Abstract—This article discusses the role of energy-efficient cloud server-on-chip solutions to reduce the total cost of ownership and the ecological impact of cloud computing data centers. A solution is envisaged exploiting a many core platform based on the new ARMv8 family, which extends the energy efficiency of the classic 32-b RISC ARM architecture to the server computing domain through a 64-b instruction set. A service-oriented Network on Chip (NoC) communication infrastructure is used to avoid inter-core communication bottlenecks. Implementation/performance figures on nano-scale CMOS technology are also provided.

Keywords: Cloud Computing; Network-on-Chip (NoC); Cloud Server-on-Chip; Many Cores; Energy and Cost Sustainability

I. INTRODUCTION

Cloud computing has rapidly emerged as a new computing paradigm representing a fundamental shift in delivering information technology services via on-demand resources. Cloud computing is where data, software (SW) applications, or computer processing power are accessed from the cloud of online resources. This permits (i) individual users to access their data and applications from any edge device, (ii) any organization to reduce ICT capital costs by purchasing hardware (HW) and SW as utility services. The fundamental unit of a cloud computing infrastructure is a server, either physical or virtual. Physical servers are discrete individual computers on which online applications can be run and data can be stored. In contrast, virtual servers allow many users to share the processing power of one physical server. Servers are individual circuit boards, known as blades, mounted within equipment racks in a data center. Most recent server generations, such as Xeon5E, integrating multiple high-end computing cores and running at several GHz, have a power consumption of hundreds of Watts per server [1]. Considering that a server farm can require hundreds of thousands servers (300K servers in a Microsoft server farm [1]), the energy cost is no more sustainable. Such high energy overhead entails also extra thermal cooling systems; therefore energy and thermal issues determine more than 30% of the total cost of ownership of cloud data center, see Fig. 1 (server amortization plan in 3 years). This is a big issue since one of the driving factors of the success of the cloud concept is a new business model based on the elimination of the up-front capital and operational expenses [2]. However, the total cost of ownership is still an obstacle for having billions of users accessing millions of services. Since a significant % of data center’s cost of ownership comes from the recurring energy costs, a new generation of energy-efficient cloud server architectures is needed.

Fig. 1: Monthly cost of cloud infrastructure dominated by power

An analysis of the US Environmental Protection Agency (EPA) states that powering the nation’s data centers, consisting of vast server grids, required roughly 100 billion kWh hours in 2011 with a cost of 7.5 billions of US$. In September 2011, Google announced that its global operations continuously draw 260 million MW of power, a quarter of the energy generated by a nuclear power plant. Moreover, as the level of carbon emissions in ICT technologies is estimated to triple by 2020, policy makers are forced to prepare legislation incentives towards green computing. The Smart 2020 report estimates the potential impact of ICT-enabled solutions to be as much as 18 percent of total global carbon emissions.

Towards the implementation of energy-efficient cloud server-on-chip, a quad-core ARM Cortex-A9 architecture has been recently proposed [8]: relying on the energy efficiency of the ARM Cortex-A9 RISC (Reduced Instruction Set Computer) architecture, cloud servers have been realized using boards, each hosting 4 of these multi-core chips with a power consumption of 20 W per board, 5 W per quad-core chip. However, the achieved performance is still not comparable to that of servers using high-end computing processors since the ARM Cortex-A9 core is limited to a 32-b architecture. To overcome the above issues a novel solution is envisaged in this paper exploiting a many core platform based on the new ARMv8 family, which extends the energy efficiency of the classic 32-b RISC ARM architecture to the server computing domain through a 64-b instruction set. A solution scalable in terms of number of cores, and hence in terms of the trade-off between power consumption and computational capabilities, is proposed. Since the on-chip communication infrastructure is the real bottleneck in computing systems with a
high number of cores, then a service-oriented Network on Chip (NoC) solution is adopted.

II. MANY-CORE PLATFORMS FOR CLOUD

A huge opportunity for power reduction is due to the fact that most server processors in use today, based on CISC (Complex Instruction Set Computer) architecture, draws about 150-200 W under normal operations and tens of W in idle mode [1,6,9]. For instance, the last generation of Itanium processor [9] for server applications, realized in 32 nm CMOS technology, requires 170 W, 540 mm² of area and 3.1 billions of transistors for its 8 64-b cores with total 54 MB of on-die cache. Much higher energy efficiency comes from RISC cores, adopted in embedded systems and mobile computing platforms, although classic 32-b devices as the ARM Cortex are not optimized for data storage and computing intensive server market. The use of ASIC (Application Specific Integrated Circuit) is envisaged in [10] to achieve ultra low-power cloud computing but in such case there will be a dramatic reduction of system flexibility. The proposed solution achieves the right trade-off between performance, flexibility and energy/cost sustainability by adopting a multi-core platform, based on energy efficient ARM-based RISC cores, with 64-b instruction set extension.

In 2005, processor market hits a wall, see Fig. 2. As described in [11] “the single-threaded free lunch was over” with a fail of major PC manufacturers (Intel, AMD, IBM) with most of their traditional approaches to boosting CPU performance. Thus key semiconductor challenge has moved from only performance to energy efficiency (measured in performance per joule) and miniaturization. In order to sustain the Moore’s law, the 2005 was the starting year to move the industry toward an architectural solution called multi-core, integrating multiple cores on a single die.

![Fig. 2: Evolution of CPUs for cloud servers](image)

During all these years we have seen 3 majors multi-core computing platform styles [12]. The first one, is a symmetric many-core processor that simply replicates a state-of-art superscalar processor on a die. The second one, is a symmetric many-core processor that replicates a smaller and more power-efficient core on the same die. The third multi-core computing style is a heterogeneous computing platform containing distinct classes of processing cores on the same die. An example of the latter style is represented by a state-of-the-art host processor integrated with an array of parallel processing elements for accelerating certain parts of an application. Heterogeneous computing has emerged to address the growing concerns of energy efficiency and silicon area effectiveness. The 2011 was a special year for small-scale heterogeneous multi-core computing platforms with the arrival of multi-core tablets (e.g., iPad2, Playbook, Kindle Fire, Nook Tablet) and smartphones (e.g., Galaxy S II, Droid X2, iPhone 4S). On the server computing side the trend is the transition to high performance multi-core platforms, providing an array of register sets, shared cache and multiple threads of execution, leading to high bandwidth and often real-time processing. Existing high end server architectures are IBM’s Power7, Oracle Sun PowerNap, AMD’s “Magny Cours” Opteron and Intel Ivy Bridge. The high computational power, Tera floating-point operations per seconds (TFLOPs), of the above platforms are paid with high energy cost, and related thermal issues reducing system reliability and creating cooling overheats in terms of cost and size. Solving the energy-efficiency problem is a key issue to reduce cost, ambient and energy impact of the cloud computing paradigm. Moreover energy-efficiency will reduce thermal problems thus increasing life time and reliability of the cloud computing infrastructures. To be noted that cloud computing will be not devoted solely to multimedia web-service for consumer/entertainment applications, but also to services to increase efficiency of the public administration or of private companies. Hence reliability, low risk of service denial, long life time and reduced energy cost are key factors for the widespread success of cloud computing.

III. PROPOSED CLOUD SERVER-ON-CHIP

a. Selection of the processing core

In this section a novel cloud computing infrastructure is introduced that considers energy awareness as key optimization factor. The proposed approach is alternative to the parallel effort carried out by Intel [13] that, in order to increase the energy carried by cloud servers, proposes the SCC (Single-chip Cloud Computer): a 2D array of 48 Pentium P54C cores each with 16 KB of L1 data cache and 16 KB of L1 instruction cache, organized as 24 dual-core tiles with 2x256 KB of L2 cache, interconnect by a 6x4 2D mesh on-chip network. The area and power costs of the Intel's SCC are 567 mm² and up to 125 W [13] respectively.

Unlike Intel's SCC our proposal for a green cloud computing platform is based on a new ARM architecture, called ARMv8 family, with the addition vs. a classic RISC-based architecture of a new "A64" 64-b instruction set. This instruction set extension will enable the use of ARM RISC cores into the server and enterprise computing space. In this context, today, there are a lot of headlines about the HP announcement to build servers with ARM processors. Since the proposed platform is based on ARM the amount of mobile and consumer legacy SW that potentially can run on cloud computing is huge. Through the use of a 64-b extended instruction set, more suited for cloud server applications than state-of-art cores of the 32-b ARMv7 family, we expect a performance gain of a factor of least 2.

Cloud computing based on ARM cores, could enable vast energy savings, since new ARM chips and platform architectures offer speed, low power consumption and require a reduced amount of cooling. Recently, Microsoft announced that
Windows 8 will support ARM architecture. Moreover, ARM 32-b RISC processors such as the multi-core Cortex-A9 can already compete with Intel Atom or Xeon processors: experiments carried out on video applications prove that in terms of computational capability the Cortex-A9 MPCore, with up to 4 processors at 500 MHz integrated on-chip, can be compared with a 1.6 GHz Intel Atom [1].

Experiments carried out with Apache 2.2 HTTP server applications prove that a dual-core Cortex-A9 processor at 1 GHz has an energy efficiency one order of magnitude higher than a quad core Intel Xeon E5430 at 2.66 GHz and even much higher than a Pentium 4 at 2.8 GHz. Energy efficiency in [1] has been measured as the total number of requests served per Joule consumed by the Apache 2.2 HTTP server running on different platforms: the Xeon E5430 was able to serve 33000 requests/s but with an energy efficiency of 413 requests/J, the Pentium 4 was able to serve 7100 requests/s with an efficiency as low as 413 requests/J while the dual Core Cortex-A9 was able to serve 4600 requests/s with an efficiency of 4600 requests/J. Furthermore the new ARMv8 architecture features the "A64": a 64-b instruction set, which will enable to use ARM cores into the server and enterprise computing space [14,15]. ARMv8 will succeed ARMv7, which is the foundation for current processors such as the Cortex-A9 and Cortex-A15. ARMv8 has also floating point processing capabilities and it represents a superset of ARMv7. All ARMv8 chips will run legacy 32-b ARMv7 code in the "AArch32" execution state, while 64-b code will be run in the "AArch64" state. By widening the integer registers in its register file to 64 bits, ARMv8 can store and operate on memory addresses in the range of millions of terabytes. This is quite interesting for server applications such as cloud computing.

b. Target technology

As far as the target implementation technology is concerned, STMicroelectronics is developing and producing high-end CMOS processors and custom System-on-chip. The silicon technology process of STMicroelectronics has a track record in the low power category of core CMOS, especially for mobile/consumer markets. These technologies can address high frequency applications as well, currently higher than 2.5 GHz in 28 nm CMOS. Concurrently STMicroelectronics is developing technologies on FDSOI (Full-Depleted Silicon-on-Insulator) [16], which are able to address both the high-performance end of the spectrum (3 GHz and more) and ultra low-power operation thanks to an improved electrostatic control on the transistor channel. The challenge is to dynamically adapt the power consumption to the required performance, thus wasting the minimum amount of energy. To this aim a vertical integration from technology to SW level virtualization layer and virtual machine management would be implemented.

Our estimation is that targeting already available 32 nm submicron CMOS technologies the basic cluster for cloud computing can be composed of at least 4 64-b ARM cores integrated on a single chip. For a server on chip application we aim at implementing a heterogeneous architecture with a Network-on-Chip connecting several heterogeneous clusters: ARM cores, IO and GPU (graphics processing unit). The ARM clusters could be divided in high-performance and low-power. Both types include a set of specialized cores targeting high performance or low power while maintaining the same instruction-set architecture (ISA). High-performance clusters will be architectured to reach higher frequencies and will be implemented using faster transistors in G (general/fast) process technology. The low-power clusters will be implemented in LP, low power process technology. LP transistors have very low leakage but can't run at super high frequencies, while G transistors on the other hand are leaky but can switch very fast. In addition a GPU cluster is architectured in such way to minimize area and leakage, while specialized IO clusters supporting Ethernet and PCIs are connected to the NoC minimizing the distance toward the ARM core clusters. Targeting more advanced 20 nm technologies, under development and expected to be available for productions in next 2 years, see Fig. 3, the number of basic clusters, and hence of basic cores in a single chip with 2D integration, should increase by a factor higher than 2 with computational power up to TFLOPS.

![Fig. 3: Evolution of STMicroelectronics CMOS technologies enabling cloud server-on-chip](Image)

The architecture will allow in the future also 3D integration, targeting a number of cores for a cloud server-in-a-single-package of several hundreds. In case of big cloud data center the green cloud platform will be further scaled adopting low latency interconnection for off-chip communication (e.g. LLI Low-Latency-Interface and C2C Chip-to-Chip standards recently proposed by the MIPI Alliance of electronic companies) and achieving thousands of cores on a single board. Therefore, the computing platform will be conceived to realize not only cloud server-on-a-chip (CSoC) solutions, with tens of 64-b ARM cores, but in case of bigger data centers the proposed architecture can be scaled to realize also cloud server-in-a-package (CSiP), with more than one hundred of 64-b ARM cores, and cloud server-on-a-board (CSoB) solutions, with thousands of 64-b ARM cores.

c. On-chip networking infrastructure

As discussed above in order to exploit the energy efficiency of ARM cores for server applications but filling the performance-gap with CISC processors, two approaches are followed:

- extension to a new 64-b instruction set;
- use of many cores exploiting parallelism; in such case an efficient on-chip communication solution has to be found.
According to [17] an important source of power consumption is moving data in interconnects. Quoting data from nVIDIA's 28nm chips, 20 pJ are the computation costs required for performing a floating point operation, and for an integer operation, the corresponding number is 1 pJ. However, getting the operands for the computation from local memory (placed 1mm away) consumes 26 pJ. If the operands need to be obtained from the other end of the die, 1 nJ is required and, if the operands need to be read from DRAM, the cost is 16 nJ.

Since computation costs require much less energy than moving operands to and from the computation units, the traditional cloud computing will implement a hierarchy of caches allowing a considerable reduction of data in and out from DDR (Double Data Rate) DRAMs. In addition, the data coherency among the different clusters will be maintained in hardware by a cache coherent Network-on-Chip. With respect to the state of the art the designed service-aware NoC, also called Interconnect Processing Unit (IPU) supports on-chip advanced service functionalities [19,20] such as cache coherency, power down management, quality of service management, memory remapping and others. Thanks to the IPU the CSoC can be partitioned in different islands each dynamically optimized in terms of power [21-24] according to the required high level service by a well-defined SW application programming interface (API).

The architecture we propose cuts down power and energy consumption by reducing the data transfer in and out of the chip. Of course, such an approach requires a coordinated vertical approach involving HW and SW. For example, at the infrastructure level, cloud management SW (e.g. OpenNebula) should minimize communication, even at the expense of additional computation. If the energy cost of moving data across the die is high, the SW Virtualization Layer should reuse local data for computations whenever possible. The compiler should be able to assess when it is energetically favorable to re-compute or to move data, and should also try to sub-divide problems to minimize communication cost.

IV. PRELIMINARY RESULTS FOR CLOUD SERVER-ON-CHIP

Fig. 4 shows the proposed green cloud architecture to realize a CSoC, with estimated complexity in 28 and 20 nm CMOS technology. The architecture hierarchically combines:

- A first cloud computing kernel represented by the cluster of 4 ARM processors of the v8 family with 64-b instruction set extension (Core v8 in Fig. 4). The architecture integrates a number of high-performance and low power clusters for a total of 4 clusters.

- In addition, an IO cluster which includes a fast PCI-Express (PCIe) and Ethernet MAC for high-speed I/O and two special links for networking with other CSoC devices.

- A second level represented by high-speed STNoC IPU encircling all clusters plus two DDR4 multi-port memory controllers for off-chip access.

- All I/O modules are connected to the NoC via I/O Memory management units (IOMMUs) translating all devices access to host memory.

From the memory hierarchy point of view, beside the L1 cache for each ARMv8 core, a L2 cache memory of 2Mbytes is shared among the 4 processors in each cluster. Thanks to the STNoC IPU, which exploits a Spidergon topology with bi-directional rings and across connections, the memory can be extended to on-chip L3 cache (4 blocks in Fig. 4) and to large off-chip DDR DRAM thanks to two memory controllers, each with 2 ports, connected through the NoC.

From preliminary synthesis in 28 nm and 20 nm CMOS technology nodes, a complexity of about 30 mm² in 28 nm for a v8-like cluster can be foreseen. In 20 nm the occupation for a cluster is halved, roughly 15 mm². The area occupation for the whole architecture is about 200 mm² in 28 nm and 100 mm² in 20 nm, i.e. from 2.5 to 5 times lower than the Itanium processor for server applications discussed in Section II. The main area limiting factor is due to the on-chip memory size for large caches at L2 and L3 hierarchy levels. In these technologies for system-on-chip integration there is a density higher than 3.5 Millions of logic gates in 28 nm and less than 0.15 mm² per Mbits for on-chip SRAM. In 20 nm CMOS technology, for the same area, the complexity of logic gates and memory bits that can be integrated is doubled. The target clock frequency is 1.8 GHz in 28 nm CMOS node and 2 GHz in 20 nm CMOS node. An accurate analysis of the power consumption of the proposed platform in 28 nm and 20 nm CMOS technologies, considering leakage and dynamic power consumption (for different functional benchmarks), is on-going but a power cost of few tens of W, well below 100 W at peak performance, is expected. The area and power results are suited to set up sustainable cloud computing applications.

Given the target ARMv8-based architecture in our roadmap the design of a parametric on-chip interconnect, scalable in terms of connected cores, is foreseen: from few cores of the basic cluster (4) to tens of cores (16 in the proposed embodiment in 28 nm CMOS technology node, see Fig. 4, but that can be increased up to a factor 4 using 20 nm FDSOI CMOS technology) for a server on chip using planar silicon technologies. The number of cores can be raised to hundreds in case of migration to 3D integration technology. The proposed architecture of Fig. 4 implements directly on silicon, in the NoC, advanced networking services to minimize the use of SW services for on-chip communication and the access to off-chip
resources. The NOC has been configured with a flit-size of 128 bits and it supports data conversion, frequency conversion, store&forward transmission, management of out of order transactions, quality of service management, cache coherency, memory remapping, security and power down states.

The SW API provided by the IPU, will enable the creation of a virtualization-based self-adapting infrastructure which dynamically optimizes energy consumption and performance across computing, storage and communications resources within clouds, ensuring that the overall platform performance and energy consumption adapts to the minimum level required to fulfill the contracted QoS for each service. The envisaged energy and performance-aware infrastructure could be extended to work across sites, including federated data centers or data center scale-out to public clouds.

Although virtualization is not the target of this paper, which is focused on the HW architecture, it is worth noting that in the classic virtualization model multiple OS instances are running on each high-powered server processor, e.g. Opteron from AMD or Xeon from Intel. In this way, a virtualized server might use few Xeon processors and a large pool of RAM to run multiple virtual OS instances. Instead for the energy-efficient servers based on 32-b ARM RISC cores, recently proposed in [5,8], the idea is having multiple boards each hosting multiple System-on-Chip devices integrating ARM cores with one-OS-per-chip. For instance, in the Calxeda solution [8] each board has 4 ARM-based multi core chips with a power consumption of 20 W per board and 5 W per chip. As a consequence, to run multiple virtual OS instances multiple boards are needed and hence multiple chips. Indeed a limit of traditional ARM architecture is that it does not satisfy the requirements to be virtualizable [25]. The main obstacle is that a number of instructions behave differently when in user mode rather than in privileged mode; for the platform to be realistically virtualizable without significant slowdowns incurred by binary patching techniques, these instructions should trap. This is facilitated by a number of new features in the ARMv8 which enable efficient SW virtualization through HW acceleration, usually referred as the "ARM Virtualization Extensions". These extensions include virtual-to-physical address translation, protection, partitioning and resource management of complex systems involving client and server devices and SW stacks into virtual machines. To this aim, a new Hypervisor processor mode is introduced, which allows each guest to have access to its own privileged mode; the processor state can be switched between guests, allowing the processor to be virtualized without expensive binary patching techniques, and with very few traps being necessary. Virtualization Extensions also allow certain instructions to be set up to trap if that is necessary to run a guest efficiently.

V. CONCLUSIONS

The role of energy-efficient cloud-server-on-chip solutions in reducing the total cost of ownership and the ecological impact of cloud computing data centers has been discussed. A green cloud computing platform, based on multi-cluster architecture with upcoming 64-b RISC processors and off-chip memory controllers and fast I/O modules, interconnected by a power-aware IPU ensuring cache coherency, could achieve TFLOPs performance. The area and power costs for CMOS implementations lead to sustainable cloud services.

REFERENCES