ScuDo

Scuola di Dottorato ~ Doctoral School

WHAT YOU ARE,TAKES YOU FAR

XXXIII Cycle

Performance/energy-optimized allocation for multi-kernel applications on multi-FPGA platforms

Junnan Shan

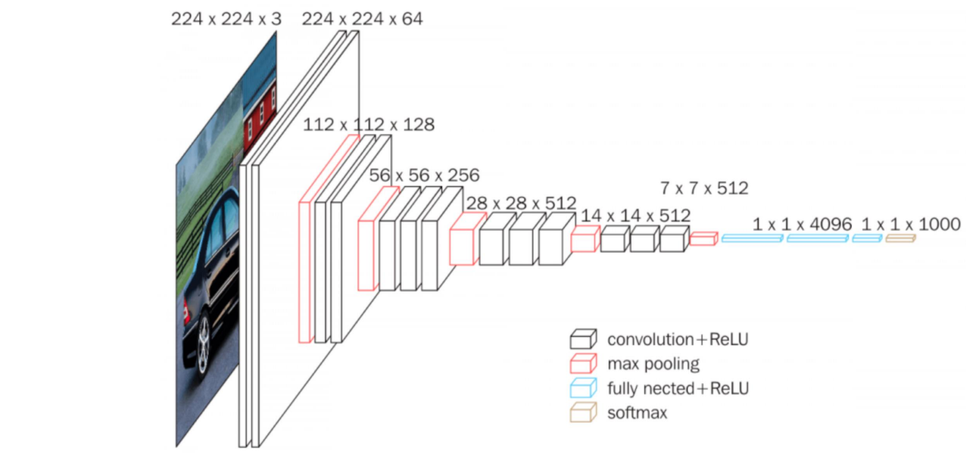
Supervisor: Prof. Mario Casu, Prof. Luciano Lavagno

Research context and motivation

- Datacenter workloads are increasing exponentially: need for high performance
- Energy is a major component of datacenter cost: need for efficient power management
- Heterogeneous nodes (CPU, GPU, FPGA) are the best way to tackle different kinds of workloads, e.g. GPUs for Convolutional Neural Network training and FPGA for CNN inferencing, due to different performance and energy characteristics
- Sigle-FPGA design is obstructed by the limited on-chip resources, we employ multiple FPGAs to increase the throughput by balancing the resource distribution
- Given a target Initiation Interval (II, i.e. inverse of throughput), our algorithm finds the best resource allocation on the minimum number of FPGAs together with dynamic frequency scaling to minimize the energy consumption

Addressed research questions/problems

- HLS tools, such as Vivado HLS from Xilinx, can only schedule fine-grained operations, such as additions and multiplications
- Kernels are often arbitrarily parallelizable, i.e. an appropriate number of them can be allocated to multiple FPGAs to maximize overall application throughput
- For example, consider a CNN like VGG net, with 22 kernels (layers) in a pipeline



- Highly computation-intensive layers require more kernel instances to be allocated to maximize throughput (inferences/sec)
- «Compilation environments» for FPGAs, such as SDAccel and SDSoc from Xilinx, do not support the designer (often more experienced in SW architecture and implementation) to allocate the appropriate computational resources at the kernel level
- Research goal: develop a Mathematical Programming scheduling model that (1) can be used to allocate kernel instances to maximize throughput (or latency) under resource and external DRAM bandwidth constraints. (2) given an II, it can choose the number of FPGAs used to allocate CUs and using the right working frequency to minimize the Energy consumption

Future work

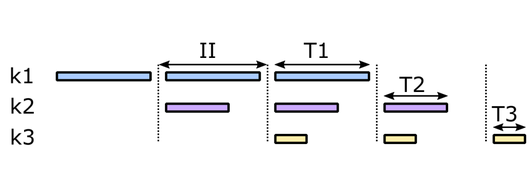
- Improve the GP model so that it can:
  - Consider various implementations of a kernel (faster/smaller)
  - Support more general task graphs (not just pipelines)
- Experiment with other more complex models with branches (e.g. GoogleNet, ResNet, etc.)

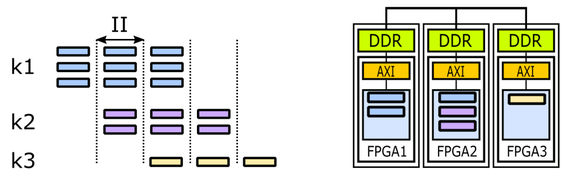
List of attended classes

- 01SGURV – Intellectual Property Rights, Technology Transfer and Hi-Tech Entrepreneurship (22/3/2018, 6)
- 01SCSIU – Machine learning for pattern recognition (28/5/2018, 4)
- 01NWNOQ – Modeling and optimization of embedded systems (9/2/2018, 6)
- 01MNFIU – Parallel and distributed computing (27/6/2018, 5)
- 01QSCIU – Reconfigurable computing (15/6/2018, 4)
- 01SHCRV – Unsupervised neural networks (9/4/2018, 6)
- External Training Activities – Innovation for Change (22/2/2018, 30h)
- External Training Activities – Sequence Models (15/4/2019, 14h)
- External Training Activities – Convolutional Neural Networks (14/4/2019, 18h)
- External Training Activities – Structuring Machine Learning Projects (25/3/2019, 4h)
- External Training Activities – Improving Deep Neural Networks: Hyperparameter tuning, Regulation and Optimization (19/3/2019, 14h)
- External Training Activities – Neural Networks and Deep Learning (25/2/2019, 16h)

Novel contributions

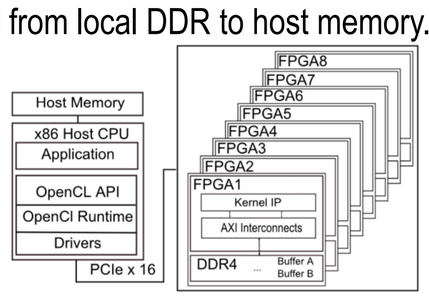
- The definition of the multi-FPGA CU allocation problem for linear kernel pipelines and its constraints
- The definition of a Non-Linear Programming model for that problem, and its solution both (1) by an exact (very expensive) MINLP solver and (2) by a GP solver, which finds an optimal non-integer solution, followed by a heuristic allocator aimed at minimizing the spreading of CUs of one kernel to multiple FPGAs
- The analysis of the quality of results for two large CNN applications, implemented on large multi-FPGA AWS F1 instances.



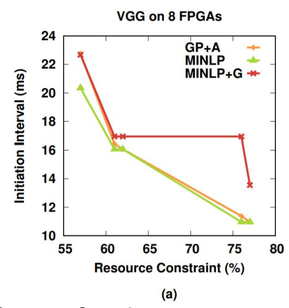


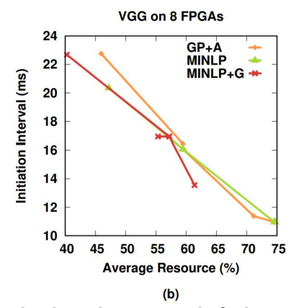
Adopted methodologies

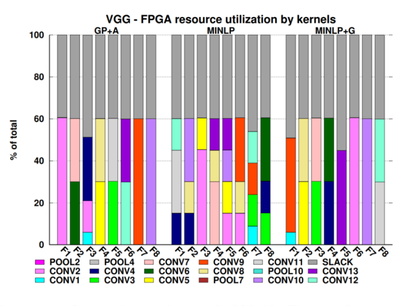
- The execution model consists of 1) Host-to-FPGA phase: data transferred from the host memory to the local DDR of the FPGA boards. We model the transfer time and energy. 2) Execution phase: all CUs of each kernel read data from local DDR, perform computing, then write data back to local DDR. We model the execution time and energy. 3) FPGA-to-Host phase: data transferred from local DDR to host memory. We model transfer time and energy



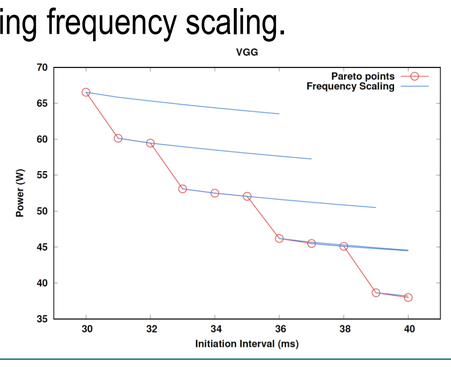
- The performance and energy model involves integer, binary and real variables, and it can be solved using a Mixed-Integer Non-Linear Programming solver. It takes around 1 hour to find a solution for a smaller problem like AlexNet. It will take 40 hours to find a solution for a large problem like VGG net, which is still not optimum
- To deal with this problem, we provide a heuristic method (using Geometric Programming and our allocator) to calculate the number of Compute Unit (CU) of each kernel and allocate them on FPGAs. Our heuristic method can generate the same or comparable solution with a much smaller execution time (few seconds)

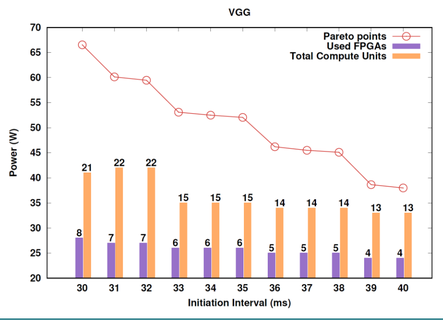






- Same for the energy optimization model, it can be solved using MINLP solver, the result shows that our model returns a much more efficient way of saving power compared to applying frequency scaling.





Submitted and published works

- Junnan Shan, Mario Casu, Jordi Cortadella, Lucianno Lavagno, Mihai Lazarescu, “Exact and Heuristic Allocation of Multi-kernel Applications to Multi-FPGA Platforms”, the 56th Annual Design Automation Conference, Las Vegas, NV, USA, 2019, pp. 3:1—3:6
- Junnan Shan, Mario Casu, Jordi Cortadella, Luciano Lavagno, “Energy-Optimized Allocation of Multi-Kernel Applications to Multi-FPGA Platforms”, Design, Automation and Test in Europe Conference, submitted.
- Osama Bin Tariq, Junnan Shan, Georgios Floros, Luciano Lavagno, Mihai Teodor Lazarescu, Christos Sotiriou, Mario Casu, Muhammand Tahir Rafiq, “Physical Aware High-Level Synthesis”, Design, Automation and Test in Europe Conference, submitted.

POLITECNICO DI TORINO

PhD program in

Electrical, Electronics and Communications Engineering