Transparent Heterogeneous Cloud Acceleration

Jessica Vandebon
Imperial College London
United Kingdom
jessica.vandebon17@imperial.ac.uk

José G. F. Coutinho
Imperial College London
United Kingdom
gabriel.figueiredo@imperial.ac.uk

Wayne Luk
Imperial College London
United Kingdom
w.luk@imperial.ac.uk

Thomas Chau
Intel Corporation
United Kingdom
thomas.chau@intel.com

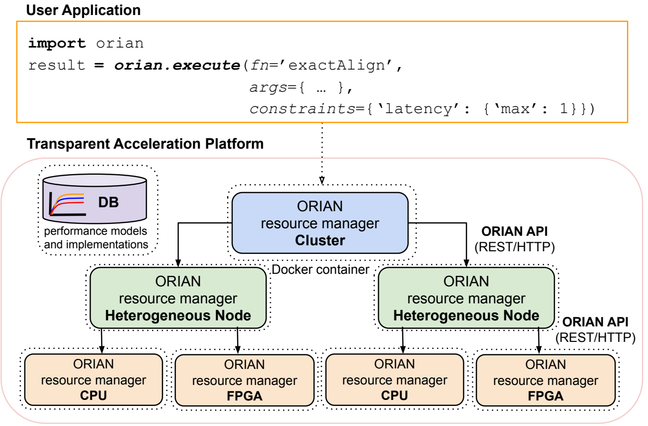
***Abstract*—This work proposes a cloud computing platform (PaaS) with a novel micro-service architecture designed to support transparent acceleration on large-scaled heterogeneous cloud infrastructures with hardware accelerators such as FPGAs.**

***Index Terms*—FPGA, PaaS, heterogeneous clouds, transparent acceleration**

Current cloud providers (most notably Amazon EC2 [1]) offer specialised accelerator resources, such as FPGAs and GPGPUs, to handle complex HPC and AI workloads. This embrace of heterogeneity is found mostly in IaaS (Infrastructure-as-a-Service) offerings, where users can request VMs with access to accelerators, but are responsible for manually managing provisioned resources. In contrast to IaaS, Platform-as-a-Service (PaaS) offerings automatically manage compute resources based on user performance and/or cost requirements, abstracting cloud tenants from this effort. However, current PaaS offerings are limited in their support for hardware accelerators. They require users not only to allocate the most appropriate type and number of compute resources, taking cost into account, but also to optimise the mapping of computations to these resources to meet performance requirements.

A key reason for this limited support of heterogeneity in PaaS is the complexity of acquiring and maintaining knowledge about the suitability of each resource to different workload types and sizes. Resources have a maximum throughput capacity for a certain computation, but the actual throughput achieved varies depending on the given workload. There is no panacea when it comes to a resource configuration that can outperform others for all workloads. For example, a sequence alignment application implementation targeting a resource configuration with 12 CPUs provides the highest throughput for workloads up to 20,000 reads, while for workloads above 20,000 reads, its throughput remains flat and is increasingly outperformed by an implementation targeting 4 FPGAs. This situation motivates the need for an automated decision making process to map jobs to the most effective resource configurations at runtime. Furthermore, such a runtime process is needed to manage cost: the most performant resource might not be the best if the performance gain is less than the cost increase.

To address the above challenges, we propose ORIAN, a heterogeneous PaaS platform that realises automatic, transparent acceleration onto large-scaled heterogeneous cloud infrastructures. We employ a library-based system, where each supported function is associated with multiple implementations targeting different resource configurations, to enable resource-oblivious applications to automatically tap on to specialised resources. Our platform operates in two modes: offline and online. When offline, library functions are profiled, and empirically derived performance models are generated for each implementation. When online, these derived performance models are used to automatically select the implementation expected to minimise execution time and/or cost for a submitted job. Our resource management architecture is made up of modular, hierarchical resource managers, organised to reflect the levels of complexity of a hardware infrastructure (data centre managers, cluster managers, node managers, and so on). All resource managers share the same interface, allowing the architecture to grow with and adapt to the underlying hardware platform.


User Application
Transparent Acceleration Platform
DB
performance models and implementations
ORIAN resource manager Cluster
ORIAN API (REST/HTTP)
Docker container
ORIAN resource manager Heterogeneous Node
ORIAN resource manager Heterogeneous Node
ORIAN API (REST/HTTP)
ORIAN resource manager CPU
ORIAN resource manager FPGA
ORIAN resource manager CPU
ORIAN resource manager FPGA

```
import orian
result = orian.execute(fn='exactAlign',
                       args={ … },
                       constraints={'latency': {'max': 1}})
```

Fig. 1. ORIAN prototype architecture.

We illustrate our approach in Fig. 1, in which a Python application invokes a managed function (`exactAlign`), specifying its arguments and a timing constraint. In this case, the transparent acceleration platform automatically decides and executes this function with the most cost efficient resource configuration under 1 second.

Our current ORIAN prototype is deployed on a hardware platform with access to up to 36 CPU cores and 28 FPGAs. Managers operate in isolated Docker containers and communicate via a generic REST/HTTP API. Future work includes extending our work to combine transparent acceleration with heterogeneous elasticity.

Acknowledgement. The support of the United Kingdom EPSRC (grant numbers EP/L016796/1, EP/N031768/1, EP/P010040/1 and EP/L00058X/1), Maxeler and Intel is gratefully acknowledged.

References

[1] "Amazon EC2." [Online]. Available: https://aws.amazon.com/ec2/