Scotch: Generating FPGA-Accelerators for Sketching at Line Rate
Martin Kiefer, Ilias Poulakis, Sebastian Breß, Volker Markl

Background & Motivation

Sketching
- Constructing stream summaries
  - e.g., Count-Min, AGMS, HLL, ...
- Various applications (e.g., AQP, ML, Network)

FPGAs
- Custom hardware defined by software
- Reprogrammable
- Circuit-level parallelism
- Data / Task / Pipeline

Sketching on FPGAs
- High guaranteed throughput
- Lower power consumption
- Boards for various interconnects and use cases

Evaluation

Throughput experiments
- S10: Intel Stratix 10 GX 2800 FPGA
  - Accelerator generated by Scotch
  - Data parallelism degree d =4/16
- GeForce: Nvidia GeForce RTX 2080 GPU
  - CUDA, Hand-tuned
- Xeon: Intel Xeon Silver 4214, 2 Sockets
- OpenMP, Hand-tuned
- Uniform data distribution

More experiments in the paper
- Resource consumption, fmax
- Power consumption
- Xilinx devices
- Comparison to hand-tuned implementation

Generated accelerators outperform parallel GPU and CPU baselines
- Data parallelism is crucial

Trade-off between sketch size and degree of data parallelism / throughput

Challenges & Approach

An FPGA expert is required!

Device-, vendor-, interconnect-specific implementations
- Reusing existing artifacts is hard

Register-Transfer Level (RTL) programming
- Concepts unfamiliar to software developers

Manual tuning
- Sketch size vs. resources & timing
- Trial and error

Automated Tuning
- Optimization in a feedback loop

Scotch: System Overview

(1) Auto-Tune proposes a sketch size to the RTL generator
(2) RTL generator creates sketching unit based on parameters and specification
(3) Vendor toolchain compiles the full accelerator (sketching unit + I/O controller)
(1) Auto-Tune algorithm analyzes the compilation report and recalibrates
(… ) repeat until optimal solution is found

Device Descriptor
IJSCN

Sketch Specification
(SketchDSL)

I/O Controller
(Design Project)

Optimal Bitstream

Evaluation

Column sketch
Matrix sketch
Row sketch

Scotch DSL Example

S10, d=4
GeForce, d=4
HLL, d=4

S10, d=1
GeForce, d=1
HLL, d=1

S10, d=16
GeForce, d=16
HLL, d=16

Scotch: Generating Accelerators for Sketching at Line Rate

Funding
This work has received funding by the German Ministry for Education and Research as BIFOLD – Berlin Institute for the Foundations of Learning and Data (01IS18025A, 01IS18037A) and Software Campus (01IS17052L).

Author Affiliations

Contact:
martin.kiefer@tu-berlin.de

Source & Data on GitHub:
martinkiefer/scotch