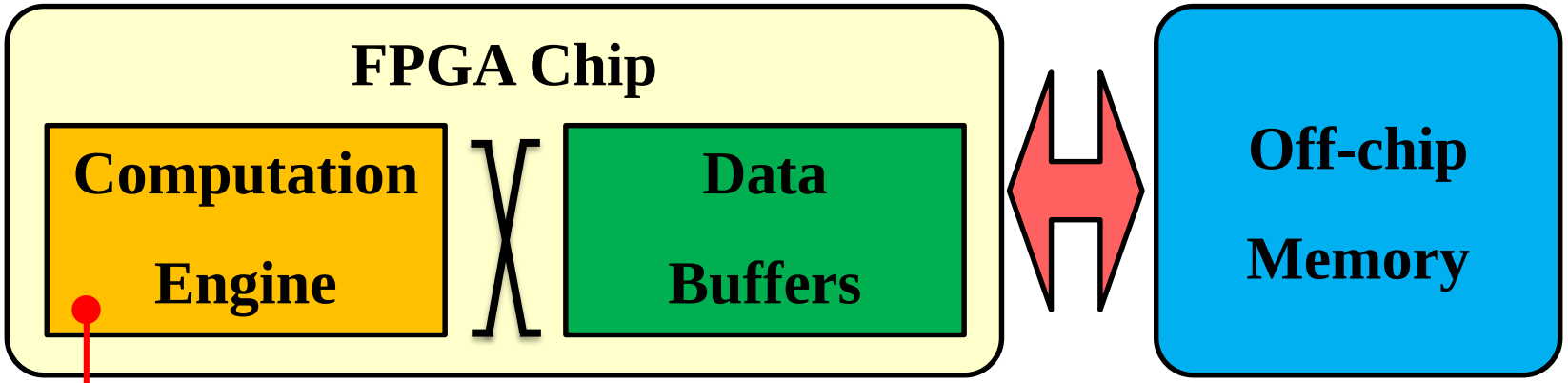
FPGA



Design Challenges and Optimizations

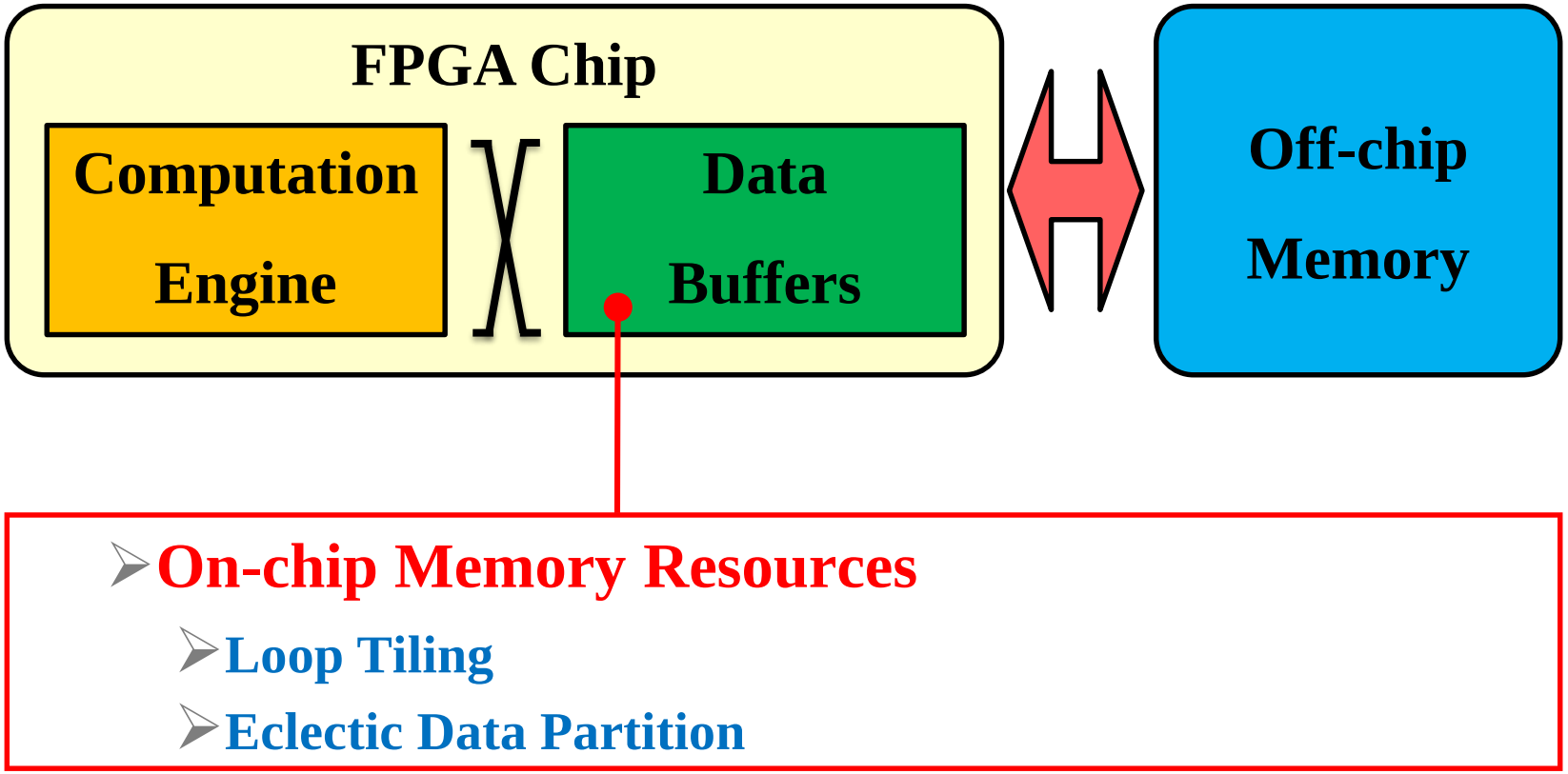


Design Challenges and Optimizations

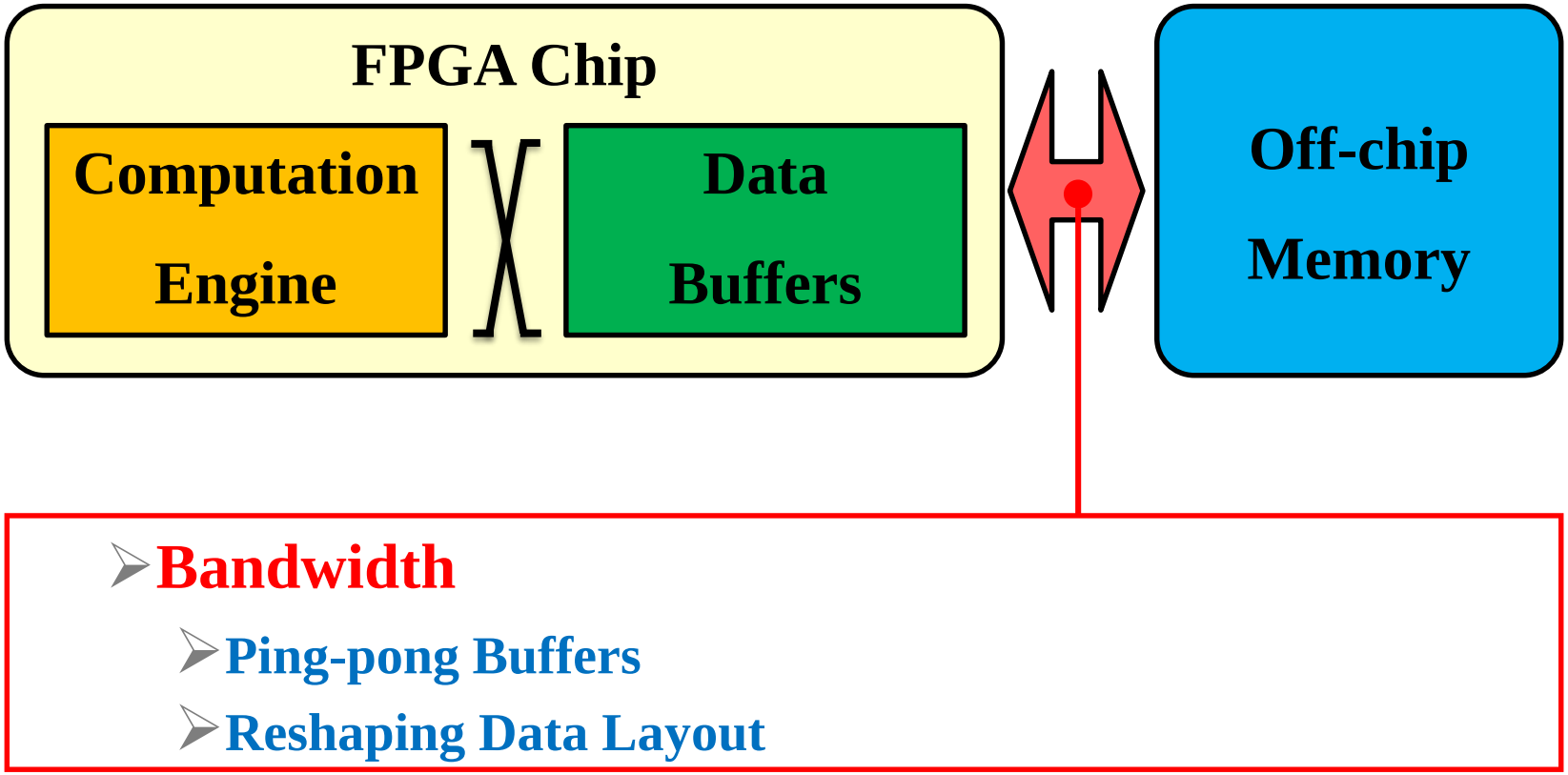


- **Computation Resources & Performance**
  - **Loop Unroll**
  - **Deep Pipeline**

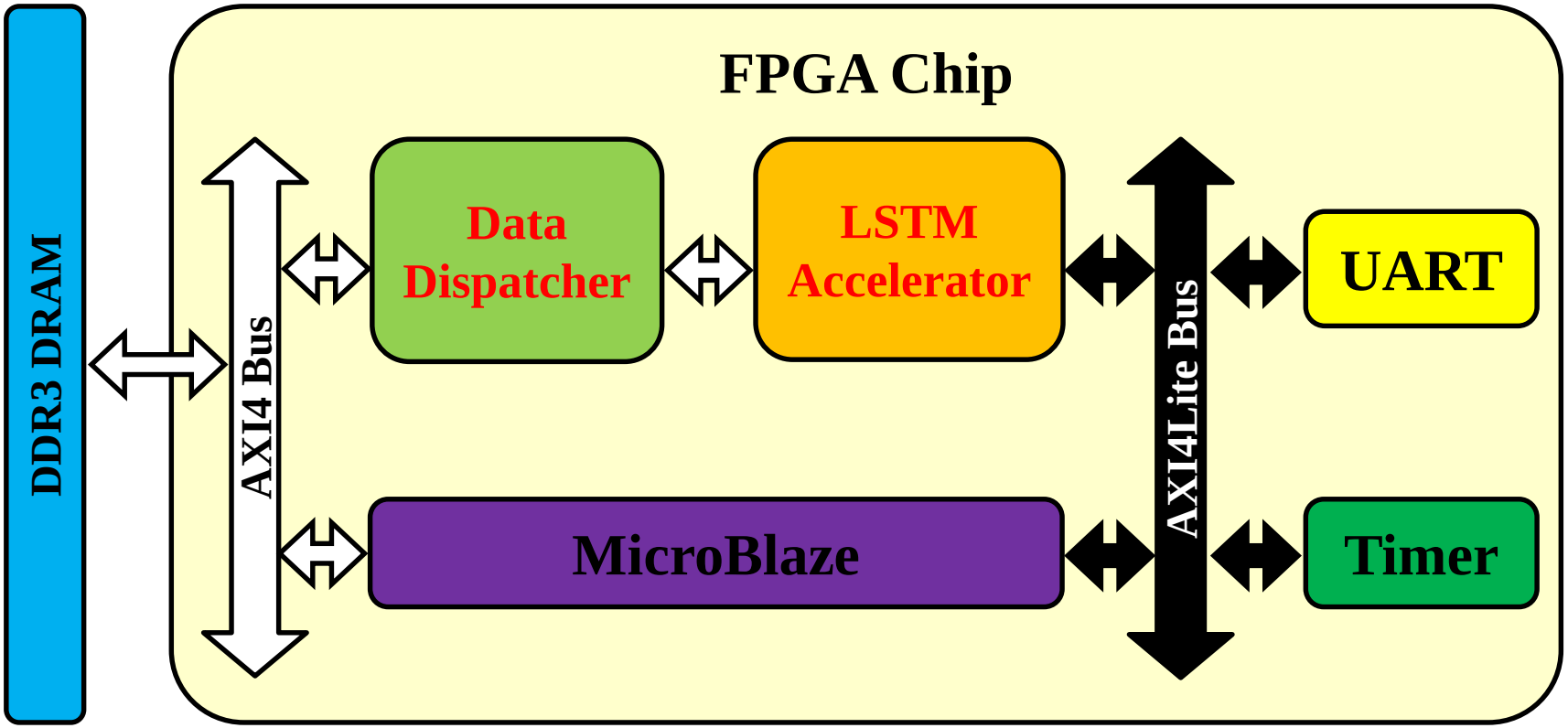
Design Challenges and Optimizations



Design Challenges and Optimizations

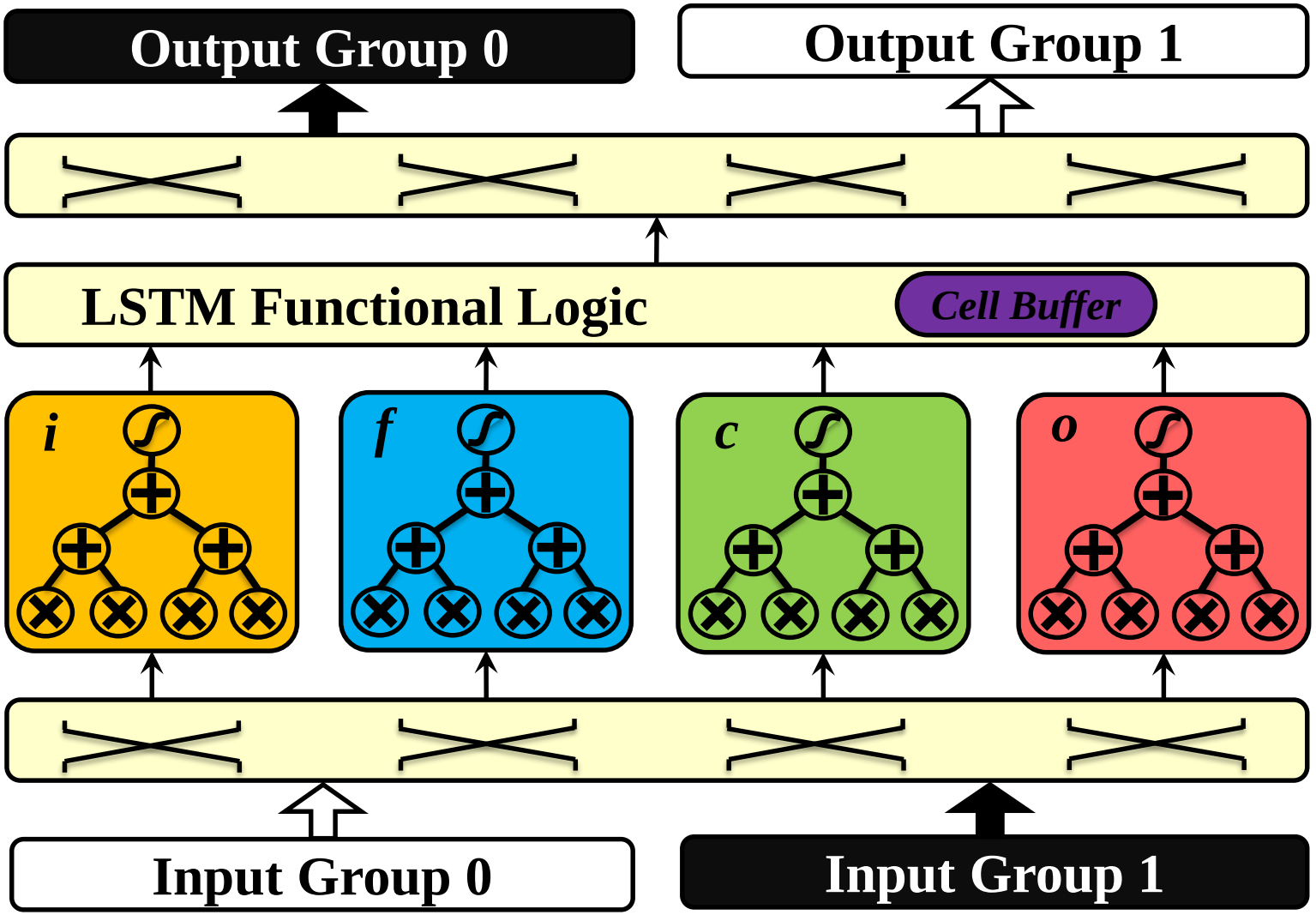


System Design

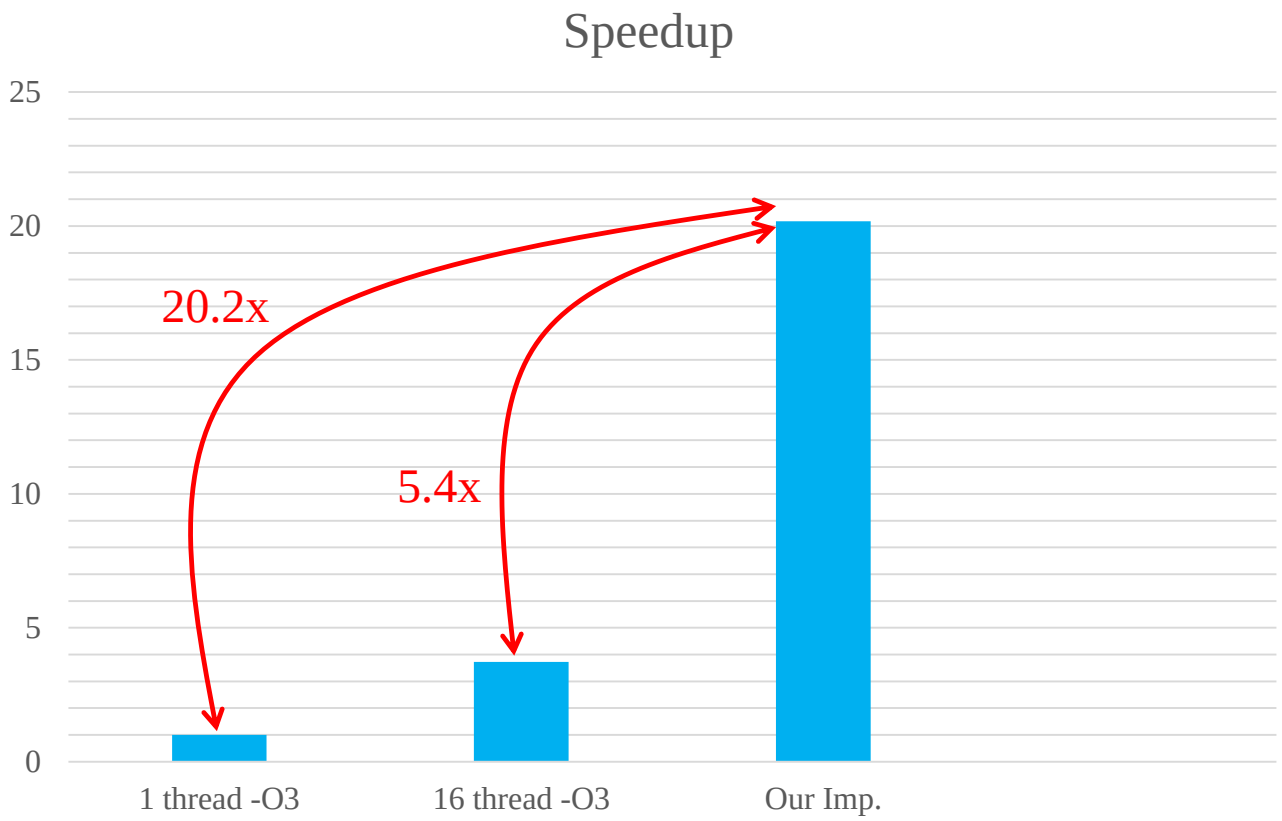


- Vivado High-level Synthesis (v2015.4)
- Vivado Design Suite (v2015.4)

Accelerator Design

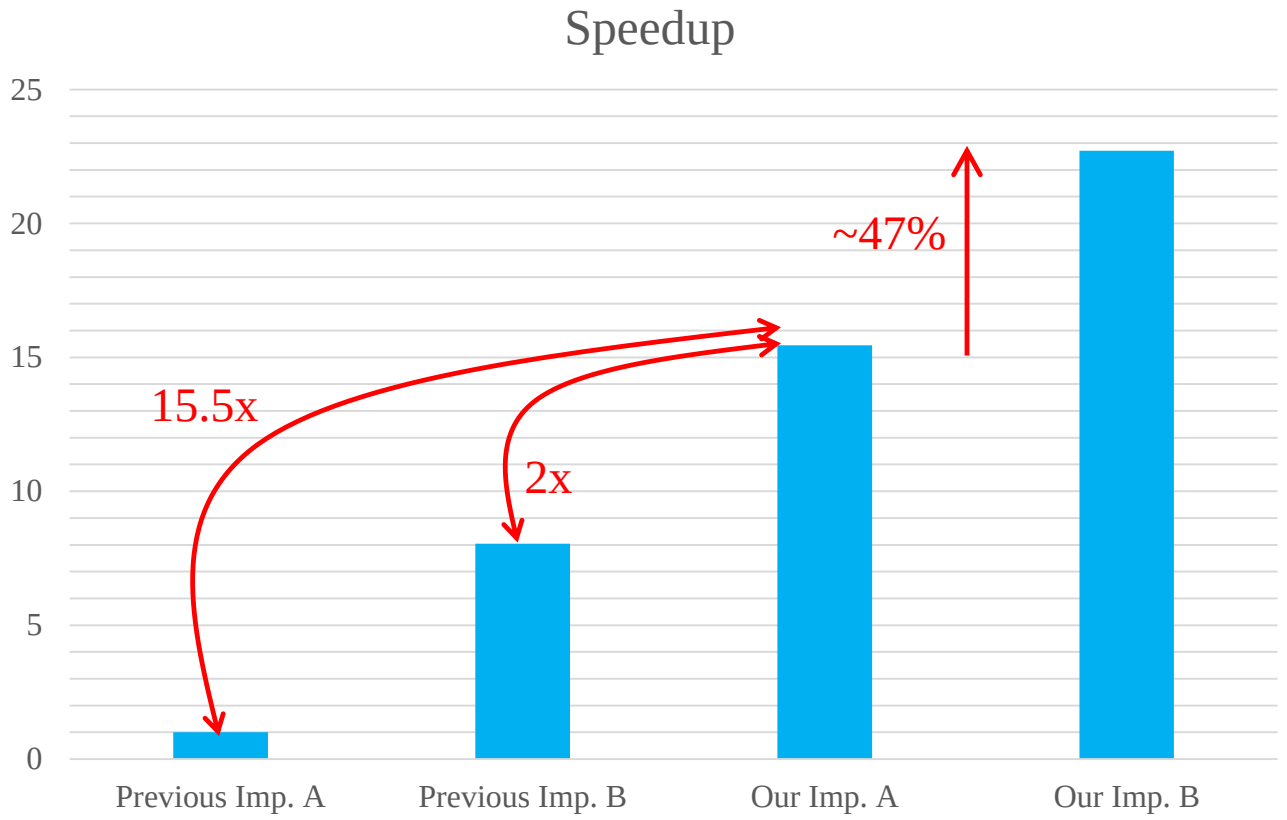


Experimental Results



Device	Model	Freq.	Development Env.
CPU	Xeon E5-2430	2.20 GHz	gcc -O3 & OpenMP
FPGA	Xilinx Virtex7-485t	150 MHz	Vivado Design Suite

Experimental Results



Imp.	Model	Freq.	Data Precision
Previous Imp.	Xilinx Zynq7020	142MHz	Fixed-16
Our Imp.	Xilinx Virtex7-485t	150 MHz	Float-32

Future Work

- **Data quantization**
  - **Low-precision fixed-point numbers**
- **Model compression**
  - **Connection Pruning**
  - **Matrix compression (e.g. SVD)**
- **General architecture**
  - **Support for all LSTM variants**

Conclusions

An accelerator for LSTM-RNN

- Optimizations for computation & communication at architecture level
- On-board implementation with high-performance computation engines & data dispatcher
- Outperforms CPU- & other FPGA- implementations

Thank You

Q & A