Ning-hui SUN, Yun-gang BAO, Dong-rui FAN, 2018. The rise of high-throughput computing. *Frontiers of Information Technology & Electronic Engineering,* 19(10):1245-1250.
https://doi.org/10.1631/FITEE.1800501

The rise of high-throughput computing

Key words: High-throughput computing; Sysentropy; Information superbahn

Corresponding author: Ning-hui SUN
E-mail: snh@ict.ac.cn
ORCID: Ning-hui SUN, http://orcid.org/0000-0002-4179-2660

Motivations

1. Emerging applications and advanced materials are the two major forces driving computer system technologies.

2. Emerging applications, such as cloud computing, artificial intelligence (AI), and the Internet of Things (IoT), have posed three major requirements, high utilization, high throughput, and low latency.

3. AI and IoT applications are often deployed in clouds costing billions of dollars, which require high utilization of cloud datacenters, high throughput, and low end-to-end latency being critical to user experience.

Main ideas

1. We propose ‘high-throughput computing’ (HTC) to refer to the collective requirements of high utilization, high throughput, and low latency.

2. We introduce a new indicator, referred to as ‘sysentropy’, to measure the degree of latency variation of a system operating at a certain utilization level and throughput.

3. We introduce two techniques that can achieve low sysentropy: the chip-level dataflow architecture SmarCo (Fan et al., 2018) and the labeled von Neumann architecture (LvNA) (Ma et al., 2015; Bao and Wang, 2017).

Technique

1. The chip-level dataflow architecture SmarCo:

Enhance control over the internal data paths of a computer system through labeling and programmable control logic mechanisms to maintain low latency when resource utilization increases.

2. The labeled von Neumann architecture (LvNA):

Fully exploit the parallelism within applications and the concurrency among requests and thus increase the utilization and throughput of computer systems.

Prototypes

High-throughput many-core processor DPU

1. The prototype chip (Fig. 4) is based on the TSMC 40-nm technology and consumes less than 4 W.
2. An accelerator board that targets high-throughput video processing scenarios on the Internet.
3. The energy efficiency is improved by more than 20 times (Compared to the mainstream Intel processor).

Fig. 4 High-throughput many-core processor DPU chip

Prototypes (Cont'd)

Labeled RISC-V

1. A physical server can be partitioned into multiple submachines and directly load the operating system (OS) without the need for software hypervisors.
2. The hardware supports real-time monitoring (<1 ms) without software overhead.
3. The system supports performance isolation, such as dynamic allocation of cache capacity and memory bandwidth.

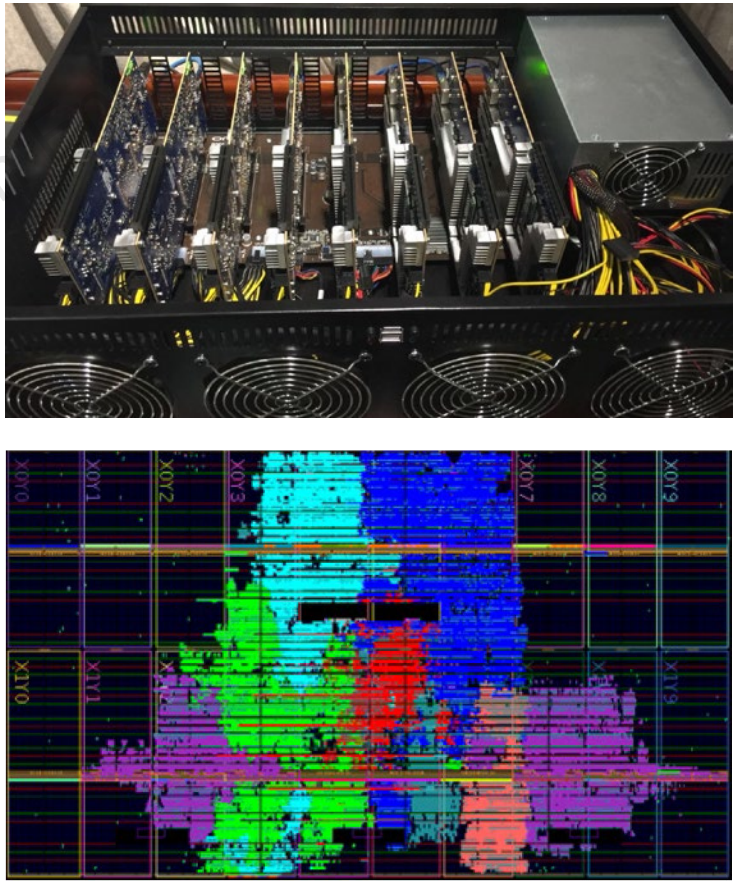

Fig. 5 Labeled RISC-V prototype system based on FPGA

Conclusions

1. HTC has distinct features that require high utilization, high throughput, and low latency, while traditional computer system designs have not yet considered these three requirements at the same time.

2. We have introduced two techniques to achieve low sysentropy: dataflow and LvNA architectures. The prototypes and experimental results preliminarily verified the feasibility of high-throughput computers.

3. High-throughput computers will also become the core devices of future information infrastructure, and will provide low-cost and high-quality information services for billions of people around the world.