Journal of Systems Architecture 000 (2016) 1–16

Contents lists available at ScienceDirect

Journal of Systems Architecture

journal homepage: www.elsevier.com/locate/sysarc

An energy-efficient system on a programmable chip platform for cloud applications

Xu Wang[a], Yongxin Zhu[a,f,*], Yajun Ha[b], Meikang Qiu[c], Tian Huang[a], Xueming Si[d], Jiangxing Wu[e]

[a] *School of Microelectronics, Shanghai Jiao Tong University, 800 Dongchuan Road,Shanghai 200240, P.R. China*
[b] *Department of Electrical and Computer Engineering, National University of Singapore, 21 Lower Kent Ridge Road, Singapore*
[c] *Department of Computer Science, Pace University, 1 Pace Plaza, Manhattan, New York City, NY, 10038, USA*
[d] *Shanghai Key Laboratory of Data Science, Fudan University, 825 Zhangheng Road, Shanghai 201203, P.R. China*
[e] *PLA Information Engineering University, P.R. China*
[f] *School of Computing, National University of Singapore, Singapore*

ARTICLE INFO

Article history:
Received 30 January 2016
Revised 8 October 2016
Accepted 22 November 2016
Available online xxx

Keywords:
Cloud computing
ECG classification
FPGA
Performance analysis
Reconfigurable architectures
Web server

ABSTRACT

Traditional cloud service providers build large data-centers with a huge number of connected commodity computers to meet the ever-growing demand on performance. However, the growth potential of these data-centers is limited by their corresponding energy consumption and thermal issues. Energy efficiency becomes a key issue of building large-scale cloud computing centers. To solve this issue, we propose a standalone SOPC (System on a Programmable Chip) based platform for cloud applications. We improve the energy efficiency for cloud computing platforms with two techniques. First, we propose a massive-sessions optimized TCP/IP hardware stack using a macro-pipeline architecture. It enables the hardware acceleration of pipelining execution of network packet offloading and application level data processing. This achieves higher energy efficiency while maintaining peak performance. Second, we propose a on-line dynamic scheduling strategy. It can reconfigure or shut down FPGA nodes according to workload variance to reduce the runtime energy consumption in a standalone SOPC based reconfigurable cluster system. Two case studies including a webserver application and a cloud based ECG (electrocardiogram) classification application are developed to validate the effectiveness of the proposed platform. Evaluation results show that our SOPC based cloud computing platform can achieve up to 418X improvement in terms of energy efficiency over commercial cloud systems.

1. Introduction

We are living in an era of cloud computing that is the backbone of embedded ubiquitous computing [1] and big data [2]. People can access powerful IT resources (e.g., computation and storage) and convenient services in the cloud via mobile devices at any time and any place. The shift of computing infrastructure from local desktop to the remote cloud improves the utilization of overall computing resources and reduces the costs associated with management of hardware and software resources for individual users.

Many services on the cloud are web-based applications, which incur a large number of concurrent requests and each of the requests involves a light-weighted task which often desires short latencies in response from energy hunger servers. For existing servers, the performance is typically limited by overheads of the network packets processing and the connection management in the NIC and OS kernel [3]. Although traditional cloud service providers such as Google and Amazon build data centers with a huge number of connected commodity computers to meet the demand of cloud computing, the ever-growing energy consumption limits the scale out of these machines due to limited energy-budget. Moreover, it comes expense related to the energy consumption and heat dissipation dominating the operating costs in such high-density computing environments. Therefore, the energy efficiency (performance per joule) becomes the key issue of building large-scale cloud computing centers [4].

Reconfigurable computing based on Field Programmable Gate Array (FPGA) technologies possesses the capability of parallel and specialized computation that can effectively exploits the task-level parallelism inherent in cloud computing with relatively low energy consumption [5]. The current trend to take advantage of FPGA in

* Corresponding author.
E-mail addresses: wang2002xu@gmail.com (X. Wang), zhuyongxin@sjtu.edu.cn, yongxin.zhu@nus.edu.sg (Y. Zhu), elehy@nus.edu.sg (Y. Ha), mqiu@pace.edu (M. Qiu), huantian@ic.sjtu.edu.cn (T. Huang), sxm@fudan.edu.cn (X. Si).

http://dx.doi.org/10.1016/j.sysarc.2016.11.009
1383-7621/© 2016 Elsevier B.V. All rights reserved.

cloud computing lies in two ways: (1) to improve data transfer performance within data centers as a TCP/IP offload engine (TOE) [6]; (2) to accelerate computation-intensive applications as a hardware accelerator [7,8]. In both two methods, FPGAs are used as slave devices hosted by commodity CPUs on the master side. However, such master-slave based architecture is not suitable for web-based cloud applications, since the high communication latency between the master CPU and the slave FPGA limits the throughput of the system.

In this paper, we analyze the inefficiency of master-slave based reconfigurable architectures in dealing with highly concurrent cloud applications by proposing a performance model, and indicate that a standalone SOPC based architecture, which tightly couples network IO handling and data processing in a single FPGA, is a more promising solution in cloud computing with both high energy efficiency and high performance catering to needs of cloud applications.

Having identified a promising standalone SOPC based architecture, we then present the design and implementation of the SOPC based architecture for cloud computing in details. The major challenge to implement such SOPC based architecture is the design of a high performance TCP/IP hardware stack that is able to support high network throughput with high concurrent connections. Since third party TCP/IP offload engines are primarily optimized for inter-server data transfer among a few TCP connections in data centers, they adopt one-TCP/IP-session-per-pipeline architecture to achieve high throughput and extremely low latency. However, such TCP/IP engines are hard to scale due to resources limitation in FPGA, which is not suitable for high concurrent cloud applications. We address this problem by adopting a macro-pipeline architecture, where all TCP sessions share a centralized memory and coarse grained pipeline. Processing modules in the macro-pipeline are operated in asynchronous mode to achieve high system level throughput, while an external SRAM is used to minimize the access latency of TCP connection information that are randomly accessed and modified by different modules concurrently.

One advantage of our design by placing FPGAs on the cloud as standalone entities to provide service is its high energy efficiency at its peak performance. In other words, it supports higher performance within a fixed energy-budget constraint in data centers. Besides the efforts in hardware design, we further propose a online dynamic scheduling scheme to reduce the runtime energy consumption of the whole cluster system. By taking advantage of fast configuration ability, a load balancing node is used to shut down or power up FPGAs according to realtime workload variance in the cloud environment, saving the energy without compromising the quality of service to users.

To verify our architecture design and scheduling scheme, two case studies including a web server cluster and a cloud based ECG (electrocardiogram) classification cluster are presented to evaluate the effectiveness of proposed architecture. Design consideration and implementation details are given to show how this architecture meets the demand of cloud computing in high throughput, high concurrency, low latency, as well as low energy consumption. The major contributions of this work include:

- Performance analysis of master-slave based reconfigurable architecture and standalone SOPC based reconfigurable architecture using an analytical model.
- A massive-sessions optimized TCP/IP offload engine (described in Section 3.A) that supports up to 100K TCP sessions under 10Gbps line rate.
- An online dynamic scheduling method (described in Section 3.B) that can reconfigure or power off FPGA nodes according to workload variance to reduce the runtime energy consumption.
- Prototype of a reconfigurable cluster system based on standalone SOPC to provide cloud services, achieving up to 38X speed up in performance and 418X improvement in energy efficiency compared to the software based cloud systems.
- Prototype verification by implementing an I/O intensive web server cluster system with detailed comparison with alternate servers.
- Prototype verification by implementing a both I/O and computation intensive ECG classification cluster system, indicating support for practical cloud applications.

The rest of the paper is organized as follows. Section 2 analyzes related work briefly. Section 3 proposes a performance model to analyze the master-slave based architecture and the standalone SOPC based architecture. Section 4 describes the design details of standalone SOPC based cluster systems for cloud applications. Section 5 presents the design and implementation of two case studies including a web server cluster and a cloud based ECG classification cluster. Section 6 evaluates performance and energy efficiency of the two application cases. Our conclusions are presented in Section 7.

2. Related works

The issue of energy consumption in cloud computing has been receiving increasing attention in recent years [9]. Researchers have considered energy minimization by consolidating the workload into the minimum of physical resources and switching idle computing nodes off, with guaranteed throughput and response time. For example, Chase et al. [10] proposed an economic approach to managing shared server resources, which improves the energy efficiency of server clusters by dynamically resizing the active server set. Similar dynamic provisioning algorithms [11] are studied for long-lived TCP connections as in instant messaging and gaming. Moreover, a queuing model [12] to dynamic provisioning technique has also been studied to obtain the minimum number of servers and a combination of predictive and reactive methods has been proposed to determine when to provision these resources. Although such scheduling methods reduce energy consumption by switching idle servers to power saving modes (e.g. sleep, hibernation or power off), the long switch latency consisting of reinitialization of the system and restoring the context in commercial servers prevents their usage in real cloud systems, especially when the workloads are unstable. Recent study [13] utilises the power gating technique for FPGA-based accelerators to efficiently reduce the energy consumption of tasks running on single FPGA. However, as far as we know, there is no related work on dynamic scheduling of FPGA based cloud computing system to reduce the energy consumption.

To ensure fast changing system configuration and sharing of systems resources, Li et al. [14] proposed a rack scale composable system using PCIe switches, which is constructed as a collection of individual resources such as CPU, memory, GPU/FPGA accelerator, disks etc., and composed into workload execution units on demand. Although such composable system is a promising architecture that can improve the utilization of overall systems resources, the relative high communication latency between master CPU and slave devices (memory,GPU/FPGA) makes it less efficient in dealing with latency-bounded applications such as web-based cloud services. Compare to it, our standalone SOPC based architecture tightly couples network IO handling and data processing in a single FPGA, which significantly reduces such communication latency. Recent study [15] shows that energy consumption in network processing can be a significant percentage of total energy consumption in cloud computing. Neither high-performance nor low-power cores provide a satisfactory energy-performance trade-