GreenDataNet

D1.2 - Survey of Key Emerging IT Trends for Data Centres

DRAFT

[Manager]
Dr. Martino Ruggiero
Embedded Systems Laboratory (ESL) - EPFL
Prof. David Atienza
Embedded Systems Laboratory (ESL) – EPFL

Rev 2.0

TABLE OF CONTENTS

REVISION SHEET

Revision Number	Date	Brief summary of changes
Rev 1.0	15/04/2014	Baseline document
Rev 2.0	01/05/2014	Final document

1. INTRODUCTION

1.1. DOCUMENT PURPOSE

The on-going transition toward information economies has created an unprecedented demand for server hardware that is used to process, store, and transmit digital data. In order to accommodate the enormous storage and computational requirements of many essential information services, as well as to leverage the cost savings associated with economies of scale, large-scale datacenters have emerged as the backbone of the IT industry. These datacenters, which often house hundreds to thousands of servers, have proliferated as a result of the rapid growth in information services and digital entertainment sectors. One side-effect of this expansion has been a dramatic increase in datacenter power consumption.

1.2. DOCUMENT SCOPE

This deliverable focuses on the key emerging Information Technology (IT) trends and approaches toward more green and energy efficient datacenters. The different approaches are reviewed and categorized with regards to their field of applicability, starting from the top layers represented by the application software virtualization, as well as the energy and thermal software management, towards the lower hardware layers, represented by innovative server architectures and cooling infrastructures.

1.3. DEFINITION, ACRONYMS AND ABBREVATIONS

1.3.1. KEY DEFINITIONS

Data centre, datacenter or datacentre

A structure, or group of structures, dedicated to the centralized accommodation, interconnection and operation of information technology and network telecommunications equipment providing data storage, processing and transport services together with all the facilities and infrastructures for power distribution and environmental control together with the necessary levels of resilience and security required to provide the desired service availability

Note: A structure can consist of multiple buildings and/or spaces with specific functions to support the primary function.

1.3.2. KEY ACRONYMS AND ABBREVATIONS

CPU	Central Processing Unit
DCIM	Data Centre Infrastructure Management
EC	European Commission
ICT	Information and Communication Technology
IT	Information Technology
PDU	Power Distribution Unit
PUE	Power Usage Effectiveness
SLA	Service Level Agreement
VM	Virtual Machine
WP	Work Package

1.4. KEY REFERENCES AND SUPPORTING DOCUMENTATION

[1] Gerald J. Popek and Robert P. Goldberg. "Formal requirements for virtualizable third generation architectures." Communications of the ACM 17.7 (1974): 412-421.

[2] Mike Nelson, Michael, Beng-Hong Lim, and Greg Hutchins. Fast transparent migration for virtual machines. In USENIX ATC. 2005.

[3] Edouard Bugnion, Scott Devine, Mendel Rosenblum, Jeremy Sugerman, and Edward Y. Wang. 2012. Bringing Virtualization to the x86 Architecture with the Original VMware Workstation. ACM Trans. Comput. Syst. 30, 4, Article 12 (November 2012), 51 pages.

[4] Ripal Nathuji and Karsten Schwan. 2007. VirtualPower: coordinated power management in virtualized enterprise systems. In SOSP ’07 265-278. http://doi.acm.org/10.1145/1294261.1294287

[5] IDC (Press Release). “Server Virtualization Hits Inflection Point as Number of Virtual Machines to Exceed Physical Systems in 2009", http://www.idc.com/about/viewpressrelease.jsp?containerId=prUK21840309, 2009.

[6] G. Dhiman et al., “vGreen: A System for Energy Efficient Computing in Virtualized Environments,” in Proc. ISLPED, 2009.

[7] Y. Chen, A. Das, W. Qin, A. Sivasubramaniam, Q. Wang and N. Gautam, "Managing server energy and operational costs in hosting centers," in SIGMETRICS, 2005.

[8] T. Heath, A. P. Centeno, P. George, L. Ramos, Y. Jaluria and R. Bianchini, "Mercury and freon: temperature emulation and management for server systems," in ASPLOS, 2006.

[9] W. Zheng, R. Bianchini, G. J. Janakiraman, J. R. Santos and Y. Turner, "JustRunIt: Experiment-Based Management of Virtualized Data Centers," in USENIX, 2009.

[10] N. Vasic, D. Novakovic, S. Miucin, D. Kostic and R. Bianchini, "DejaVu: Accelerating Resource Allocation in Virtualized Environments," in ASPLOS, 2012.

[11] B. H. McKeown, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee and N. McKeown, "ElasticTree: Saving Energy in Data Center Networks," in NSDI, 2010.

[12] M. Gupta and S. Singh, "Greening of the Internet," in SIGCOMM, 2003.

[13] N. Vasic and D. Kostic, "Energy-Aware Traffic Engineering," in e-Energy, 2010.

[14] N. Vasic, P. Bhurat, D. Novakovic, M. Canini, S. Shekhar and D. Kostic, "Identifying and using energy-critical paths," in CoNEXT, 2011.

[15] S. Yaldiz, A. Demir, S. Tasiran, P. Ienne, and Y. Leblebici. Characterizing and Exploiting Task-Load Variability and Correlation for Energy Management. In Proceedings of the 3rd Workshop on Embedded Systems for Real-Time Multimedia (ESTIMedia '05), pages 129-140, New York, NY, USA, September 22-23, 2005.

[16] Zeynep Toprak-Deniz, Yusuf Leblebici, and Eric Vittoz. On-Line Global Energy Optimization in Multi-Core Systems Using Principles of Analog Computation. In IEEE Journal of Solid-State Circuits, volume 42, number 7, pages 1593-1606, 2007.

[17] A. Andrei, M. Schmitz, P. Eles, Z. Peng, B.M. Al-Hashimi, Overhead-Conscious Voltage Selection for Dynamic and Leakage Energy Reduction of Time-Constrained Systems, DATE 2004, pp.105-118, Paris- France.

[18] E. Pakbaznia et al., “Minimizing Data Center Cooling and Server Power Costs,” in Proc. ISLPED, 2009.

[19] N. Tolia et al., “Unified Thermal and Power Management in Server Enclosures,” in Proc. InterPACK, 2009.

[20] X. Wang et al., “Coordinating Power Control and Performance Management for Virtualized Server Clusters,” IEEE T-PDS, vol. 22, no. 2, pp. 245-259, 2011.

[21] G. Dhiman et al., “vGreen: A System for Energy Efficient Computing in Virtualized Environments,” in Proc. ISLPED, 2009.

[22] D. Meisner et al., “Power Management of Online Data-Intensive Services,” in Proc. ISCA, 2011.

[23] L. Parolini et al., “A Cyber-Physical Systems Approach to Energy Management in Data Centers,” in Proc. ICCPS, 2010.

[24] M. M. Sabry, et al., “Energy-Efficient Multi-Objective Thermal Control for Liquid-Cooled 3D Stacked Architectures”, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (T-CAD), Vol. 30, Nr. 12, pp. 1883-1896, December 2011.

[25] C. E. Bash et al., “Dynamic Thermal Management of Air Cooled Data Centers,” in Proc. ITHERM, 2006.

[26] D. C. Hwang et al., “Energy Savings Achievable Through Liquid Cooling: a Rack Level Case Study,” in Proc. ITHERM, 2010.

[27] J. Koomey. Growth in Datacenter Electricity Use 2005 to 2010. Analytics Press, Oakland, CA, USA, August 2011.

[28] B. Zhai, D. Blaauw, D. Sylvester, K. Flaunter. Theoretical and Practical Limits of Dynamic Voltage Scaling. In Design Automation Conference (DAC), 2004.

[29] K. Lim, P. Ranganathan, J. Chang, C. Patel, T. Mudge, and S. Reinhardt. Understanding and Designing New Server Architectures for Emerging Warehouse-Computing Environments. In International Symposium on Computer Architecture (ISCA), 2008.

[30] R. St. Amant, D. Jiménez, and D. Burger. Low-power, high-performance analog neural branch prediction. In International Symposium on Microarchitecture (MICRO), 2008.

[31] Patel, C.D. A Vision of Energy Aware Computing from Chips to Datacenters. in The International Symposium on Micro-Mechanical Engineering - ISMME2003-K15. December 1-3, 2003. Tsuchiura and Tsukuba, Japan.

[32] Koomey, J.G. Growth in Datacenter Electricity use 2005 to 2010. August 1, 2011. http://www.analyticspress.com/datacenters.html.

[33] EIA, Monthly Energy Review - DOE/EIA-0035(2011/07). July 2011, U.S. Energy Information Administration.

[34] Koomey, J.G., Estimating Regional Power Consumption by Servers: A Technical Note. 2007, Lawrence Berkeley National Laboratory: Oakland, CA.

[35] Miller, R. Uptime Institute: The Average PUE is 1.8. May 10th 2011. http://www.datacenterknowledge.com/archives/2011/05/10/uptime-institute-the-average-pue-is-1-8/

[36] Saini, M. and Webb, R.L., Heat Rejection Limits of Air Cooled Plane Fin Heat Sinks for Computer Cooling. IEEE Transactions on Components and Packaging Technologies, 2003. Vol. 26(1): pp. 71-79.

[37] Pawlish, M. and Varde, A.S. Free Cooling: A Paradigm Shift in Datacenters. in 5th International Conference on Information and Automation for Sustainability (ICIAFs). 2010.

[38] Samadiani, E., Joshi, S., and Mistree, F., The Thermal Design of a Next Generation Datacenter: A Conceptual Exposition. Journal of Electronic Packing, 2008. Vol. 130: pp. 041104-1 - 041104-8.

[39] Greenberg, S., Mills, E., and Tschudi, B., Best Practices for Datacenters: Lessons Learned from Benchmarking 22 Datacenters. ACEEE Summer Study on Energy Efficiency in Buildings, 2006: pp. 3-76 - 3-87.

[40] Hannemann, R. and Chu, H., Analysis of Alternative Datacenter Cooling Approaches, in InterPACK '07. 2007: Vancouver, BC, CA.

[41] Sabry, M.M., Coskun, A.K., Atienza, D., Simunic, T., and Brunschwiler, T., Energy-Efficient Multi-Objective Thermal Control for Liquid-Cooled 3D Stacked Architectures. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (T-CAD), 2011. Vol. 30(12): pp. 1883-1896.

[42] Scott, A.W., Cooling of Electronic Equipment. 1974, Wiley: New York. p. 204-227.

[43] Bash, C.E. Analysis of Refrigerated Loops for Electronics Cooling. in Pacific Rim/ASME Int. Electron. Packag. Tech. Conf. Exhibition (IPACK'01). 2001. Kauai, HI.

[44] Peeples, J.W., Vapor Compression Cooling for High Performance Applications. Electronics Cooling, 2001. Vol. 7: pp. 16-24.

[45] Heydari, A. Miniature Vapor Compression Refrigeration Systems for Active Cooling of High Performance Computers. in 8th Intersoc. Conf. Thermal Thermomech. Phenom. Electron. Syst. (I-THERM). 2002.

[46] Schmidt, R.R. and Notohardjono, B.D., High-End Server Low-Temperature Cooling. IBM Journal of Research and Development, 2002. Vol. 46: pp. 739-751.

[47] Maveety, J.G., Brown, M.F.W., Chrysler, G.M., and Sanchez, E.A. Thermal Management for Electronics Cooling using a Miniature Compressor. in Int. Microelectron. Packag. Soc. (IMAPS). 2002. Denver, CO.

[48] Phelan, P.E. and Swanson, J. Designing a Mesoscale Vapor-Compression Refrigerator for Cooling High-Power Microelectronic. in Inter Soc. Conf. Thermal Thermomech. Phenom. Electron. Syst. (I-THERM). 2004. Las Vegas, NV.

[49] Suman, S., Fedorov, A., and Joshi, Y. Cryogenic/Sub-Ambient Cooling of Electronics: Revisited. in Inter Soc. Conf. Thermal Thermomech. Phenom. Electron. Syst. (I-THERM). 2004. Las Vegas, NV.

[50] Olivier, J.A., Marcinichen, J.B., Bruch, A., and Thome, J.R., Green Cooling of High Performance Micro Processors: Parametric Study between Flow Boiling and Water Cooling. International Journal of Thermal Sciences and Engineering Application, 2011. Vol. 3: pp. 041003.1-041003.12.

[51] Marcinichen, J.B., Olivier, J.A., Oliveira, V., and Thome, J.R., A Review of On-Chip Micro-Evaporation: Experimental Evaluation of Liquid Pumping and Vapor Compression Driven Cooling Systems and Control. International Journal of Applied Energy, 2011. Vol. 92: pp. 147-161.

[52] Marcinichen, J.B., Thome, J.R., and Michel, B., Cooling of Microprocessors with Micro-Evaporation: A Novel Two-Phase Cooling Cycle. International Journal of Refrigeration, 2010. Vol. 33(7): pp. 1264-1276.

[53] Marcinichen, J.B., Olivier, J.A., and Thome, J.R., On-Chip Two-Phase Cooling of Datacenters: Cooling System and Energy Recovery Evaluation. International Journal of Applied Thermal Engineering, 2011. DOI: 10.1016/j.applthermaleng.2011.12.008.

1.5. DOCUMENT OVERVIEW

The rest of the document is organized as follows. Section 2 presents and reviews the software approaches for managing consolidation exploiting virtual machines inside the datacenters. Energy and thermal management techniques are presented in Section 3. Section 4 focuses on innovative trends in hardware server architecture design towards more green IT. Section 5 review the state-of-the-art of innovative cooling Infrastructure and approaches. Section 6, finally, closes the document with the conclusions.

2. VIRTUAL MACHINES MANAGEMENT

Virtualization is the fundamental mechanism that uses a combination of multiplexing, aggregation and emulation to decouple a logical resource from its physical instantiation [1]. The former is semantically associated with a particular workload, application, or customer, whereas the latter consumes actual energy. The decoupling occurs both in space (multiplexing, partitioning, scheduling) and in time (e.g. virtual machine migration [2]). The introduction of hardware support for virtualization has largely eliminated the inefficiencies of early systems [3]. More recently, power management has been considered in the context of virtualized deployments [4]. However, although the majority of datacenter servers are virtualized today [5], large-scale analytics are largely deployed without virtualizing either the compute or the data involved. The authors of [6] outline the impact of a characterized workload of a virtual machine (VM) on server physical resources utilization. This work shows that homogeneous VM workload mapping on the same physical resources does not result in lower energy consumption.

Identifying the optimal resource allocation in virtualized environments is a challenging endeavor. State-of-the-art approaches either rely on analytical models or experiment-based management of virtualized datacenters [7] [8] [9]. However, both approaches suffer from an unacceptable delay once the workload changes as they feature a long adjustment period. During this period, the system may violate a service level objective (SLO) if running with an inefficient configuration, or waste resources otherwise.

To simplify and accelerate the management of virtualized resources, other approaches have been proposed [10] that cache and quickly reuse the results of previous resource allocation decisions once the workload changes. On the network side, several existing efforts [11][12][13], reduce network energy consumption by using traffic consolidation. Unfortunately, computing the minimum network subset is computationally hard. Authors in [8] address the high re-computation rate by proposing a hybrid approach which pre-computes a few energy-critical paths off-line, and uses them selectively at runtime.

A new research trend in this topic is developing techniques for automatically identifying the optimal allocation of computing and networking resources, by relying on detailed workload characterization and self-optimizing techniques [14]. The objective is to achieve high degrees of utilization on the active nodes, enabling the inactive nodes to enter low-power states. To that end, workload characterization enables to cluster workloads into workload-compliant groups that make it possible to aggregate large numbers of virtualized workloads onto the soon-to-arrive massively multicore machines. One important issue that arises in this situation is resource interference due to competing workloads on the same physical platform.

3. ENERGY AND THERMAL MANAGEMENT

Managing energy and power consumption in data centers has been a major research topic in the last five years. This interest has led to improving energy efficiency through either software-based optimization or software-controlled power scaling at hardware level. On one hand, from a components-based perspective, dynamic voltage scaling (DVS) is one of the preferred approaches for trading energy dissipation for performance in datacenter components based on workload predictions [15], and account for relative timing dependencies among all tasks running in the system [16]. Several approaches also consider minimizing the overall dissipated energy in a multiple-PE system under timing constraints [17], as well as considering the cold air inlet temperature [18] or the fan and blades activities [19], but these approaches do not formalize the optimization with considerations of the dynamic workload characteristics. Finally, the use of several cascade controllers, where each controller is responsible for a specific element (e.g., fans, virtual machine migration, DVFS, etc.) and control theory [20] [21] has been proposed, but this approach does not yield an optimum management at system level for complete datacenters because such optimizations techniques are designed for worst case server workload and cooling scenarios.

On the other hand, researchers have started to propose thermal-aware job scheduling schemes at different software levels to minimize heat recirculation costs in servers and datacenters by applying scale-based hierarchical control [22][23], rule-base filters [24], heuristic-based approaches [25], or programming methods [26]. However, these thermal management techniques do not include important electrical control knobs and cooling flow controls, and are not aligned with services and resource utilization requirements of Data Analytics.

4. DATA PROCESSING ARCHITECTURES

One side-effect of the datacenters expansion has been a dramatic increase in server power consumption. A recent report to the U.S. Congress noted that datacenter server power consumption has grown at a 17% compound annual rate from 2000 to 2006. By 2010, global datacenter energy usage was estimated at around 1.5% of the world-wide energy production capacity [27] and is widely expected to grow at an exponential pace in the coming decade.

Increasing demand for servers is partly responsible for the mounting energy expenditures. However, a major culprit behind the datacenter energy crisis is the server hardware itself, particularly the processor chips that power these servers. Historically, on-chip voltages dropped with each technology generation, effectively countering the growing energy demands associated with the doubling of transistor density per Moore's law. However, physical limits of nano-scale devices have largely brought an end to voltage scaling. As a result, chips have become power-limited – a development that threatens the continuation of Moore's law, the vehicle fueling the digital revolution.

Several techniques have been proposed to improve the energy-efficiency of servers. At the circuit level, voltage and frequency scaling (DVFS) has proven effective, but its usefulness will become increasingly limited as the gap between core and threshold voltages shrinks [28]. At the architecture level, researchers have advocated the use of lower-power processors and memories, such as those found in laptops and embedded systems [29]. While such components offer better efficiencies than today's server chips, the resulting systems fall significantly short of the optimal due to their general-purpose architectures and lack of server-grade features.

Existing server chips are designed with a broad range of application domains in mind, including scientific, entertainment, and data processing. The resulting architectures are completely general purpose, offering reasonable performance for a variety of application classes, but at a cost of poor area and energy efficiency. To overcome the efficiency limitations of today's designs, researchers and industries are developing fully programmable processor architectures that are tuned to the demands of prevalent data-processing workloads. To that end, core- and chip-level organizations are developed that maximize silicon efficiency through tight coupling of execution resources and memories via a light-weight communication substrate optimized for the workload requirements.

Contemporary data-intensive workloads, such as web search and business analytics, manipulate vast quantities of data. In many cases, memory accesses dominate the execution time with low arithmetic intensity, which often occurs during common indexing operations, such as search and sort. Such memory-intensive operating regimes leave the processor idle as it waits for memory operations to complete, resulting in low utilization and poor energy efficiency. Several research groups are proposing augmenting the server cores with specialized hardware for off-loading common indexing operations, thereby enabling dramatic reductions in processor energy consumption at comparable or higher performance levels.

While some operations, such as database transactions, require absolute fidelity, others are amenable to reduced degrees of precision. Recent research has uncovered the potential for significant energy savings at the micro-architectural level through inexact (analog) computation [30]. Other work has shown that certain applications, particularly in the media processing space, are tolerant of various degrees of imprecision. Several approaches build on these insights by investigating the potential for systematic use of approximation in servers, from hardware structures to system and application software, with the objective of saving energy in data-intensive domains by shedding processor load.

Finally, other approaches are investigating integration options at the chip and system level to improve efficiency from the perspective of performance and energy consumption. One particularly acute problem this research is trying to address is that of the "bandwidth wall", whereby package pin limitations are impeding the performance scalability of many-core architectures. To that end, options in the form of 3D stacked memories acting as data caches, as well as optical chip-to-chip and chip-to-memory interfaces that may offer greater bandwidth and lower data transfer latencies, have been examined.

5. COOLING INFRASTRUCTURE

The demands for higher processor speeds and miniaturization have led to a steady rise in the heat flux generated by the electronics. This increase in power density must be analyzed not only at the level of the electronic components, but also at the system- and datacenter levels – following the path from the local point of electrical power consumption/heat dissipation locally on a chip, all the way through the system enclosure to the room and out to the environment. Patel [31] highlighted two clear objectives for future thermo-mechanical solutions – to facilitate effective heat transfer from high power density chips and systems in order to maintain a specified temperature on the device, and to facilitate the heat removal efficiently by minimizing the energy used to remove the dissipated heat. A third objective, no less important in magnitude and which is starting to become extremely relevant due to environmental aspects, is to explore the potential of waste heat recovery. A fourth objective is to put a cost on this energy consumption and heat dissipation “trail” from chip to the environment, such that specific technological investments can be evaluated and justified. The ultimate objective is to model this entire process such that energy management and optimization schemes can monitor and operate the entire system to achieve green computing, big data management and data storage.

To put this problem in perspective, the energy usage of datacenters in the US was estimated to make up about 2% of the total electricity production in 2010 (growing 15% annually), which represents about 82 billion kWh at an annual cost of approximately $6.1 billion [32][33]. Cooling of datacenters can represent up to 45% [34][35] of this total consumption using current cooling technologies (air cooling). This means an estimated 37 billion kWh usage for 2010 in the US, with an annual cost of $2.7 billion, just for cooling.

Steady performance increases in microprocessors have pushed the limits of air cooling. Future microprocessors are expected to have heat fluxes on the order of 100 W/cm2; meanwhile, Saini and Webb [36] proved that the maximum heat removal capacity of air cooling technology is 37 W/cm2. Issues such as these highlight the need for alternative solutions to air cooling. The partial solution adopted by thermal designers of datacenters is the confinement of the cooled air inside the rack cabinets and the rack cabinets inside of containers, as an attempt to maximize the cooling performance and to reduce the overall thermal resistance between the chip and the external environment. However, these advanced cooling systems are designed independently of the server architecture [37]. Thus, their cooling capabilities are over-provisioned to cope with worst-case scenarios. Additionally, these systems lack the support for variable cooling provisioning at run-time at server, rack, and datacenter level with respect to their influence on global energy consumption and environmental impact.

Cooling management needs to implement new cooling technologies, such as liquid or two-phase cooling, directly in the server itself, eliminating the poor thermal performing air as a coolant all together [31] [38][39][40], or relying on the use of outside cold air and/or water for cooling instead of electricity (i.e., free cooling [37]). It is also worth highlighting that the existing cooling control approaches only act on the cooling infrastructure, without coordinating both computing and cooling systems at server, rack, and datacenter level. Additionally, reuse of datacenter's waste heat in secondary applications instead of being disposed into the atmosphere must be taken into consideration.

That said, there is a clear research trend for a detailed study of these new cooling and energy recovery strategies, that will ultimately enable holistic mechanisms for multi-level heat control (server, rack, datacenter), as well as a tight integration of dynamic computing cooling needs and cooling systems power costs [41]. It should be possible to provide more efficient heat transfer solutions from the chips, memories, etc. by completely eliminating air as a means of heat transfer, while also reducing energy consumption for driving the cooling system and considering the waste heat recovery. Some examples of design and evaluation of these new cooling strategies can be found in [42-53].

6. CONCLUSIONS

Energy use is a central issue for data centers. The GreenDataNet EU Project aims to reduce the environmental impact of the data explosion allowing urban data centers to radically improve their energy and environmental performance. GreenDataNet intends to design, validate and demonstrate a system-level optimization solution allowing a network of urban data centers to collectively improve their energy and environmental performance, and act as a resource for smart grids.

The present document focused on the key emerging Information Technology (IT) trends and approaches toward more green and energy efficient datacenters. We presented and reviewed several approaches and techniques, dealing with and categorized in several domains, namely virtual machines management, energy and thermal management, data processing architectures and cooling infrastructure.

Starting from and leveraging on this knowledge, the GreenDataNet project will develop innovative technologies that will allow urban data centers to decrease their average Power Usage Effectiveness (PUE) from an average of 1.6-2.0 today to less than 1.3.