Technical Disclosure Commons

Defensive Publications Series

September 13, 2018

Automatic Hardware Acceleration of Computational Hotspots

N/A

Automatic hardware acceleration of computational hotspots

ABSTRACT

As Moore's law slows down, CPUs offer less annual incremental performance per watt, while demand continues to increase. To maintain economic computational performance in the face of increasing demand, data centers are deploying hardware accelerators specialized for particular tasks, e.g., machine vision, video compression, etc. Creating custom application specific integrated circuits (ASICs) for specialized tasks is expensive and requires a highly skilled team.

The techniques of this disclosure describe a workload-identifying process that automatically identifies computational hotspots amenable to hardware acceleration. FPGA designs for such workloads may be machine generated. The FPGA design is tested, and if found worthy by the measure of economic return-on-investment (RoI), taped out as an ASIC. The RoI is fed back to the workload-identifying process, which uses such feedback to improve identification of economically relevant computational hotspots.

KEYWORDS

- Hardware accelerator
- Design automation
- Rapid prototyping
- Data center
- Machine-generated design
- Machine learning
- FPGA

BACKGROUND

As Moore's law slows down, CPUs offer less incremental performance per watt year over year, while demand continues to escalate. It is widely understood that this will likely result in an increased heterogeneity of the data center, e.g., an increase in the number of distinct hardware accelerators specialized for particular tasks such as machine vision, video compression, etc.

Creating application specific integrated circuits (ASICs) for specialized computational tasks is a fairly involved and expensive process that requires a skilled design team. Indeed, even identifying workloads that are likely to benefit from ASICs is not yet a clear-cut process and requires human experts. Once an ASIC is fabricated, introducing it to the compute fleet is another challenging task.

DESCRIPTION

The techniques of this disclosure perform automatic identification of computational workloads (hotspots) that are amenable to hardware acceleration, e.g., by field programmable gate arrays (FPGAs) or ASICs. An FPGA design may be auto-generated and optimized. The workload is tested on an instrumented test cluster, e.g., a simulated ASIC design running on FPGAs. If tests show likely good performance, as modeled e.g., by off-loading of CPU tasks, degree of acceleration attained, return on investment (RoI), etc., then an ASIC is scheduled for design.

If the task-offloading, RoI analysis, etc. indicate that ASIC fabrication is worthwhile, the ASIC is designed to be hot-swappable and compatible with form factors of the machines of the data center, e.g., the ASIC may have a generic hot-swap PCI-Express mount. The ASIC may be deployed into the data center computers, with swaps and/or decommissioning of machines performed automatically, e.g., by robots, and acceptance testing performed. Measured RoI is fed back to the automatic process that identifies computational hotspots. Such feedback is utilized to improve the identification of economically relevant computational hotspots.

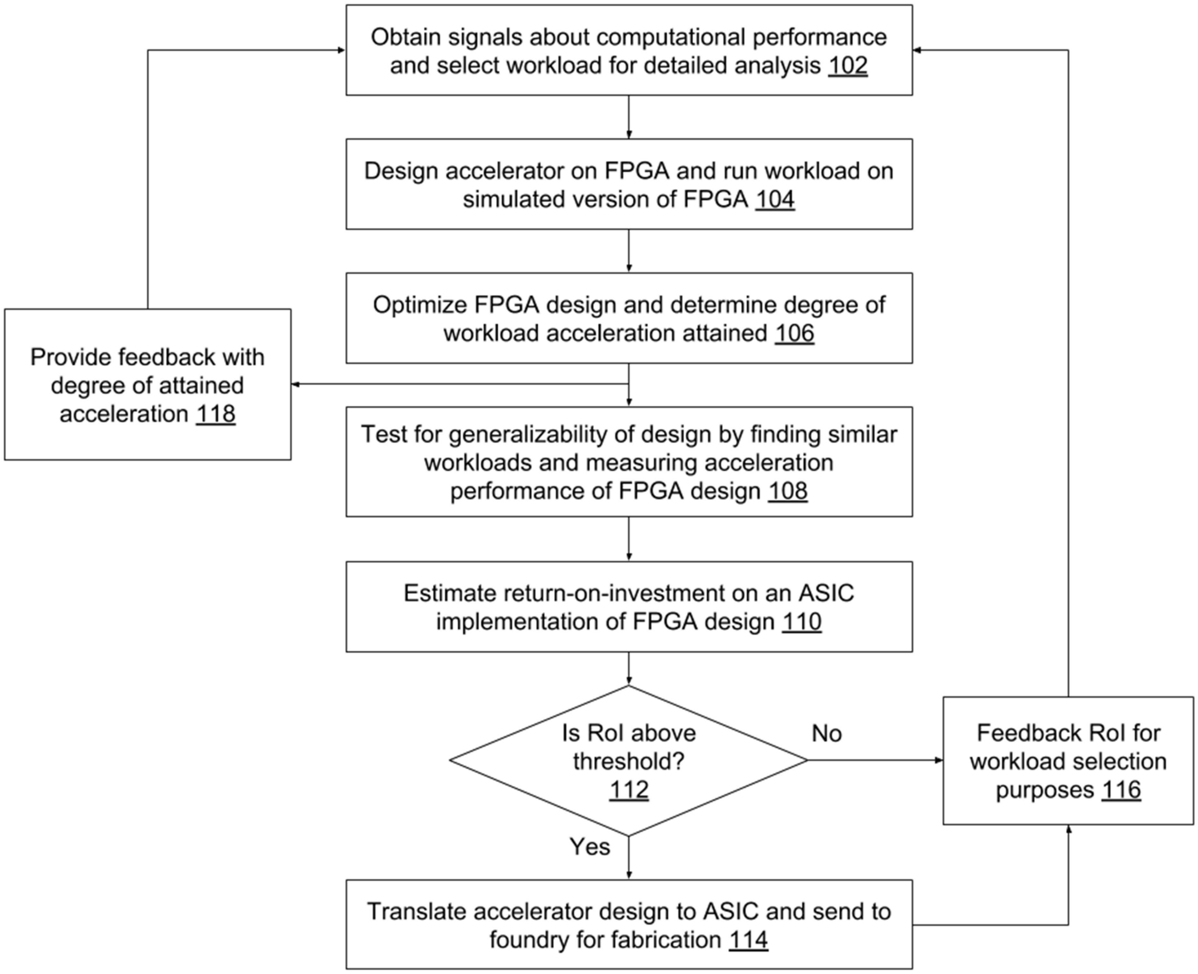
Obtain signals about computational performance and select workload for detailed analysis 102
Design accelerator on FPGA and run workload on simulated version of FPGA 104
Optimize FPGA design and determine degree of workload acceleration attained 106
Provide feedback with degree of attained acceleration 118
Test for generalizability of design by finding similar workloads and measuring acceleration performance of FPGA design 108
Estimate return-on-investment on an ASIC implementation of FPGA design 110
Is RoI above threshold? 112
No
Feedback RoI for workload selection purposes 116
Yes
Translate accelerator design to ASIC and send to foundry for fabrication 114

Fig. 1: Automatic detection of computational hotspots and design of hardware accelerators

Fig. 1 illustrates automatic detection of computational hotspots and design and implementation of hardware accelerators for identified hotspots, per techniques of this disclosure. A primary picker process (also known as “observer process” or simply as “picker process”) obtains signals about computational performance for workloads across a fleet of servers and selects a given workload for detailed analysis (102). The picker process uses machine learning techniques to identify computational hotspots. The picker process at first makes naive selections, e.g., nearly arbitrary workloads with basic, set-by-human heuristics such as:

```
“select processes that
(a) consume more than 80% of their CPU quota,
(b) for more than one hour, and
(c) have been run on at least 5000 computational-load units
    in the past week.”
```

For the purposes of selecting computational hotspots, a unit of computational load may be defined, e.g., as one standard processor-core worth of compute power. Over time, and with feedback from downstream stages, the selection capability of the picker process improves per the mechanisms described herein.

Having selected a target workload or computational hotspot as worthy of hardware acceleration, an FPGA is designed to match the workload, and the workload is run on a simulated version of the FPGA (104). The FPGA design is thoroughly instrumented, such that rich signals about computational run-paths and constraints are exposed. Running the workload on an instrumented test cluster reveals details of the workload, such as:

- workload constraints or gatings due to availability of particular resources, e.g., memory bandwidth, network access, CPU cycles, etc.;
- rates of cache invalidation;
- types of computing units being exercised by the workload, e.g., adders, multipliers, memories, etc.;
- statistics of flows of instructions and data vis-a-vis computing units;
- indications (or not) of parallelizability, e.g., multi-core or single-core use; etc.

It is worthwhile noting that even if an actual hardware accelerator does not eventually result, automatic profiling studies, as described above, can point to directions for improvement in the data center, e.g., the procurement and installation of more memory rather than CPU for memory-constrained processes, routing of certain tasks to servers with superior memory bandwidth, etc.

The FPGA design is optimized and the degree of optimization, e.g., the degree of workload acceleration, is determined (106). The degree of attained acceleration for the workload is fed back to the picker process (118) to train the picker process to better identify optimizable workloads.

To test for generalizability of the discovered optimization, a secondary picker process is utilized to find similar workloads across the server fleet. The similar workloads are run through the same simulated FPGA design (108). The degree of similarity of acceleration to that for the original workload is determined. The simulated optimization efficiency as measured, e.g., by sameness of acceleration as original workload, is used by the secondary picker process as training parameter. At this point, an estimate is obtained of the percentage of fleet tasks that are amenable to acceleration, and to what degree.

A financial solver makes an estimate of the return-on-investment (RoI) for an ASIC implementation of the FPGA design (110). The financial solver uses discounted cash flow analysis to determine RoI and time-to-RoI-positive, and includes in its analysis factors such as estimates of predicted future growth of workloads impacted by the optimization, costs of introducing a new accelerator, risks of errors and omissions, yield risk, supply chain risk, etc.

If the financial solver determines that the RoI is above a certain threshold (112), the accelerator design/architecture is translated to an ASIC platform and sent to a foundry for fabrication (114). The thresholds and performance analyses for FPGA design and full-fledged ASIC design may be different. Thus, it is possible that there is economic value in offloading some workloads to an FPGA, but not to an ASIC. The estimated RoI is fed back to the primary picker process (116) in order to train it to find workloads worth optimizing. In this manner, the RoI estimate acts as a cost function for the picker process. The ultimate efficacy of the picker process is indicated in terms of RoI to enable improvements to the ability of the picker process to identify workloads that are not merely optimizable but workloads that generate monetary savings when hardware-accelerated.

Using RoI as a signal for the machine learning models within the picker process eliminates, for example, complicated ASIC designs that perform reasonably well when measured in terms of off-loading but are too expensive to fabricate. It is worthwhile noting that the automated workload identification and FPGA design techniques described herein may discover a combination of off-the-shelf ASICs that work well with the types of identified workloads. This is also an economically relevant discovery, as it points to the existence of a hardware accelerator without the need to fabricate one for specific workloads.

An ASIC, as created via the techniques described herein, is fabricated in a form factor that matches the machines of the data center to enable quick deployment. Simultaneous with ASIC fabrication, a test program is automatically created to validate the ASIC.

Although the initial workload selections of the picker process may be naive, with feedback received in the form of accelerated performance and RoI generated, over time the ability of the picker process in the identification of workloads amenable to acceleration is improved.

The techniques of this disclosure also apply to evaluation of IP cores that are being considered for integration into systems-on-chips. In particular, the claims of a vendor of third-party IP core can be stress-tested to determine the actual degree of acceleration and the RoI attributable to that IP core.

The custom ASIC, as fabricated with the techniques described herein, may be mounted on a common accelerator card. The common accelerator card derives from a framework that enables many different kinds of ASICs to be mounted on it. It has defined input-output parameters, e.g., 4x lanes of PCI Express 4.0, power characteristics, cooling characteristics, physical size, latching, reparability, etc. These cards are designed for hot-plugging into the machines of the data center, or to bus-attached accelerator housing. The cards are instrumented for automation with fiducials and easy-access latch points.

Mounting of the custom ASIC to this common accelerator card may be done either at a contract manufacturer, or onsite at the datacenter. At the datacenter, acceptance testing is performed by an accelerator test station that exercises the accelerator with a test program. Such acceptance testing can also take place at the foundry or at the integrator and/or applied to a random subset of shipments. Accelerators that have passed testing may be inserted into accelerator hot-plug spots by robots or humans, inserting accelerators into machines that the primary picker has determined as likely to benefit from acceleration.

To maximize utilization as loads shift, accelerators can be physically moved via robot or person from one machine to another. If workloads shift across geographies, e.g., due to jurisdictional requirements to geographically constrain computing or data, accelerators may be cross-shipped.

The individual steps of this disclosure are useful independent of other parts of the disclosure. For example, identification of computational hotspots is useful for data center resource management. Even when an automatic FPGA/ASIC design is not executed, a design team can still develop a solution to address an identified computational hotspot. The RoI and performance/cost analyses have independent utility as an estimate for costs and time for development of the FPGA or ASIC, which can then feed into decisions about which designs to pursue. The RoI and performance/cost analyses can indicate economic value for FPGA insertion, but not for full-fledged ASIC fabrication.

CONCLUSION

As Moore's law slows down, CPUs offer less annual incremental performance per watt, while demand continues to increase. To maintain economic computational performance in the face of increasing demand, data centers are deploying hardware accelerators specialized for particular tasks, e.g., machine vision, video compression, etc. Creating custom application specific integrated circuits (ASICs) for specialized tasks is expensive and requires a highly skilled team.

The techniques of this disclosure describe a workload-identifying process that automatically identifies computational hotspots amenable to hardware acceleration. FPGA designs for such workloads may be machine generated. The FPGA design is tested, and if found worthy by the measure of economic return-on-investment (RoI), taped out as an ASIC. The RoI is fed back to the workload-identifying process, which uses such feedback to improve identification of economically relevant computational hotspots.