FPGA-BASED MULTI-CORE PERFORMANCE ANALYSIS OF ARCHITECTURAL TECHNIQUES

K. Ashokkumar¹, P Gowrishankar², V Balaprakash³ and S Sudha⁴

¹³⁴ Research Scholar, Department of Electronics & Communication System, Hindustan CAS
²Prof and Head, Department of Electronics & Communication System, Hindustan CAS

Abstract- Since the very beginning of integrated circuits development, processors were invented with ever-increasing clock frequencies and sophisticated in-build optimization strategies. Physical constraints put up multiple barriers in achieving high performance computing within power budgets. In the past, processor design trends were dominated by increasingly complex feature sets, higher clock speeds, growing thermal envelopes and increasing power dissipation. Recently, clock speeds have tapered and thermal and power dissipation envelopes have remained flat. However, the demand for increasing performance continues which has fueled the move to integrated multiple processor (multi-core) designs. This paper discusses this trend towards multi-core processor designs, the design challenges that accompany it and a view of the research required to support it.

Keyword- Multicore, ASIC, Pipelining, FPGA.

I. INTRODUCTION

Applications that require energy efficient high performance computing are becoming increasingly commonplace. These applications often comprise multiple computing tasks and are frequently key components in many systems such as: wired and wireless communications, multimedia, large-scale multi-dimensional signal processing (e.g., medical imaging, synthetic aperture radar), some large-scale scientific computing, remote sensing and processing; and medical/biological processing. Many of these applications are embedded and are strongly energy constrained (e.g., portable or remotely located) and cost-constrained. In addition, many of them require very high throughputs, often dissipate a significant portion of the system power budget, and are therefore of considerable interest.

There are several design approaches to achieve high performance computing solutions such as ASICs, programmable processors/DSPs and FPGAs. ASICs can provide very high performance and very high energy efficiency, but they have little programming flexibility. On the other hand, programmable processors/DSPs are easy to program but their performance and energy efficiency is normally 10–100 times lower than ASIC implementations. FPGAs fall somewhere in between. One-time fabrication costs for state of the art designs (e.g., 90 nm CMOS technology) are roughly one million dollars, and total design costs of modern chips can easily run into the tens of millions of dollars. Programmable processors are the platforms which allow high one-time fabrication costs and design costs to be shared among a variety of applications/users; but higher performance and energy efficiency are expected to be achieved using architectural innovations. IBM first introduced multi-core processor chip, Power4 in 2001 [15] through which designers were able to achieve much greater communication bandwidth and resulting performance. In mid-2006, Intel reached new levels of energy-efficient performance with their Intel CoreTM2 Duo processors using 65 nm technology and latest microarchitecture [2]. Although it has been a frequently used architecture, numerous challenges involve accordingly and they must be addressed by the researchers. The rest of this paper is organized as follows. Section II briefly gives some major advantages of multi-core processor. It is concluded that multi-core processors become the standard for delivering greater performance, improved performance per watt and new capabilities across...
different electronic applications. Section III describes the leading interconnect challenges in multi-core processors. Challenges incurred by design automation and verification and software adaptability have been briefly studied in section IV and V respectively. Finally section VI wraps up the paper.

II. PERFORMANCE CONSIDERATION

Increasing the processor clock frequencies and using wide issue processor architectures have worked well to improve performance but recently have become significantly more challenging. Deeper pipelining is one of the key techniques to increase the clock frequency and performance, but the benefit of the deeper pipeline is eventually diminished when the inserted Flip-Flop’s delay is comparable to the combinational logic delay. Moreover, deeper pipeline stage increases cycles-per-instruction (CPI) and impacts negatively to the system performance. Performance increase by microarchitecture alone is governed by Pollack’s Rule, which states that performance increase is roughly proportional to square root of increase in complexity. In other words, if you double the logic in a processor core, then it delivers only 40 per cent more performance. Researchers found that the depth per pipeline is approximately 8 Fanout-4 (FO4) inverter delays to obtain highest performance which corresponds to 20–30 pipeline stages. The pipeline delay of some modern processor is already close to 10 FO4 so that the deeper pipelining technique for high performance is reaching its limit. Also, increasing pipeline stages necessitates more registers and control logic, thereby further increasing design difficulty as well as power consumption. The optimum pipeline depth for maximum energy efficiency is about 22.5 FO4 delay (about 7 stage pipeline), using BIPS3/Watt as the metric—BIPS are billions of instructions per second. Invariably, a faster functional unit (such as an arithmetic logic unit or memory interface) is larger and more expensive than a slower functional unit. Ullman shows that an implementation is theoretically bounded by if volume (I/O, power dissipation, and so on) limits it; or AT2 in implementation that are communications limited. So depending on the type of functional unit and circuit implementation, doubling processing speed could increase die size by 2 to 4 times

\[ A \times T^n \]  (1)

where n is typically between 1 and 2. Besides exploiting fabrication process advancements, computer architects employed wide issue processor architectures such as VLIW and superscalar approaches for high performance computation. While their benefit is also quickly diminished when the issue width is more than 10; since most applications don’t have so many independently parallel executable instructions in per fetch.

III. MULTI-CORE SYSTEMS

To address the challenges in performance, power and future technologies, innovations on computer architecture and design are needed; and multi-core systems are one of the most, or the most promising technology. As was also pointed out by a computer architecture group at EECS department of UC Berkeley the “shift toward increasing parallelism is not a triumphant stride forward based on breakthroughs in novel software and architectures for parallelism; instead, this plunge into parallelism is actually a retreat from even greater challenges that thwart efficient silicon implementation of traditional uniprocessor architectures”. Deep submicron fabrication technologies enable very high levels of integration such as a dual-core chip with 1.7 billion transistors, thus reaching a key milestone in the level of circuit complexity possible on a single chip. A highly promising approach to efficiently use these circuit resources is the integration of multiple processors onto a single chip to achieve higher performance through parallel processing, which is called a multi-core system or a chip multiprocessor. Multi-core systems can provide high energy efficiency since they can allow the clock frequency and supply voltage to be reduced together to dramatically reduce power dissipation during periods when full rate computation is not needed. Giving a simple example, assuming one uniprocessor is capable of
computing one application with clock rate $F$ and voltage $V$, and consuming power $P$; now if using a dual core system and assuming the application can be partitioned into two cores without any overhead, then each core only needs a clock rate $F/2$ and the voltage can be reduced accordingly; assuming a linear relation between voltage and clock frequency and the voltage is reduced to $V/2$, then the power dissipation of the dual core system will only be about $P/4$. Multi-core systems also potentially provide the opportunity to independently control each processor’s clock and voltage to achieve higher energy efficiency, if different processors are in separate clock and voltage domains. Furthermore, multi-core systems are suitable for future technologies. The distributed feature can potentially constrain the wires into one core and eliminate the global (long) wires. The multi-core systems also provide flexible approaches to treat each core differently by adjusting the mapped application, supply voltage, and clock rate; to utilize each core’s specific features due to variations. For example, when one processor in the chip is much slower than the others, a low workload can be mapped on it without effecting system performance. The multi-core systems have high scalability since a single processor can be designed and the system can be obtained by combining multiple processors. Thus the systems can be easily adjusted according to the required performance and cost constraints by changing the number of cores; which is much easier than changing the issue-width or the pipelining of uni-processors. A multi-core CPU combines multiple independent execution units into one processor chip, in order to execute multiple instructions in a truly parallel fashion. The cores of a multi-core processor are sometimes also denoted as processing element or computational engine. According to Flynn's taxonomy, the resulting systems are true multiple-instruction multiple-data (MIMD) machines, able to process multiple threads of execution at the same time. Achieving processing speed up by using parallel execution units is obviously nothing new. The concept of the first multiprocessor computer goes back to the 60's with initial architectures such as the ILLIAC IV. The resulting challenges were also discussed to large extend, sometimes even 25 years ago. What is the difference now? The true paradigm shift today is not the realization of CMP (Chip Multiprocessor) architectures, but their spreading in all computer markets. Embedded systems, mobile phones, desktop systems and server systems now include multiple cores out of the box. A parallel computer is no longer a dedicated hardware setup for special purposes. It is commodity. While parallel computing of the past was only intended for a specific set of problem domains, it is now relevant for every scientific, industrial or private application running on a computer. Parallel computing now becomes visible for millions of industrial software developers in practice, who usually lack experience in the creation and handling of fine-grained parallel activities. For academia, the shift will influence both future teaching and research in computer science and software engineering. Many traditional areas such as computer hardware architecture, programming languages, design patterns, scheduling theory and parallel algorithms will gain more attention in the future, since the upcoming research challenges demand answers from these fields. In the following text, an attempt is made to provide a high-level overview about some of the identified issues in this area. I base the argumentation on one fundamental statement: "40 years of parallel computing need to be considered."

IV. PROMISING ASPECTS OF MULTI-CORE PROCESSORS

The key driving force to adopting multi-core processor architecture was to address power and cooling challenges. Figure 1 gives performance comparison between a single core and multi-core processor [2].

This analysis which is performed based on Intel tests using the SPECint2000 and SPECfp2000 benchmarks reports that multi-core processors perform much better than a single core processor and it is projected that relative advantage of multi-core system will enhance over the next couple of years.
A. Controlling power consumption by multi-core processor

Historically chip manufacturers have met the demand for increasing processor speed by boosting up the operating clock frequency along with the higher integration density. This approach has resulted in uncontrollable heat dissipation in current technology node. With heat rising incrementally faster than the rate at which clock signal propagates through the processors, it has prompted the processor designers for alternative methodologies. Multi-core processors take advantage of a fundamental relationship between power and frequency. By incorporating multiple cores, each core is able to run at a lower frequency, dividing power among them normally given to a single core. The result is a big performance increase over a single core processor. It can be observed that increasing clock frequency by 20% to a single core delivers a 13% performance gain, but requires 73% greater power. Conversely, decreasing clock frequency by 20% reduces power usage by 49%, but causes only 13% performance loss [2]. If a second core is added into the single core architecture, it results in a dual-core processor that at 20% reduced clock frequency; it can effectively deliver 73% more performance while using approximately the same power as a single-core processor at maximum frequency.

B. Efficient usage of chip area by Caches and Memory modules

As stated earlier, each single processor core in a multicore architecture has its unique L1 cache and all processors in the die share a common L2 cache. Therefore, number of caches and memories required become less than if single core processor is used for the equal number of jobs that need to be performed. For example, Intel Advanced Smart Cache works by sharing the L2 cache among cores so that data are stored in one place that each core can access. Sharing L2 cache enables each core to dynamically utilize even up to 100% of available L2 cache, thus optimizing cache resources [2]. Intel® Smart Memory Access improves system performance by optimizing available data bandwidth from the memory subsystem and hiding the latency of memory accesses through two techniques: a new adaptability called memory disambiguation, and an instruction pointer-based pre-fetcher that fetches memory contents before they are requested [2].

C. Performance enhancement by multi-threading technology

Along with parallel processing method, multi-threading technology is extensively used in single core processor. According to this approach, on a single processor, multithreading generally works on the principle of time-division multiplexing which is much similar to the parallel execution of multiple tasks where the processor switches between different threads [13]. This context switching happens so fast that it creates the illusion of simultaneity to an end user. On a multiprocessor system, threading can be achieved via multiprocessing, where as different threads and processes can run simultaneously on different processor cores. Threading a task in parallel processing machines thus not only increases the number of tasks executed per unit time but also enhances the accuracy of the task. Consequently, it is obvious that significant performance improvement can be achieved using multicore systems coupled with advances in memory, I/O, and storage devices.

V. DESIGN AUTOMATION CHALLENGES IN MULTICORE PROCESSOR

CAD tools are very important for the verification and optimization of high density integrated circuits from early phase of performance analysis to the final. The verification of a large block has been one of the major challenges in electronic design automation [2]. As the push for more functionality in a single core along with increasing power gating, variable frequencies and asynchronous communication, especially in ASIC design has grown over the years, the verification process has been more complex. Different functioning processors in a multi-core unit have different types of working principles. Therefore, their internal architecture is also different which causes onerous challenges to the verification teams to individually check on each processor. It complicates and creates additional delay to the chip
production cycle [7]. Each core can be functioning at a different mixture of instructions leading to wide variations of dynamic power dissipation. Therefore, accurate analysis by a CAD tool becomes very hard to predict this thermal behavior. In the early phases of processor design, chip layout, power analysis and packaging are divided among different teams. Considering all realistic assumptions about architecture, packaging and market requirements as well as alternatives, a performance model is formed. In physical architecture, each functioning cell is modeled as a block. A particular block generally acts as standard for reuse so many times in a single processor chip. Thus it saves design challenges and time. However, due to different performance requirements in different processor cores in a multi-core system, there is no standard method for minimizing layout optimization. A number of unique controlling blocks are required to properly design the whole multi-core system and it is not possible to implement all blocks using a common technology. This becomes more important when performance and power efficiency are key design criteria. In the final phase, a rigorous testing is performed on each processor to ensure the desired performance. There is no common test platform for multi-core processor systems.

VI. CONCLUSION

It is worth mentioning that introduction of multi-core processors has opened many new doors for high performance integrated circuits. While device scaling is approaching its limit, multi-core processors have continued improving the performance of CMOS technology by splitting the instruction executions among different functioning cores with the addition to parallel processing and multithreading technique. However, like any technology, with the advancement to the successive generations, numerous challenges such as interfacing different cores, innovative CAD tools and software adaptability for new architecture will be more complex and these issues must be addressed for ensuring the optimum performance of multi-core processors. Besides, single core chips will continue to compete since they have well established and inexpensive technology for manufacture and therefore, they will even be popular for low-priced PCs and servers.

REFERENCES

[1] J. Cong, L. He, C-K. Koh, and P. Madden, “Performance optimization of VLSI interconnect,” The VLSI Journal Integration, Vol. 21, pp. 1–94, November 1996.
[2] www.intel.com
[3] International Technology Roadmap for semiconductors, Semiconductor Research Corporation, 2008
[4] Deepak C Sekar and James D Meindl, “The impact of multi-core architectures on design of chip-level interconnect networks,” International Interconnect Technology Conference, pp. 123-125, June 2007
[5] Isci, C.; Buyuktosunoglu, A.; Chen, C.-Y.; Bose, P.; Martonosi, M, “An Analysis of Efficient Multi-Core Global Power Management Policies: Maximizing Performance for a Given Power Budget” Annual IEEE/ACM International Symposium on microelectronics, pp. 347-358, December 2006
[6] Rajarshi MukherjeeT and Seda Ogrenci Memik “Physical Aware Frequency Selection for Dynamic Thermal Management in Multi-Core Systems” International conference on Computer Aided Design, pp. 347-352, November 2006
[7] John A Darringer, “Multi-Core Design Automation Challenges,” Design automation conference, pp. 760-764, June 2007
[8] H. Peter Hofstee, ”Future Microprocessors and Off-Chip SOP Interconnect,” IEEE Tran. On advanced packaging, Vol. 27, No. 2, May 2004
[9] A Chandrakasan, S Sheng and R Brodersen, “Low power CMOS digital design,” Journal of Solid state Cir, April 1992
[10] Rakesh Kumar, Victor Zyuban, Dean M. Tulsan, “Interconnections in multi-core architecture: understanding mechanisms, overheads and scaling,” International Symposium on Computer Architecture, pp. 1-11, 2005