Learning Accurate Architectural Simulations

Reverse Engineering the NVIDIA Volta V100[1] with Black-Box Optimisation

Mark Sheppard (ms14979@bristol.ac.uk) Supervisor: Simon McIntosh-Smith

Motivation

Architectural simulations of CPUs, GPUs and accelerators are useful for diagnosing performance issues in HPC and conducting research into heterogeneous systems. However, making these accurate is tricky when the internal details are hidden!

Goals

- Demonstrate that black-box optimisation is an effective technique for enhancing the accuracy of architectural simulations
- Deliver simulator for a subset of the NVIDIA Volta V100 with state-of-the-art accuracy

1. Build Parameterised Simulation

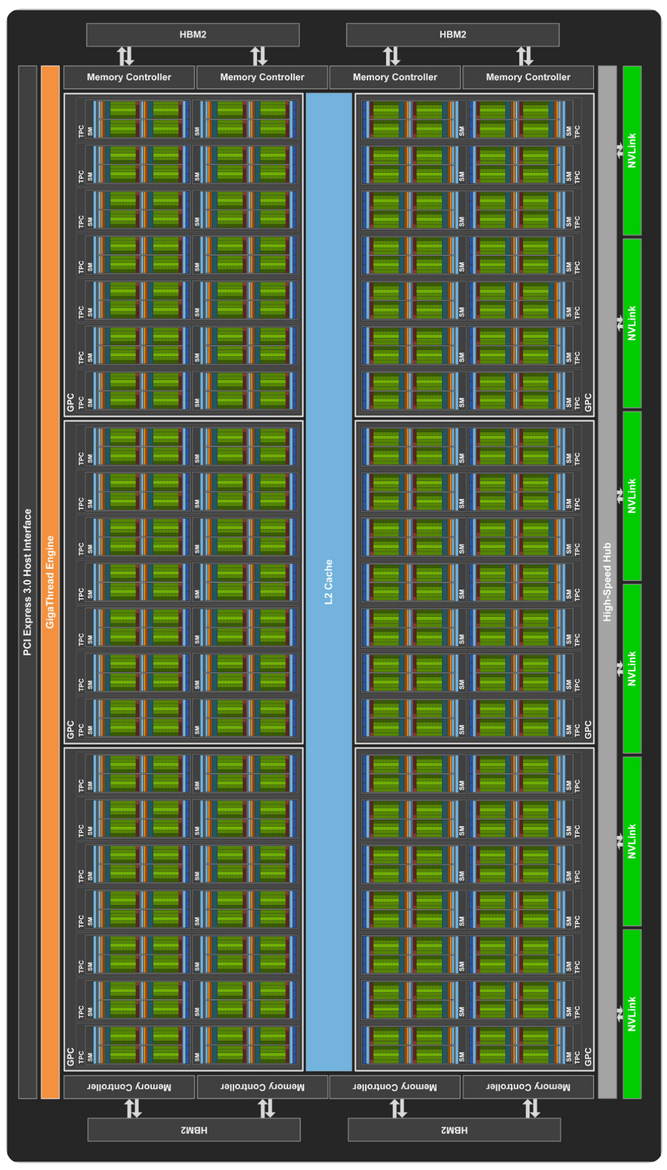
- Build approximate simulation of the NVIDIA Volta V100 with available information
- Parameterise and estimate unknowns (eg. L1 cache latency, FP32 pipeline depth)

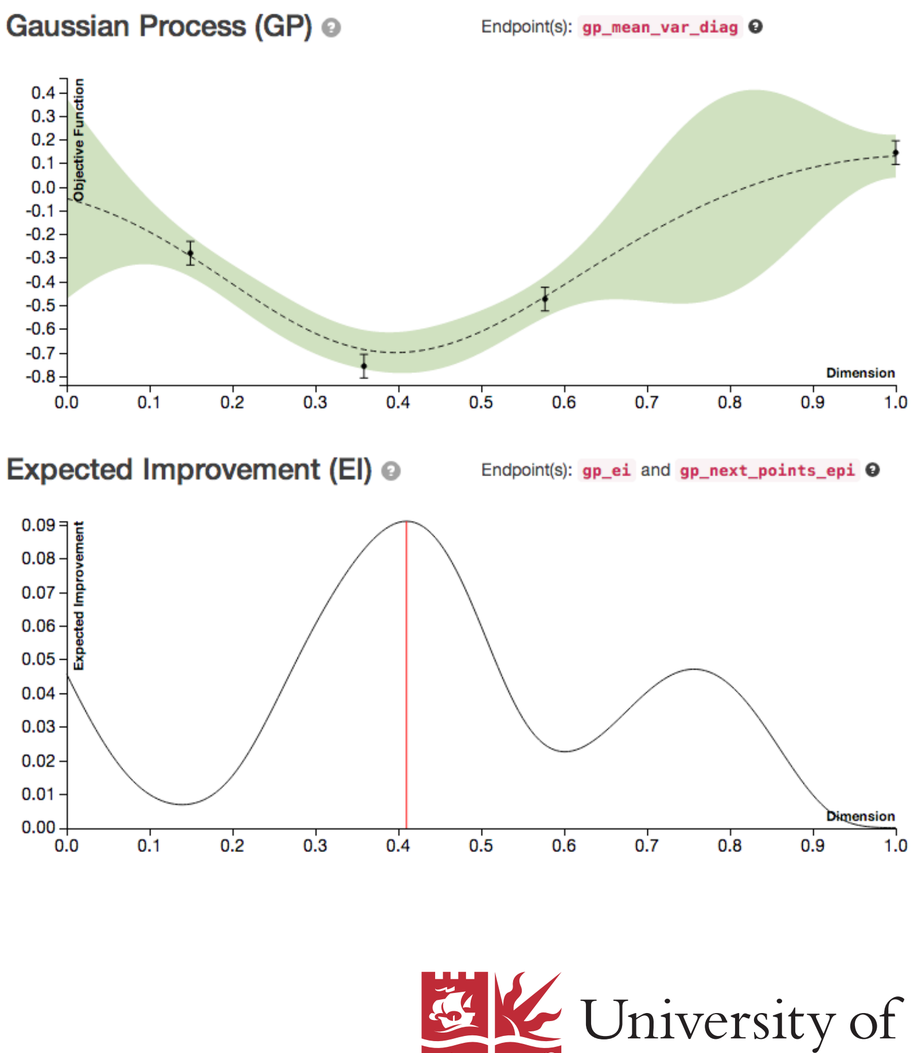
2. Generate Benchmark Kernels

- Parameter learning stage requires large training set that thoroughly exercises the target device in order to be accurate - common benchmarks are not enough!
- Randomly generate syntactically correct and semantically interesting kernels

3. Learn Simulation Parameters

- Evaluate simulation parameters by comparing kernel execution time to real GPU, with objective function measuring inaccuracy - this is highly time consuming
- Use Yelp's Metric Optimisation Engine[2] (MOE) employing black-box optimisation to find global minimum with small number of evaluations:
  1. Build Gaussian Process modelling previously evaluated points
  2. Evaluate at point(s) of greatest expected improvement in objective and repeat





[1] NVIDIA Tesla V100 GPU Architecture. NVIDIA Corporation. (2017) [2] MOE (Metric Optimisation Engine). Yelp Inc. (2014)

University of BRISTOL