Contemporary Alternatives to Traditional Processor Design in the Post Moore's Law Era

Andy Kuszyk
Manchester Metropolitan University
Manchester, UK
andy.kuszyk@stu.mmu.ac.uk

Mohammad Hammoudeh
Manchester Metropolitan University
Manchester, UK
m.hammoudeh@stu.mmu.ac.uk

ABSTRACT

Over the last forty years, Moore's Law has held as a general rule of thumb for the progress made in the Central Processing Unit (CPU) industry. This law has broken down over the last decade as the design of processors using traditional techniques has begun to approach physical limitations. However, despite this set back in the advancement of traditional processing technologies, alternatives have started to present themselves from all corners of industry and academia. Many of these are radical changes to the way processors have been designed and deployed in the past. This paper reviews three promising and contemporary approaches at continuing to increase the performance of tomorrow's CPUs, despite the physical limitations constraining their design today.

CCS CONCEPTS

• **Hardware** → *Memory and dense storage*;

KEYWORDS

Moore's law, central processing unit, domain specific hardware, quantum computing, 3D chips.

ACM Reference Format:
Andy Kuszyk and Mohammad Hammoudeh. 2018. Contemporary Alternatives to Traditional Processor Design in the Post Moore's Law Era. In *ICFNDS'18: International Conference on Future Networks and Distributed Systems, June 26–27, 2018, Amman, Jordan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3231053.3231099

1 INTRODUCTION

Moore's Law [11, 12] predicted an exponential increase in computing power throughout the final quarter of the twentieth century and on into the twenty first century. Whilst this was both an exciting and improbable prospect in the long term, industry kept pace remarkably well until the 1990s [14]. It has now become apparent that we are approaching the physical limitations of Complementary Metal-Oxide-Semiconductor (CMOS) devices [14] and that the industry must look to new technologies and approaches to maintain its momentum in increasing the computational power of new devices. These physical limitations are a result of transistors approaching the nanometre scale and therefore being subject to thermodynamic and electromagnetic noise [2].

Although Moore's Law did prove sustainable for almost three decades in the previous century, there were significant challenges in doing so both physically and commercially. These challenges were overcome, but should also be important metrics for any emerging technologies and new approaches, because if they do not overcome the same problems then they may not prove robust in the long term or commercially viable. Among the challenges faced by the industry were those of:

(i) power consumption[2]; (ii) manufacturing cost[14]; (iii) large-scale adoption; (iv) versatility at a wide range of tasks.

Over the last two decades, the demands placed on the CPU by the technology industries have grown enormously, with the corresponding increase in processing power barely keeping up. Whereas, at the end of the 1990s, we had relatively low performance domestic and commercial workstations with a low requirement for distributed computing power via the Internet, we are now facing a very different landscape. Today, high performance mobile devices, capable of complex computational tasks such as facial recognition and real-time video processing, are ubiquitous. Further more, the Internet age has given rise to increasingly large volumes of data being collected by internet giants and start-ups alike, with the key to commercial success lying in analysing and understanding this veritable mountain of data. Machine learning applications are also becoming ubiquitous, with wide-spread use of deep neural networks to understand and interpret these large data sets consuming vast amounts of computation. Furthermore, these services and capabilities are being offered to everyone via cloud computing platforms, so that yesterday's high-end applications are becoming today's basic features. Supplying faster and faster processors at scale is a fundamental requirement for the growth of cloud computing, with the commodification of hardware virtualisation being supported by an increasing number of data centers. Packing more computing power into each server rack in these data centers is the key to providing hardware as a service solutions for tomorrow's software companies in a sustainable and profitable way. The question remains, however: where does the next advancement come from to fulfil this future potential? This amazing growth in the demand for ever more powerful computational resources leads us to a place today where the industry is almost waiting for a paradigm shift to accelerate us on our way in the information revolution that we have surely only just begun.

This paper presents some of the most promising innovations that have come out of the research community over the last five years, which represent a spectrum of different approaches to keeping up with an ever present demand from businesses. Following this

ICFNDS'18, June 26–27, 2018, Amman, Jordan

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/3231053.3231099

introduction, Section 2 overviews three different alternatives to traditional CPU design, namely: (i) 3D chip design; (ii) Quantum computing; (iii) Domain specific hardware. Each of these emerging fields represents a novel approach to improving the computing power of the next generation of machines, without necessarily following the industry's previous vector of CMOS device miniaturisation Section 3 summarises these approaches, comparing each technique against the others. Finally, Section 4 outlines future opportunities in the processing industry along with concluding remarks about the future of CPUs.

2 ALTERNATIVES TO THE TRADITIONAL CPU

2.1 3D Chip Design

Traditional processing chips consist of a single layer of active components mounted on a 'die' which connects the unit to other components. A 3D chip is an evolution of this approach, whereby multiple layers of active components are stacked (most commonly vertically) together into a device consisting of multiple dies [9]. Stacking components vertically has a number of advantages such as lower overall power usage and reduced latency between processor and memory, as well as the obvious benefits of packing more processing power per device [9]. This technique has been advocated by both academics and members of industry for more than a decade [9], but has faced significant challenges around manufacturing processes and automated testing that are specific to the more complicated process of introducing multiple layers of components into a single device.

Building multiple processing layers into one CPU can be achieved in a number of different configurations, and often memory layers are included alongside processing layers [5]. This can result in major performance improvements, because a significant limiting factor in today's processors is the so-called von Neumann bottleneck [5], which is a result of processors' reliance on fast access to rapid memory caches. A limitation on the computational power provided by a processing chip is often the rate at which it can read data from memory and this is the 'bottleneck' in question. This limitation can be overcome, at least in part, with 3D chips by interleaving memory layers with processing layers, so that memory and processor are both active in the same device. Figure 1 illustrates some of the many different architectures for 3D chip design and also demonstrates that this field is already quite advanced, with many of these techniques being manufactured commercially [9].

3D chips have actually been in commercial production for some time in less complex applications, such as solid state hard drives and DRAM [5]. As a result, they have potential for improving the performance of commercially available CPUs, because many manufacturing hurdles have already been overcome. Indeed, 3D chips have already been successfully trialled for industrial processing applications. Even ten years ago, early trials indicated that a 3D chip design could outperform a comparable, traditional chip by 14%, whilst using 55% less power [8]. Furthermore, large cloud computing consortiums, such as the EuroCloud project, have actively trialled 3D chips as part of an attempt to reduce running costs and, importantly, energy consumption [7].

However, the manufacturing process for 3D CPU chips is more complicated than that employed for other devices. The production of traditional 2D chips is accompanied by a sophisticated automated testing process, to ensure that the resultant devices are compliant with their original design parameters. Given the low fault tolerance of digital systems, this automated testing process is essential in the large-scale production of processing chips. This testing process is considerably more complicated with 3D chips, whilst being no less essential. Additional complexity comes from the fact that circuits buried deep in a device are often more difficult to test than those that are easily accessible [9]. In a 3D chip, with multiple, stacked layers of chips, the interactions between the devices complicate the process of testing individual circuits in each of the active components. Whilst these challenges are yet to be fully overcome and do threaten 3D chips with higher manufacturing costs than their 2D counterparts, the topic of testing 3D chips is an area of on-going research [9]. As a result, there is still optimism that 3D chips can become a mainstream, mass-produced alternative to traditional designs.

2.2 Quantum Computing

Quantum Computers have been an exciting new area of computing research for several decades and are a result of the realisation that useful information processing can be achieved using the outcomes of Quantum Mechanics, namely that quantum particles (atoms, ions and electrons [1]) can exhibit all possible states available to them prior to observation [10]. Furthermore, it is understood that a so-called Quantum Computer would be capable of storing, transmitting and processing large amounts of information in parallel, although probably not in a way that we would currently recognize and possibly not for applications that are undertaken at the moment. Indeed, it is not even understood exactly what this kind of computer would look like and what sort of problems it would be applicable too [10]. Answers to these questions are the subject of ongoing research, and have been for many years.

Whereas a classical computing system operates on binary states, 0 or 1, a quantum system would operate on both 0 and 1 and a superposition of both states, consisting of the probability of being in either 0 or 1 as well [1]. This vastly increases the number of states the computer can process at a given time, although does result in a small proportion of errors that require additional processing to remove [1]. There are enumerable different approaches towards constructing a functioning Quantum Computer [1], although one recent paper [3] succinctly illustrates a straightforward architecture (see Figure 2) involving isolated quantum processors whose communication is mediated by microwaves and who, collectively, form a computational device capable of electrical inputs and outputs. It's worth noting that this entire enclosure would still need to be refrigerated at cryogenic temperatures [1], which does not bode well for wide-spread adoption of this technology, or for low power usage.

Although this technology is not ready for commercial applications at present, it remains an exciting and busy area of research with hopes that commercially viable products may enter the market in five to ten years [1]. Many successful and high profile trials have taken place, including D-Wave and IBM's attempts [13], although

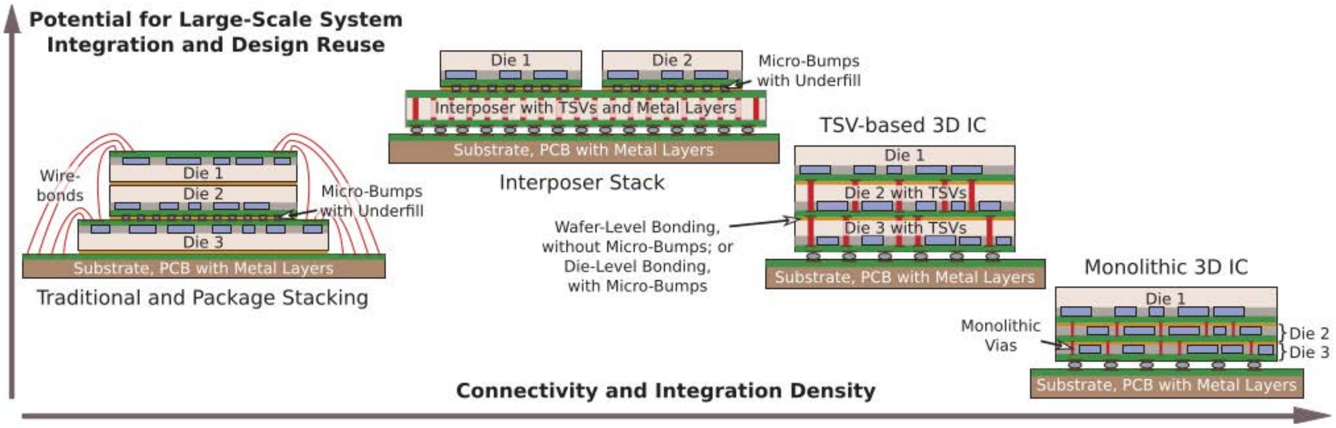
Potential for Large-Scale System Integration and Design Reuse
Die 1
Die 2
Micro-Bumps with Underfill
Interposer with TSVs and Metal Layers
Substrate, PCB with Metal Layers
Interposer Stack
TSV-based 3D IC
Die 1
Die 2 with TSVs
Die 3 with TSVs
Substrate, PCB with Metal Layers
Wafer-Level Bonding, without Micro-Bumps; or Die-Level Bonding, with Micro-Bumps
Wire-bonds
Die 1
Die 2
Die 3
Micro-Bumps with Underfill
Substrate, PCB with Metal Layers
Traditional and Package Stacking
Monolithic 3D IC
Die 1
Monolithic Vias
Die 2
Die 3
Substrate, PCB with Metal Layers
Connectivity and Integration Density

Figure 1: An illustration of different 3D chip implementation architectures, taken from [9].

these have still yet to prove that Quantum Computers are ready for wide-scale adoption.

One of the main challenges that Quantum Computers will face in practice is that they will, at least to begin with, be highly specialised for specific tasks and will not have a general instruction set [1]. An example of this is the factorization of an integer into its prime factors, as suggested in Shor's Algorithm [1]. Whilst this may have very specific uses (breaking RSA encryption for example!), it does mean that they are likely to be integrated into existing super-computer architectures, rather than becoming the ubiquitous computer devices we are familiar with today. Indeed, it is clear that whilst we may see the introduction of powerful Quantum Computers in the coming years, these will be neither cheap to run (with high design, manufacturing and operating costs) or general purpose. Quantum Computers are a tangential change in the way we construct computing devices, but they will not help us to overcome the limitations of Moore's Law in the short or medium term.

2.3 Domain Specific Hardware

The challenges presented by expanding software demands and slowing processor development is especially prevalent in the field of machine learning, where a recent increase in the use of Deep Neural Networks (DNNs) to analyse large and complex datasets has resulted in a huge increase in the requirement for computing power. In 2006, Google concluded that no additional requirement for computing resources existed in order to service the increasing demand of machine learning applications (such as voice, speech and text recognition) and that the increasing demand from these services could easily be supported by their spare processing capacity [6]. However, in 2013 this conclusion was reversed with the realisation that double the number of data centers would be required if every mobile user performed just 3 minutes of voice searching per day [6]. Clearly this presented an untenable solution, but no obvious alternative existed in the industry to exceed Moore's Law and provide additional processing power so rapidly. To this end, Google embarked on a 15 month project to create a domain specific processing unit that satisfied the demands of DNNs in particular and that was optimised for this use case. The result was the Tensorflow Processing Unit (TPU), so named for its support for Tensorflow, an open source DNN software package written by Google [6].

As a result of the TPU's optimisation for the specific computational task required for the computation of DNNs (matrix multiplication), it was able to run Tensorflow programs approximately 15 times faster than on a server level GPU [6]. In addition to this increase in speed, the deployment of the TPU (which is now a widespread component in Google's data centers, and a feature in their commercial cloud platform) has also seen correspondingly significant reductions in power usage. The TPU itself is a custom processing unit mounted on a PCI card (see Figure 3) that can be added to existing commodity hardware easily and cheaply. This generic design means that the TPU can be retro-fitted into existing data center hardware as well as being easily added to off-the-shelf server hardware.

Google are not alone in pursuing domain specific hardware solutions to the computational challenges posed by ever more complex DNNs. The Integrated Systems Laboratory in Zurich developed an embedded processing unit capable of supporting Convolutional Neural Networks (a flavour of DNNs, commonly used in image processing) in 2015 that showed major improvements over traditional CPU and GPU hardware [4].

Given the huge improvements in performance that Google realised by the introduction of the TPU as well as the relatively short time over which they developed and introduced it, it is clear that domain specific processing units could provide a veritable gold mine of un-tapped computational potential. However, the use case for these domain specific processors are necessarily niche, since their huge performance gains are a result of their customisation towards a specific use case. Domain specific hardware has already, and will continue to, meet the demand of tomorrow's high performance applications, but it does not provide a general purpose solution to the relative decline in CPU advancements.

3 SUMMARY AND COMPARISON

Each of the approaches outlined above posed a different kind of solution to the demise of Moore's Law. These approaches fall into

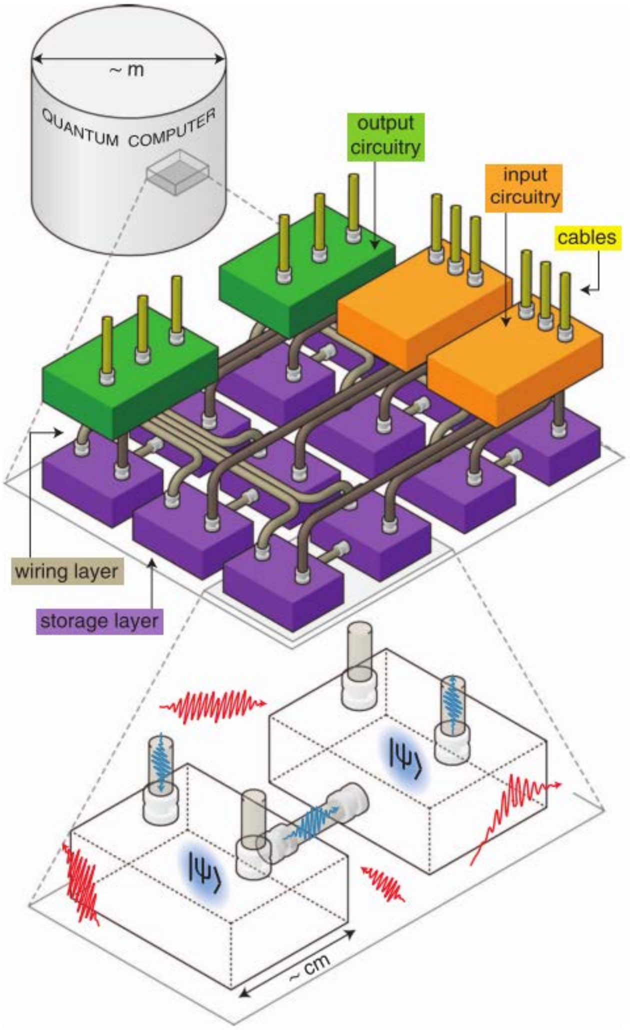

Figure 2: A possible design of a quantum computer, making use of multiple, isolated quantum computing units[3].

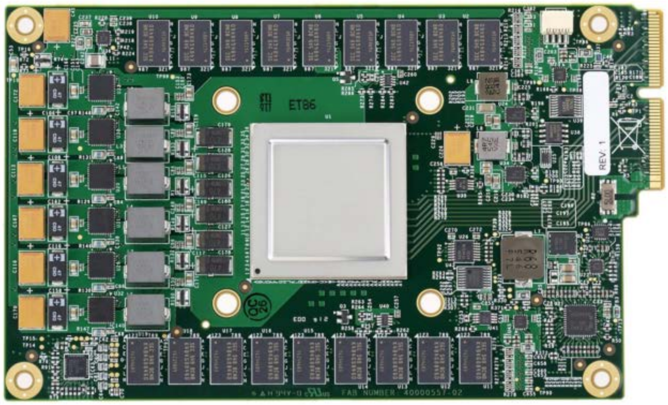

Figure 3: The Tensorflow Processing Unit, a custom processing unit designed for running Deep Neural Networks[6].

two main categories: (i) General purpose processors, capable of replacing today's ubiquitous computing devices; (ii) Use-case specific processors optimised for a particular problem or application.

3D chips have the potential to yield general purpose solutions that could be used in commodity hardware in both the domestic and commercial sectors. They are an evolution of our current approach to building general purpose processing units and, as such, share many advantages with current approaches (such as established manufacturing processes and well established testing paradigms). However, complications with mass production and testing of 3D chips is currently a barrier to their widespread adoption, although it seems likely that these challenges will be overcome in the next few years.

Unlike 3D chips, Quantum Computing and Domain Specific Hardware provide highly specialised solutions to the problem of diminishing processing power. Due to the high demands of specific applications, such as cryptography and machine learning, these approaches have been developed to provide improvements in the short-term to the ever increasing demands of the internet age. That being said, their applications are restricted to the particular problem domain they have been developed for and they are unlikely to yield general purpose domestic or commercial improvements to processing power. Quantum Computing in particular is probably years away from a viable commercial solution, whereas Domain Specific Hardware is already proving promising for machine learning applications.

Table 1 summarises the main characteristics of each of these approaches and compares them to one another qualitatively, outlining their benefits and limitations.

4 CONCLUSIONS

This paper has reviewed three contemporary approaches to alternative CPU techniques. 3D Chips are an evolution of traditional approaches that are promising for both their short and medium term benefits, as well as their suitability for wide scale production and adoption. Quantum Computers are an entirely different approach to the field which may yield a step change in computational devices in the long term. However, research into their construction and application continues and their viability as wide scale alternatives to traditional techniques in the short term is unlikely. Domain specific CPUs present an exciting avenue for short term advances, although they are, by their very nature, highly specialised and unsuitable for wide spread use in commodity or domestic hardware.

Whether it be super computers, commodity hardware or specialised servers, the likelihood is that all three approaches will have significant benefits over the coming decade, although 3D chips will probably provide the most significant impact on the every day computing devices we increasingly depend upon.

REFERENCES

[1] C. G. Almudever, L. Lai, X Fu, N Khammassi, I Ashraf, D. Iorga, S. Varsamopoulos, C. Eichler, A. Wallraff, L. Geck, A. Kruth, J. Knoche, H. Bluhm, and K. Bertels. 2017. The Engineering Challenges in Quantum Computing. *2017 Design, Automation and Test in Europe (DATE)* (2017), 836–845. https://doi.org/10.23919/DATE.2017.7927104

[2] John Augustine, Krishna Palem, and Parishkrati. 2017. Sustaining Moore's Law Through Inexactness. (2017), 1–11.

[3] T. Brecht, W. Pfaff, C. Wang, Y. Chu, L. Frunzio, M. H. Devoret, and R. J. Schoelkopf. 2015. Multilayer microwave integrated quantum circuits for scalable quantum

Table 1: Summary of the different approaches to alternative processor design.

Technique	Category	Previous Work	Target Market	Benefits	Limitations
3D Chips	General Purpose	Debenedictis, et al. [5]; Knechtel, et al. [9].	Mass-market domestic; commercial; servers.	General purpose enough to replace traditional CPUs; existing manufacturing processes for simple devices.	Complex testing process holding back mass-market manufacturing.
Quantum Computing	Domain Specific	Almudever, et al. [1]; Singh and Singh [13]; Brecht, et al. [3].	Super-computer; scientific.	Step-change in processing power, capable of solving some of today's most intractable problems.	Highly specialised, unlikely to be a generalised computational device in near future; high manufacturing and running costs.
Domain Specific Hardware	Domain Specific	Jouppi, et al. [6]; Cavigelli, et al. [4].	High performance servers; data centers.	Already in use; demonstrable performance improvements; benefits can be realised today.	Implicitly can not provide general purpose CPUs; only solves niche problems.

computing. *Nature Publishing Group* (2015). https://doi.org/10.1038/npjqi.2016.2 arXiv:1509.01127

[4] Lukas Cavigelli, David Gschwend, Christoph Mayer, Samuel Willi, Beat Muheim, and Luca Benini. 2015. Origami. *Proceedings of the 25th edition on Great Lakes Symposium on VLSI - GLSVLSI '15* (2015), 199–204. https://doi.org/10.1145/2742060.2743766 arXiv:1512.04295

[5] Erik P. Debenedictis, Mustafa Badaroglu, An Chen, Thomas M. Conte, and Paolo Gargini. 2017. Sustaining Moore's Law with 3D Chips. *Computer* 50, 8 (2017), 69–73. https://doi.org/10.1109/MC.2017.3001236

[6] Norman P. Jouppi et al. 2017. In-Datacenter Performance Analysis of a Tensor Processing Unit. *Proceedings of the 44th Annual International Symposium on Computer Architecture - ISCA '17* (2017), 1–12. https://doi.org/10.1145/3079856.3080246 arXiv:1704.04760

[7] Sachin Idgunji, Ali Saidi, Yiannakis Sazeides, Bushra Ahsan, Nikolas Ladas, Chrysostomos Nicopoulos, Isidoros Sideris, Almutaz Adileh, Michael Ferdman, Pejman Lotfi-kamran, Pol Marchal, and Nikolas Minas. 2010. EuroCloud: Energy-conscious 3D Server-on-Chip for Green Cloud Services. (2010), 3–4.

[8] Taeho Kgil, Shaun D'Souza, Ali Saidi, Nathan Binkert, Ronald Dreslinski, Steven Reinhardt, Krisztian Flautner, and Trevor Mudge. 2006. PicoServer: Using 3D Stacking Technology To Enable A Compact Energy Efficient Chip Multiprocessor. *Asplos'06* 34, 5 (2006), 117–128. https://doi.org/10.1145/1168919.1168873

[9] Johann Knechtel, Ozgur Sinanoglu, Ibrahim (Abe) M. Elfadel, Jens Lienig, and Cliff C. N. Sze. 2017. Large-Scale 3D Chips: Challenges and Solutions for Design Automation, Testing, and Trustworthy Integration. *IPSJ Transactions on System LSI Design Methodology* 10 (2017), 45–62.

[10] Thaddeus D. Ladd, Fedor Jelezko, Raymond Laflamme, Yasunobu Nakamura, Christopher Monroe, and Jeremy L. O'Brien. 2010. Quantum Computing. (2010). https://doi.org/10.1038/nature08812 arXiv:1009.2267

[11] Gordon E Moore. 1965. Cramming more components onto integrated circuits. *Electronics* 38, 8 (1965), 114–117. https://doi.org/10.1109/jproc.1998.658762

[12] Gordon E Moore. 1975. Progress in Digital Integrated Electronics. (1975).

[13] Jasmeet Singh and Mohit Singh. 2017. Evolution in Quantum Computing. *Proceedings of the 5th International Conference on System Modeling and Advancement in Research Trends, SMART 2016* (2017), 267–270. https://doi.org/10.1109/SYSMART.2016.7894533

[14] Thomas N. Theis and H.-S Philip Wong. 2016. The End of Moore's Law: A New Beginning for Information Technology. June 2016 (2016), 41–50.