KOCL: Kernel-level Power Estimation for Arbitrary FPGA-SoC-accelerated OpenCL Applications

James J. Davis  
Imperial College London  
james.davis@imperial.ac.uk

Joshua M. Levine  
Imperial College London  
josh.levine05@imperial.ac.uk

Edward A. Stott  
Imperial College London  
ed.stott@imperial.ac.uk

Eddie Hung  
Imperial College London  
e.hung@imperial.ac.uk

Peter Y. K. Cheung  
Imperial College London  
p.cheung@imperial.ac.uk

George A. Constantinides  
Imperial College London  
g.constantinides@imperial.ac.uk

This work presents KOCL, a fully automated tool flow and accompanying software, accessible through a minimalist API, allowing OpenCL developers targeting FPGA-SoC devices to obtain kernel-level power estimates for their applications via function calls in their host code [1]. KOCL is open-source, available with example applications at https://github.com/PRiME-project/KOCL. In order to maximise accessibility, KOCL necessitates no user exposure to hardware whatsoever.

Contrary to reliance upon worst-case operating assumptions, knowledge of fine-grained power consumption can support power-aware prototyping and facilitate the deployment of adaptive energy-saving strategies including DVFS, task migration and power gating. Since the provision of multiple power islands within SoCs is often impractical, particularly for reconfigurable devices, our approach provides this information by relating circuit switching activity to power.

When targeted to FPGA-SoCs, such as the Altera Cyclone V we used to evaluate KOCL, OpenCL kernel code is compiled into bespoke hardware accelerators, one per kernel, prior to application execution. KOCL adds additional steps to Altera’s OpenCL tool flow in order to augment kernel accelerators with instrumentation to measure the switching activity of their most power-indicative signals. Although KOCL is specific to Altera’s tools and devices, a Xilinx equivalent would be largely similar in structure and operation.

Probabilistic circuit simulations are performed to rank signals from highest to lowest predicted activity. These so-called ‘vectorless’ simulations are fast and do not require user provision of test vectors. Area-efficient counters are added to the highest-ranked signals, with control logic instantiated to allow for runtime read-back. All of these steps are automated, with none contributing significantly to compilation times.

During execution, KOCL’s software runs alongside OpenCL host code in order to build and continuously update an online power model of the hardware. Kernel-level activity and FPGA-wide power measurements are fed into the model, which apportions the total power between the kernel accelerators present. Users can query realtime power estimates of a particular kernel by simply passing its name to the KOCL\texttt{get()} function within KOCL’s API.

Since KOCL’s power model is online—that is, trained using real data at runtime—it is able to adapt to workload and environmental changes dynamically. Unlike alternative techniques that make use of offline models, KOCL is able to compensate for sources of power behavioural change including input data, operating modes, noise, voltage, frequency, temperature, variation and degradation automatically.

Experimentation has confirmed that KOCL is both a low-overhead and high-accuracy power estimation technique [2, 3]. Typically, relatively few signals need to be monitored in order to achieve good results. With just eight counters per kernel, 10mW absolute accuracy—the difference between modelled and measured kernel-level power—has been found to be obtainable, with errors as low as 5mW achievable when additional counters are used. With low numbers of instruments, area and power overheads are only a few percent.

ACKNOWLEDGEMENTS

This work was supported by the EPSRC-funded PRiME Project (grant number EP/K034448/1).
http://www.prime-project.org

The authors also acknowledge the support of Imagination Technologies and the Royal Academy of Engineering.

REFERENCES