HOW GREEN IS YOUR CLOUD?

A 64-b ARM-based Heterogeneous Computing Platform with NoC Interconnect for Server-on-chip Energy-efficient Cloud Computing

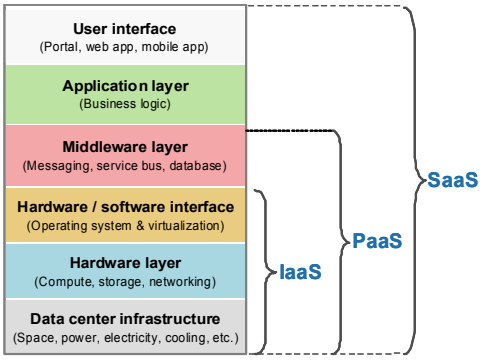
Sergio Saponara[2], Marcello Coppola[1] and Luca Fanucci[2]
[1]*STmicroelectronics, Grenoble, France*
[2]*Dip. Ingegneria della Informazione, Università di Pisa, Pisa, Italy*

Keywords: Green Cloud Computing, Energy-efficiency, Hardware Platform, Network on Chip (NoC), Multi-core Systems.

Abstract: This position paper discusses the role of energy-efficient cloud-server-on-chip (CSoC) solutions to reduce the total cost of ownership and the ecological impact of cloud computing data centers. A green cloud computing platform, based on a multi core architecture with upcoming 64-b ARM processors of the ARMv8 family, interconnected by a service-aware Network on Chip (NoC) ensuring cache coherency, could reduce costs (due to energy consumption and extra cooling systems) and increase system reliability (by avoiding thermal issues) of cloud data centers. Implementation figures on 28 nm and 20 nm silicon technology nodes from STMicroelectronics are provided.

1 NEED OF GREEN CLOUD COMPUTING

Cloud computing is where data, software applications, or computer processing power are accessed from the cloud of online resources. This permits individual users to access their data and applications to any edge device. As well as to any organization to reduce their capital costs by purchasing hardware and software as utility services. Cloud Computing, also referred as utility computing, as has rapidly emerged as a new computing paradigm representing a fundamental shift in delivering information technology services via on-demand computing resources.



Source: Morgan Stanley Research

Figure 1: Layers in the Cloud Computing Architecture.

According to Fig. 1 cloud computing is generally defined as consisting of 3 layers or set of services: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS).

These layers are separating the end-user to the physical infrastructure of the data center. All this allows users to run applications and store data online. However each layer offers a different level of user flexibility and control. In particular, IaaS is the set of services closest to the hardware, which allows users to run any applications on the cloud hardware of their choice. They include typical services at the operating system-level creating abstractions of "Resource Clouds" consisting of managed and scalable resources using enhanced virtualisation capabilities. Abstract resources are provided via a service interface allowing for:

- Data & Storage Clouds dealing with reliable and timely access to data;
- Computing Clouds providing scalable, reliable and timely access to computational resources.

The fundamental unit of the cloud computing infrastructure is a server, which can be physical or virtual. Physical servers are discrete individual computer on which online applications can be run

Saponara S., Coppola M. and Fanucci L..
HOW GREEN IS YOUR CLOUD? - A 64-b ARM-based Heterogeneous Computing Platform with NoC Interconnect for Server-on-chip Energy-efficient Cloud Computing.
DOI: 10.5220/0003959801350140
In *Proceedings of the 2nd International Conference on Cloud Computing and Services Science* (CLOSER-2012), pages 135-140
ISBN: 978-989-8565-05-1

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 centre.

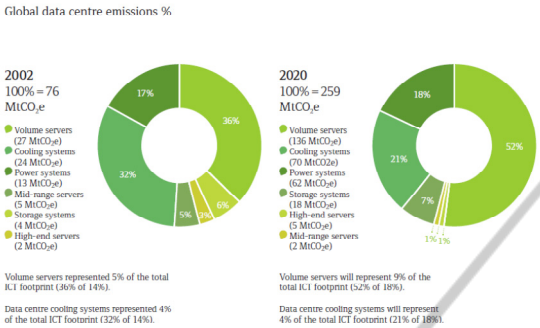


Figure 2: % of emissions for different classes of systems.

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 (Schubert, 2010). However, the total cost of ownership (TCO) is still an obstacle for having billions of users accessing millions of services. In fact, a significant fraction of a data center's TCO comes from the recurring energy costs consumed during its lifecycle. Beside the cost of energy (Fan et al., 2007; White et al., 2004; Shah et al., 2008), other issues are driving the R&D focus on cloud computing platforms towards low-power: the ecological impact of cloud servers and the thermal management since high power consumption needs costly and heavy cooling systems and the thermal stress reduces the reliability and the life time of the computing platforms (Mullas et al., 2009; Reda, 2011). The problem of power "Green IT" is one of the most popular areas in research today.

In particular, energy-efficient cloud computing is one of the key research item since analysis of the US Environmental Protection Agency (EPA) states that powering the nation's data centers consisting of vast server grids, power supply and cooling infrastructures is growing to unprecedented levels, ~100 billion KW hours by 2011 costing $7.5B.

In September 2011, Google said that its global operations continuously draw 260 million megawatts of power, roughly a quarter of the energy generated by a nuclear power plant. Considering that most server processors in use today draw about 160 watts under normal operations and 80 watts even when they're idle, there is a huge opportunity for power reduction. Moreover, as the level of carbon emission in IT and communication technologies is estimated to triple by 2020, see Fig. 2, IT professionals continue to raise public awareness and 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. Broad adoption of cloud computing can stimulate innovation and accelerate the deployment of these enabled solutions. Cloud computing may have a major impact on global carbon emissions. In the new era of energy efficient performance, the question how cloud computing could be implemented in practice the constraints of efficient energy utilization.

2 STATE-OF-ART MANY-CORE PLATFORMS

In 2005, PC market hits a wall. As described in (Sutter, 2005), "the Free Lunch was Over". The major PC manufacturers from Intel and AMD to IBM, have run out of gas 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 extreme miniaturisation. Higher still performance is also required for enterprise computing and communications. 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. During all these years we have seen 3 majors multi-core computing platform styles (Woo et al., 2008; Kumar et al., 2003a; Kumar et al., 2003b). The first one, it is a symmetric many-core processor that simply replicates a state-of-art superscalar processor on a die. The second one, it 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. One example is to have a state-of-the-art host processor as the host 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 in which we completed to small-scale heterogeneous multi-core computing platform with the arrival of multi-core tablets (e.g., iPad 2, Playbook, Kindle Fire, Nook Tablet) and smartphones (e.g., Galaxy S II, Droid X2, iPhone

4S). On the server computing side we have assisted to the recent transition to high performance multi-core platforms, providing an array of register sets, shared cache and multiple threads of execution, leading to power-efficient, high bandwidth (and often real-time) processing similar to that of past supercomputers and/or embedded systems. Existing high end server architectures, such as IBM's Power7 high-end server, Oracle Sun PowerNap processor and AMD's "Magny Cours" Opteron processor incarnation, benefit from powerful out-of-order cores, large on-chip cache hierarchies and/or rapid non-intrusive power model transitions between a high-performance active state and a near-zero-power idle state. This trend will carry on in 2012 with quad- and eight-core solutions.

However the high computational capabilities, Tera operations per seconds, of the above platforms are paid with high power and energy cost, and related thermal issues reducing system reliability and creating cooling overheads 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 and entertainment applications, but also to services to increase efficiency of the public administration or of private companies, both SME and big ones. Hence reliability, low risk or service denial, long life time, low ambiental impact, reduced energy cost are all key factors for the widespread adoption of the cloud computing paradigm.

3 GREEN CLOUD PLATFORM WITH MANY-CORE 64-B ARM PROCESSORS AND SERVICE-AWARE NOC INTERCONNECT

In this position paper a novel cloud computing infrastructure is introduced that considers energy awareness as key optimization factor. At the lowest levels, special hardware support for power efficiency and proper mechanisms for supporting it at the hypervisor level within the infrastructure of an IaaS provider are investigated. Particularly, the following objectives will be addressed.

The green cloud computing will be based on a new ARM architecture with the addition of a new "A64" 64-b instruction set, which will enable to use ARM 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 mobile and consumer legacy software 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. It was announced earlier this year that Windows 8 will support ARM architecture (Osborne, 2011). Moreover, ARM Cortex-A9 MPCore can already compete with a 1.6GHz Intel Atom (Humphries, 2010), even when running at 500MHz. Last but not least, the new ARMv8 architecture features the "A64": a 64-b instruction set, which will enable to use ARM core into the server and enterprise computing space (Grisenthwaite, 2011; Goodacre, 2011). 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 capability (Goodacre, 2011).

ARMv8 will effectively be a superset of ARMv7. All ARMv8 chips will run legacy 32-bit ARMv7 code in the "AArch32" execution state, while 64-bit 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 particularly interesting for server applications such as cloud computing.

As far as the target implementation technology is concerned, STMicroelectronics is developing and producing in Europe high-end CMOS processors and custom System-on-chip integrated circuits (ICs). The silicon technology process of STMicroelectronics has a track record in the low power category of core CMOS, specially devoted to mobile/consumer applications. These technologies are able to address high frequency applications as well, currently as higher than 2.5 GHz in 28 nm CMOS. Concurrently STMicroelectronics is developing technologies on Full-Depleted Silicon-on-Chip (FDSOI) (Skotnicki et al., 2011), which are

able to address both the high performance end of the spectrum (3 GHz and more) and ultra low power operation thanks in particular to an improved electrostatic control on the transistor channel. The challenge is to dynamically adapt the power consumption to the required performance at each moment such as wasting the minimum of energy. To achieve such a goal, a fully vertical integration from technology to software level virtualization layer and virtual machine management would be implemented.

Our estimation is that targeting already available 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 an heterogeneous architecture with a network on chip of several clusters. All ARM clusters have the same ISA but they will be implemented some using G (general/fast) process technology and other 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.

Targeting more advanced 20 nm technologies the number of basic clusters, and hence of basic cores in a single chip with 2D integration, should increase by a factor at least 2 resulting computational power up to Tera operations per seconds.

The architecture will conceived so that it 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 so that it can be used to realise 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.

According to (Dally, 2001), an important source of power consumption is moving data in interconnects. Quoting data from nVIDIA's 28nm chips (see Fig. 3), 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 (situated 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. As shown in (Kumar et al., 2005) the interconnect network of an eight-core microprocessor might consume as much energy as one core and as much area as three cores.

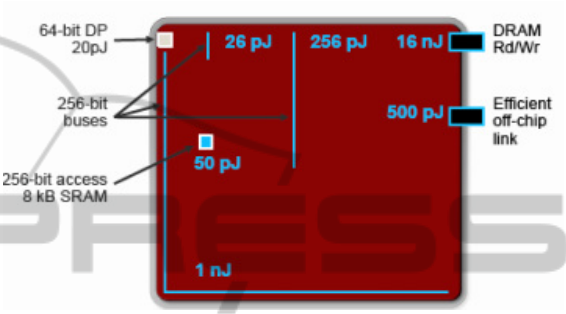


Figure 3: Computation cost is significantly lower than communication cost in 28nm nVIDIA chips (Dally, 2001).

The energy required for computation is significantly smaller than the energy needed for interconnects. In addition, this trend is expected to get worse with scaling.

In conclusion, computation costs require much less energy than moving operands to and from the computation units. Thus, the traditional cloud computing SoC architecture which extensively uses off-chip SDRAM as the main memory for buffering and/or for transferring data information needs rethinking and a well-deserved optimization. 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 touching hardware and software. For example, at the infrastructure level, cloud management software (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 Virtualization Layer SW should reuse local data for computations whenever possible. The compiler should be able to figure out when it is energetically favourable to recompute or to move data, and should also try to sub-divide problems to minimize communication cost.

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 (Coppola et al., 2008; Saponara et al., 2012) such as cache coherency, power down management, quality of service management, memory remapping and others. Thanks to the IPU the multi-core server-on-chip can be partitioned in different islands each dynamically optimized in terms of power (Saponara et al., 2007; Saponara et al., 2011a; Saponara et al., 2011b; Vitullo et al., 2008) according to the required high level software service by a well defined software API.

4 GREEN CLOUD ARCHITECTURE IN SUBMICRON CMOS

Figure 4 shows the proposed green cloud architecture enabling the realization of a cloud server-on-single-chip, with estimated complexity in 28 nm and 20 nm CMOS technology by STMicroelectronics. The architecture is hierarchically organized in a combination of:

- a first cloud computing kernel represented by the cluster of 4 ARM processors of the V8 family with 64-bit instruction set extension (referred as Core v8 in Fig.4);

- a second level represented by high-speed STNoC IPU (Coppola et al., 2008) encircling 4 clusters for a total of 16 ARMv8 processors plus two DDR4 multi-port memory controllers for off-chip access, fast PCI-Express (PCIe) and Ethernet MAC for high-speed I/O and two special links for networking with other server-on-chip devices. All I/O modules are connected to the STNoC via I/O Memory management units (IOMMUs) that translate all devices access to host memory.

From the memory hierarchy point of view, beside the L1 cache for each ARM V8 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 of-chip DDR DRAM thanks to two memory controllers, each with 2 ports, connected through the NoC.

From preliminary synthesis in CMOS silicon technology considering 28 nm and 20 nm technology nodes, a complexity of about 30 mm^2 in 28 nm for a V8 like cluster can be foreseen. In 20 nm the occupation for a cluster is halved, roughly 15 mm^2.

The area occupation for the whole architecture is about 200 mm^2 in 28 nm and 100 mm^2 in 20 nm.

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

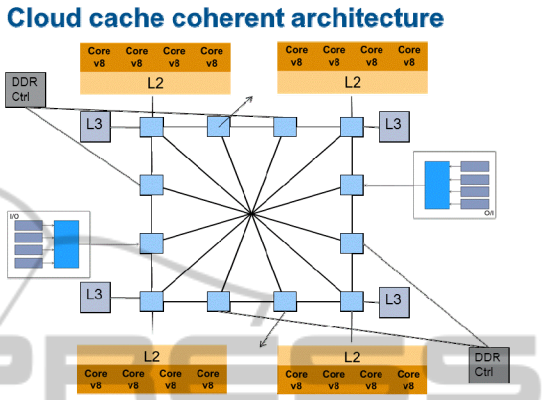


Figure 4: Architecture of the green cloud for server on a single chip realizations.

of logic gates in 28 nm and less than 0.15 mm^2 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.

Given the target ARM-based architecture we have foreseen the design of a parametric on-chip interconnect scalable in terms of connected cores; from few cores of the basic cluster (4) to tens of cores (16 in the proposed embodiment but that can be increased by 2 or by 4 using 20 nm 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 in hardware, in the NoC, advanced networking services to minimize the use of software 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, security and power down states.

In addition, the software 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 fulfil the contracted QoS for each service defined in terms of measurable KPIs. Finally, 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.

5 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 in this position paper. A green cloud heterogeneous computing platform, based on a multi-clusters architecture with upcoming 64-b ARM processors of the ARMv8 family and off-chip memory controllers and fast I/O modules, interconnected by a power-aware IPU ensuring cache coherency, could achieve a better performance gain per watt. In addition the input/output (IO) has been integrated; thereby a latency and power reduction will be envisaged mainly due better sharing of data.

REFERENCES

M. Coppola, M. Grammatikakis, R. Locatelli, G. Maruccia, L. Pieralisi, Design of cost efficient interconnect processing units, CRC Press Book, September 2008

T. Cucinotta et al., "QoS Control for Pipelines of Tasks Using Multiple Resources," IEEE Tran. on Computers, Vol. 53, N. 3, pp. 416-430, 2010

W. Dally, keynote talk at IEEE Int. Parallel and Distributed Proc. Symp. (IPSPS) 2001

Dong Hyuk Woo et al., "Extending amdahl's law for energy-efficient computing in the many-Core Era", IEEE Computer 2008, pp. 24-31

X. Fan, et al., "Power Provisioning for a Warehouse-sized Computer", ACM ISCA '07

J. Goodacre, White paper, Technology preview: the ARMv8 architecture, pp.1-10, 2011

R. Grisenthwaite, ARMv8 technology preview, 2011.www.arm.com/files/downloads/ARMv8_Architecture.pdf

M. Humphries, "ARM posts Cortex-A9 vs Atom performance video, Intel should be worried". on-line: http://www.geek.com/articles/chips/arm-posts-cortex-a9-vs-atom-performance-video-intel-should-be-worried-2010016/

R. Kumar, et al., "A multi-core approach to addressing the energy-complexity problem in microprocessors", Workshop on Complexity-Effective Design, June 2003

R. Kumar, et al., "Single-ISA Hetero geneous Multi-Core Architectures: The Potential for Processor Power Reduction", IEEE/ACM Int. Symp. Microarchitecture, 2003, pp 81-92

R. Kumar, V. Zyuban and D.M. Tullsen, "Interconnections in Multicore architecture: Understanding Mechanisms, Overheads, and Scaling" Proc. 32nd Ann. Int'l Symp. Computer Architecture (ISCA 05), ACM, 2005

F. Mullas et al., "Thermal Balancing Policy for Multiprocessor Stream Computing Platforms," IEEE Tran. on CAD, Vol. 28, N. 12, 2009

B. Osborne, "Next version of Windows will support ARM-based systems". on-line:http://www.geek.com/articles/mobile/next-version-of-windows-will-support-arm-based-systems-2011015/

S. Reda, "Thermal and Power Characterization of Real Computing Devices", IEEE Tran. on Emerging and Selected Topics in Circuits and Systems, pp. 76-87, vol. 1, N. 2, 2011

S. Saponara et al., "Architectural-level power optimization of microcontroller cores in embedded systems", IEEE Tran. Ind. Electr., vol. 54, n. 1, 2007

S. Saponara et al., "Coverage-driven Verification of HDL IP cores", Springer Lecture Notes in Electr. Engineering, vol. 81, pp. 105-119, 2011

S. Saponara et al. "A multi-processor NoC-based architecture for real-time image/video enhancement", Journal of Real-Time Image Processing, pp. 1-15, doi 10.1007/s11554-011-0215-8, 2011

S. Saponara et al., "Design of a NoC Interface Macrocell with Hardware Support of Advanced Networking Functionalities", SI - Networks-on-Chip, IEEE Transactions on Computers 2012

L. Schubert "The Future of Cloud Computing: Opportunities for European Cloud Computing Beyond 2010" EC, Information Society and Media,http://cordis.europa.eu/fp7/ict/ssai/docs/cloud-report-final.pdf

Shah et al., "Optimization of global data center thermal management workload for minimal environmental and economic burden", IEEE Trans. Components and Packaging Tech., vol. 31, n. 1, pp. 39-45, 2008

T. Skotnicki et al., "Competitive SOC with UTBB SOI", IEEE SOI Conference 2011

Herb Sutter, "The free lunch is over: a fundamental turn toward concurrency in software", Dr. Dobb's Journal, vol. 30, n. 3, 2005

F. Vitullo et al., "Low-complexity link micro architecture for mesochronous communication in Networks on Chip", IEEE Tran. on Computers, vol. 57, n. 9, pp 1196-1201, 2008

White et al., "Energy resource management in the virtual data center", IEEE Int. Symp. Electronics and the Environ., pp. 112- 116, 2004