Power-Aware FPGA Mapping of Convolutional Neural Networks

Alexander Montgomerie-Corcoran[1], Stylianos I. Venieris[2], Christos-Savvas Bouganis[1]

[1]*Dept. of Electrical and Electronic Engineering, Imperial College London, London, UK*

[2]*Samsung AI Center, Cambridge, UK*

alexander.montgomerie-corcoran15@imperial.ac.uk, s.venieris@samsung.com, christos-savvas.bouganis@imperial.ac.uk

Main Outcomes

Enabling power-aware design space exploration by exposing power consumption within a CNN-to-FPGA mapping framework. The main results are:

- **93.4% accuracy for power consumption of LeNet across design points on the zynq ZC7020 chip.**
- **20.1% power reduction for purely throughput-driven designs of AlexNet for the zynq ZC7045 chip.**

Power Consumption Modelling

Despite the extensive existing efforts in power consumption modelling, there is still a gap between accuracy and estimation speed when targeting CNN accelerators on FPGAs. To this end, a novel power modelling methodology is proposed tailored to FPGA-based CNN systems. Our method overcomes the limitations of existing tools and combines high accuracy with fast estimation by exploiting two key observations:

1. *strong statistical patterns in the feature maps*
2. *parametrisation of commonly used CNN hardware modules.*

To use a model to guide design space exploration, the model must account for the most significant sources of power consumption. As such, the following components are identified as the most significant:

- **Dynamic Power** (of PL)
- **Static Power** (of PL)
- **Memory-Interfacing Power** (from PL to DDR)

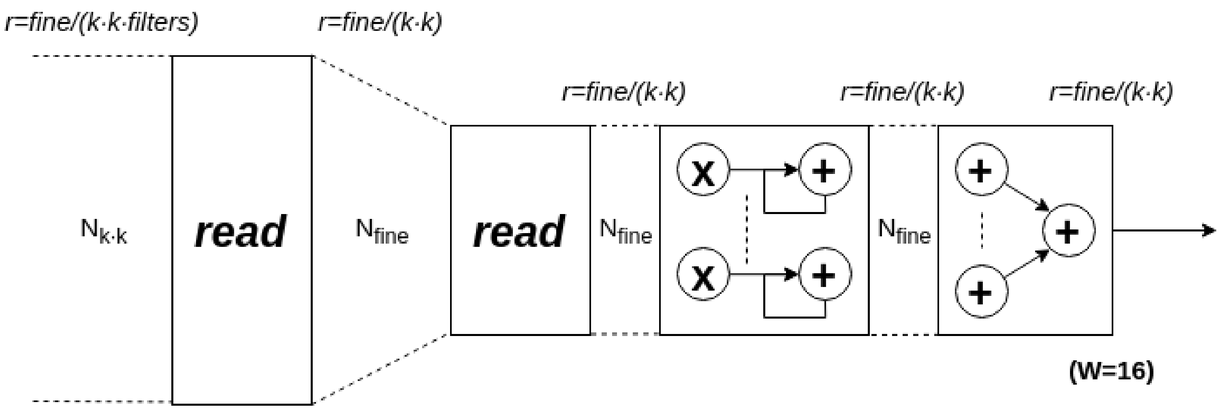

Figure 1) Diagram of a Convolution module.

The power model in this work is derived for the fpgaConvNet [1] framework, which performs DSE on modules. A linear model is derived for these modules using information about the operations they are comprised of as well as the signal activity into these modules. This is illustrated in Fig. 1 where all the parametrized operations are given as well as the related performance models for them.

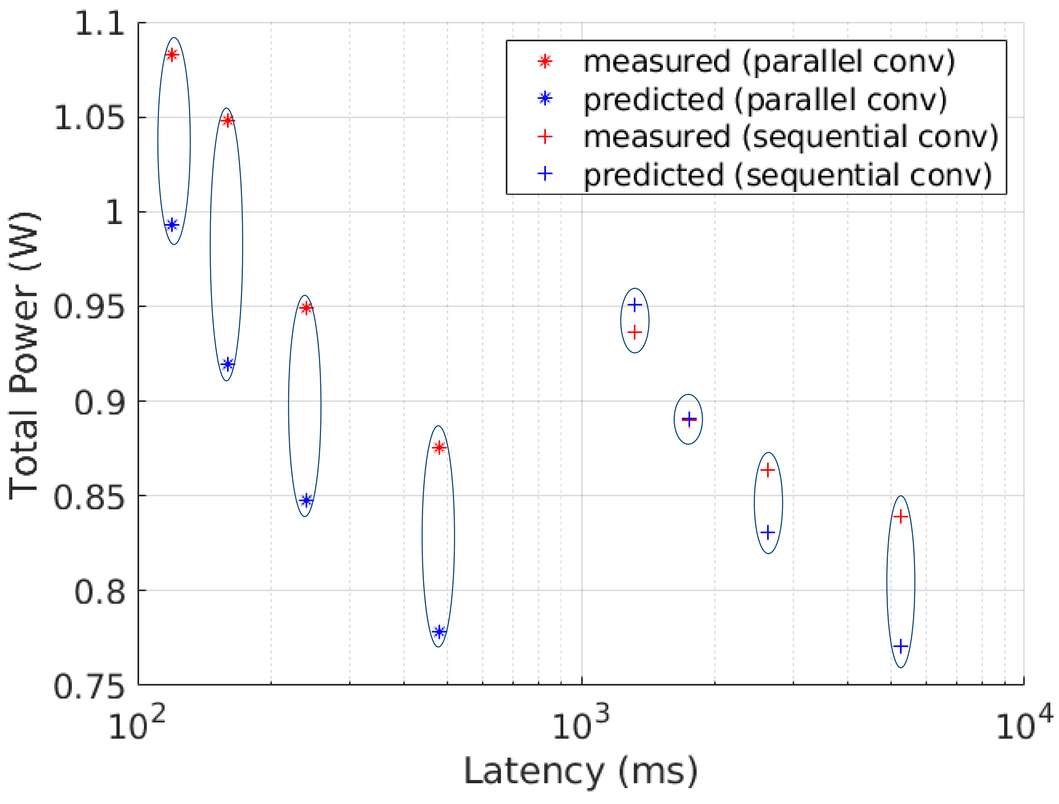

Figure 2) Measured and predicted power for the 1st layer of AlexNet.

The accuracy of the model is evaluated in Fig. 2 for the first layer of AlexNet. The different design points represent varying throughput objectives. The power model is able to stay below an error of 13% at most. This diagram also highlights the existence of power-efficient design points as well the models effectiveness in exposing them.

Another important aspect to the power model for use in DSE is the evaluation time. Having described power consumption through a small set of parameters, the model is able to be evaluated quickly. From Table 1, the power model shows competitive accuracy with very fast evaluation time.

Tool	Eval. Time	Error (%)
XPE	< 1 second	624.37
Vivado	> 30 minutes	559.86
Vivado (saif)	> 1 hour	5.01
Proposed Solution	**< 1 second**	**21.50**

Table 1: Comparison of Power Modelling Tools

Design Space Exploration

With a power-modelling methodology in place, power-driven designs can be explored. Within the embedded space, power is a crucial aspect, and characterising and limiting power consumption can play a key role to the configuration of the final design.

The power model is incorporated within a DSE technique (Simulated Annealing). Throughput-driven DSE is shown in Fig. 3 alongside power consumption. A clear pareto-optimal front can be seen between throughput and power-consumption, highlighted by the red line. By adding power consumption as a constraint, the model is able to find power-efficient designs.

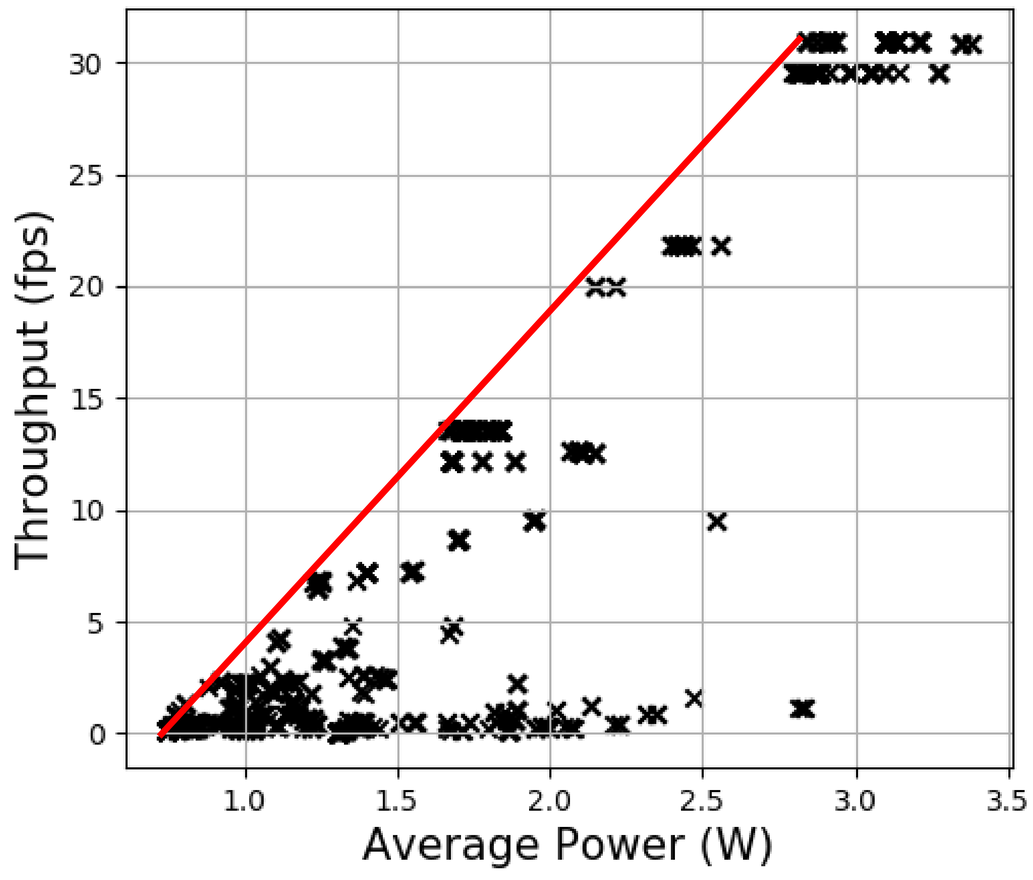

Figure 3) DSE with a throughput objective for AlexNet on ZC706.

Conclusion

This work brings power consumption to the forefront of the fgpaConvNet [1] framework, and promotes methods which can be used across other frameworks. In this way, low-power implementations of CNNs will be realisable for a host of platforms with harsh power constraints.

References

[1] *S. I. Venieris and C.-S. Bouganis, "fpgaConvNet: A Framework for Mapping Convolutional Neural Networks on FPGAs," in FCCM, 2016*