Future Generation Processors
A Technology Perspective

Sapan Kr. Gupta, Rahul Pandey, Manuj Kumar
M.Tech, Microelectronics and Embedded Technology
JAYPEE Institute of Information Technology, Noida

Abstract

Future generation microprocessors are expected to exhibit much heavier loads, faster response along with a new power management trend. This paper deals with some parameters and technological aspects that may even force to change the basic technology considerations for future generation microprocessor. The paper describes about the scope for the current available featuring technology for next generation processor and the long term validation for the same..

1. Introduction

Microprocessor design progresses from tens of millions of transistors on a chip to approximately a billion transistors on a chip using 22nm technology and finer process technologies, the microprocessor designer faces unprecedented circuit, software and featuring algorithm challenges over the future generations of microprocessors. This paper describes the changes in the design environment, current technological modifications, and parameter specifications along with the processes availability and suitability that will be necessary to develop increasingly complex processors.

Today Intel, Microsoft, Bradcomm and many technical leaders are working on different technologies available in microprocessor such as scaling, speed barrier, power management, multicore, parallelism, and even on alternate composition options for Si. This paper presents a broad view for the different technological features available and an analysis about the modifications and innovations that may become a part of microprocessor in future taking different parameters and specifications into consideration.

2. Circuit and software challenges

In the circuit specifications power and energy efficiency is having a trend for continuous dynamic voltage and frequency scaling along with power gating and reactive power management. In future it may be switched into discrete dynamic voltage and frequency scaling along with near threshold operation with proactive fine grain power and energy management

- For Gradual, temporal, intermittent, and permanent faults Guard-bands, yield loss and core sparing, design for manufacturability resilience with hardware/software co-design, dynamic in-field detection, diagnosis, reconfiguration and repair, adaptability, and self-awareness may be implemented.

- Speed binning of parts, corrections with body bias or supply voltage changes, tighter process control with Dynamic reconfiguration of many cores by speed may be found in future circuits.

- Data parallel languages and "mapping" of operators, library and tool-based approaches with a coordination of new high-level languages, compositional and deterministic frameworks are needed for the software heterogeneous parallelism.

- Manual control, profiling, maturing to automated techniques (auto-tuning, optimization) with new algorithms, languages, program analysis, runtime, and hardware techniques may be a solution to energy-efficient data movement.

- Algorithmic, application-software approaches, adaptive checking and recovery with new hardware-software partnerships that minimize rechecking energy may be proven very efficient for Resilience.

- For automatic fine-grain hardware management, self-aware runtime and application-level techniques that exploit architecture features for visibility and control is a necessity for energy management for the software architecture.

3. Parameter specifications and   Customization

In the future some basic parameters that are foundation of present technological aspects may vary. Present needs may result a new view for the technology. Here some present parameters and technologies are discussed in a new respect highlighting their loop-holes and modifications with long term applications.

3.1 Scaling

Scaling is to reduce all the dimensions of a transistor by a factor. In addition to the dimensions decreasing, a number of other parameters must increase or decrease. Constant electric field scaling is required to be used as this provides constant power density (due to thermal conductivity of the substrate).

Parameters	Effect of scaling
Doping Concentration	k
Voltage	1/k
Electric Field	1
Carrier Velocity	1
Depletion layer width	1/k
Capacitance	1/k
Current I_d	1/k
Circuit Delay Time	1/k
Power dissipation per	$1/k^2$
Power delay product	$1/k^3$
Circuit Density	k^2
Power Density	1

Table 1. Scaling Parameters Analysis

Transistor dimensions are scaled 30% every 2 years
- Area shrinks by 50% every two years which is base of moore's law.

- In every two years 40% performance increases due to delay reduction and frequency increase.

- To keep electric field constant in the transistor supply V is reduced by 30% so power by 50% and energy by 65% every two years.

3.2 Composition: Si to Graphene

Graphene is the thinnest and strongest material ever isolated. Graphene offers the seductive prospect of fast, nanoscale electronic devices. However, making and manipulating graphene is a tough task, and the techniques are not yet in place to fabricate devices commercially. Nevertheless, there are many other potential applications for graphene, which could reach the market over the next few years.

The electrical charge carriers in graphene move, unimpeded, at speeds 10–100 times faster than in today's silicon chips – and at normal temperatures. Furthermore, graphene is stable in air, transparent and flexible. Being carbon, the source material should be cheap and plentiful. It is not surprising that industry is excited by graphene's technological potential.

The charge carriers are clearly a little unusual, and are a result of interactions within graphene's rich, and symmetrical periodic arrangement of electrons spread across its hexagonal lattice, which creates waves of electric charge known as quasiparticles. These behave a bit like photons of light – that is, as if they are massless.

One extraordinary but previously untested quantum electrodynamics phenomenon that has recently been demonstrated in graphene is the Klein paradox, which predicts that relativistic charged particles can tunnel without hindrance through any energy barrier, high or wide. This happens as such particles generate a "ghost-version" of their corresponding antiparticle that has opposite charge. The antiparticle does not see the barrier and so passes through it, creating the normal particle again on the other side.

In the case of graphene, the quasiparticles and their holes are the equivalent to pairs of particles and antiparticles such as the electrons and positrons studied in particle-physics experiments.
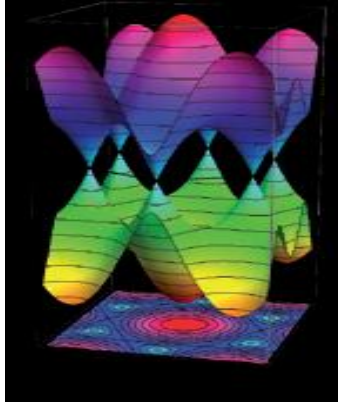
Figure 1. A visual representation of the unusual energy/momentum relationship of the charge carriers in graphene, which gives rise to its unusual quantum behavior.

3.3 Several Core Chip

3.3.1 Need

Graphene The evolutionary approach is to continue today's trend with a few large processor cores, each employing 20 to 100 million logic transistors, and a large shared cache. Performance increase is with accordance to the Pollack's rule, which states that performance increase is approximately proportional to the square root of complexity i.e. if you the logic is doubled in a processor core, then it delivers only 40% more performance.

Two smaller processor cores, instead of a large monolithic processor core, can potentially provide 70-80% more performance, as compared to only 40% from a large single core. Multicore processors have a several benefits as each processor core may posses its own optimized supply voltage and frequency, power consumption is reduced as each processor core can be individually turned on or off, easier to manage heat sink across the die and reliability and leakage current is improved.

If the technology gets to scaled further transistor performance may not get suitable and required increment due to subthreshold and leakage current resulting in unreasonable power consumption, here it's needed to go beyond the present multicore system

3.3.2 Several core chip Working

Even though a small core has lower performance than a large complex core, if a lot of small cores are integrated then the total throughput of the system is increased, performance of a smaller core reduces as square-root of the size, but power reduction is linear, resulting in smaller performance degradation with much larger power reduction and the total throughput is increased.

3.3.3 Design Consideration and Performance analysis

In present scenario, designs have to operate at the highest possible frequency to deliver high performance, but not anymore. Hence Instead of domino logic simple and robust static CMOS logic, which consumes much lower power, must be used. This'll reduce the total power consumption.

The main limitation of a several core chip is parallel speeding for fewer applications. This limitation is true if only one application is running on the system at an instant, and parallelization of a single application across all the cores is done. If an application is parallelized to both less core system and a several core system then the performance of less core system is higher but as the applications used are increased, the performance of several core systems is increased. Practically there are multiple applications running, each with multiple tasks and multiple threads, and thus there exists opportunity to harvest the performance of a several core system.

4. Current Technological Modifications

The revolution of the embedded systems and almost all the technologies is powered by ever-faster systems demanding high performance of processor. To keep up with this demand we cannot rely entirely on traditional approaches to processor. Therefore a lot of new technological aspects are introduced and will keep on introduced. We discuss about how some current technological features might be modified in future:

4.1 22-10-7-5 Lithography Technology Continuation

Current industry is dealing with 22nm lithography technology node with a further expected reach to 10nm technology soon. The main hurdle to this is control on the leakage current as the component size decreases, the gaps between components get smaller which may be responsible for numerous issues. At the same time high drive current is a necessity to enhance the performance. According to ITRS and research and development pipeline espoused, 2015 will see the production of semiconductor components based on a 10nm process size, quickly followed by 7nm and 5nm.

4.2 Power Management

For the power management the recent technology feature Turbo Boost has been proved very efficient which automatically allows processor cores to run faster than the base operating frequency if the processor is operating below rated power, temperature, and current specification limits. Turbo Boost technology can be engaged with any number of cores or logical processors enabled and active, resulting the increased performance in single and multi-threaded loads.

This technology feature may also bring with Demand Based Switching which's going to launch in Core i3 and Core i7 processors soon. In DBS applied voltage and clock speed of a microprocessor are kept at the minimum necessary levels for optimal performance of required operations. The concept behind DBS is to monitor the processor's use by application-level workloads. A further enhanced feature for the power management is Enhanced SpeedStep technology which may also capture the market in following years.

4.2 Hetrogeneous Parellelism

General purpose programming models available today can yield high performance but are too low level to be accessible for an average programmer. Therefore, a heterogeneous architecture consisting SIMD core, threaded core, specialized core with domain embedding language may result into various scientific engineering, virtual worlds and personal robotics applications. This enables to very high architecture maturity and programming accessibility and may capture the entire market till 2020 as predicted by Microsoft.

4.3 Hyper Threading

Hyper-Threading Technology makes a single physical processor appear as multiple logical processors. . In this, there is one copy of the architecture state for each logical processor, and the logical processors share a single set of physical execution resources.

To take advantage of hyper-threading performance, serial execution can not be used. In this threads are non-deterministic and involve extra design and the threads have increased overhead. More recently Hyper-Threading has been criticized as being energy inefficient. Specialist low power CPU design company ARM has stated SMT can use up to 46% more power than dual CPU designs. These issues may also result in switching to a new technological aspect.

Conclusion

As microprocessors have evolved from simple single-issue architectures to the more complex multiple-issue architectures, many more resources have become available to the microprocessor with a lot of new features. In these architectures and technology innovation a lot of new perspective are yet to be added and many issues to be resolved.

References

[1] G Intel White Paper "Built-in Protection in Laptop PCs Improves Compliance with New Healthcare Rules"

[2] Intel Architectural Server White Paper "Power Management in Intel® Architecture Servers"

[3] Mark Bohr and Kaizad Mistry Intel White Paper "Intel's Revolutionary 22 nm Transistor Technology"

[4] Intel Technology Journal,Volume 06,Issue 01,Published February 14, 2002,ISSN 1535766X "Hyper-Threading Technology"

[5] Millind Mittal, Alex Peleg and Uri Weiser Intel White Paper "MMX™ Technology Architecture Overview"

[6] Intel White Paper November 2008 "Intel® Turbo Boost Technology in Intel® Core™ Microarchitecture (Nehalem) Based Processors"

[7] ISA White Paper"Fifty year of microprocessor technology advancements: 1965 to 2015"

[8] IOP "Graphene: A new form of carbon with scientific impact and technological promise"

[9] Shekhar Borkar Intel Corp. Microprocessor Technology Lab "Thousand Core Chips—A Technology Perspective"
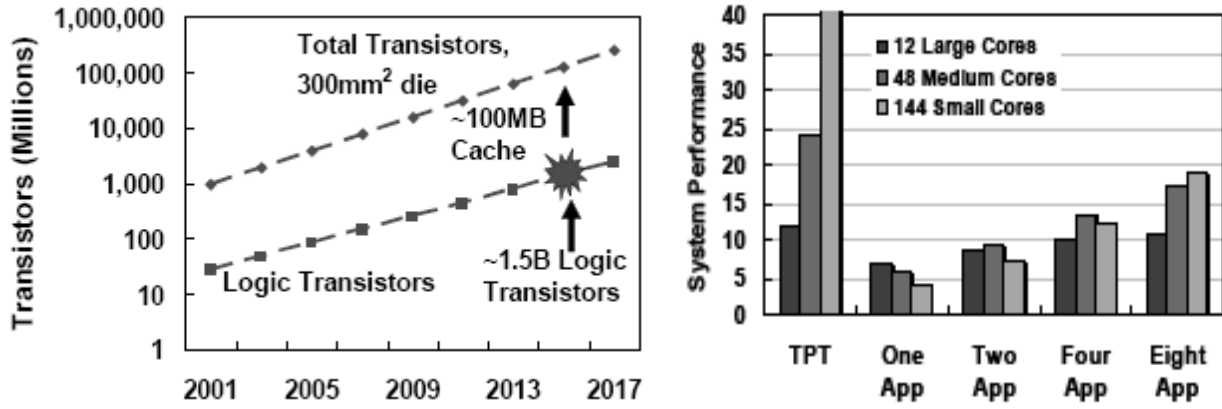
Figure 2.   Transistor Capacity Analysis and Performance of different cores