The MANGO FET-HPC Project: An Overview

José Flich*, Giovanni Agosta†, Philipp Ampletzer‡, David Atienza Alonso§,
Alessandro Cilardo¶, William Fornaciari†, Mario Kovač‖, Fabrice Roudet**, Davide Zoni†
*Universitat Politècnica de València, †DEIB – Politecnico di Milano, ‡Pro Design GmbH,
§ESL – École Politechnique Fédérale de Lausanne (EPFL), ¶Centro Regionale Information Communication Technology SCRL,
‖University of Zagreb, **Eaton Industries SAS
Email: *jflich@disca.upv.es, †name.surname@polimi.it, ‡philipp.ampletzer@prodesign-europe.com, §david.atienza@epfl.ch,
¶acilardo@unina.it, ‖mario.kovac@fer.hr, **FabriceRoudet@eaton.com

***Abstract*—In this paper, we provide an overview of the MANGO project and its goal. The MANGO project aims at addressing power, performance and predictability (the PPP space) in future High-Performance Computing systems. It starts from the fundamental intuition that effective techniques for all three goals ultimately rely on customization to adapt the computing resources to reach the desired Quality of Service (QoS). From this starting point, MANGO will explore different but interrelated mechanisms at various architectural levels, as well as at the level of the system software. In particular, to explore a new positioning across the PPP space, MANGO will investigate system-wide, holistic, proactive thermal and power management aimed at extreme-scale energy efficiency.**

***Keywords*—High Performance Computing, Customization, Energy Efficiency**

I. Introduction

High-Performance Computing (HPC) as we know it today is experiencing unprecedented changes, encompassing all levels from technology to use cases. On one hand, the escalating quest for performance/power efficiency is increasingly requiring deep application-based customization of the underlying computing architecture. On the other hand, new delivery models, such as outsourced and cloud-based HPC, are dramatically widening the amount and the type of HPC demand, posing both promising opportunities and big challenges for future HPC. In fact, cloud enables resource usage and business model flexibility, but it also deeply impacts architecture, as platforms must inherently support virtualization and be ready for large-scale capacity computing, serving many unrelated, competing applications with different workloads.

Moreover, HPC is increasingly faced with a QoS-sensitive computing demand, coming from that particular class of applications whose correctness depends on both performance and timing requirements, and the failure to meet either of them is critical. Examples of such time-critical applications include financial analytics, online video transcoding, and medical imaging. Such applications require some form of time predictability, in addition to performance and power efficiency. Time-predictability and QoS are relatively unexplored areas in HPC – traditional HPC focuses on throughput and resource usage optimization, in a trade-off with power-efficiency requirements. Extending the traditional optimization space, MANGO aims at addressing what we call the PPP space: *power*, *performance*, and *predictability*. In fact, predictability, power, and performance appear to be three inherently diverging perspectives on HPC.

In this scenario, the essential objective of MANGO is to achieve extreme resource efficiency in future QoS-sensitive HPC through ambitious cross-boundary architecture exploration. The research will investigate the architectural implications of the emerging requirements of HPC applications, aiming at the definition of new generation high-performance, power-efficient, deeply heterogeneous architectures with native mechanisms for isolation and QoS. MANGO will follow a disruptive approach challenging several basic assumptions, exploring new manycore architectures specifically targeted at HPC.

A. The MANGO Approach

The performance/power efficiency wall poses the major challenge faced nowadays by HPC. Looking straight at the heart of the problem, the hurdle to the full exploitation of today computing technologies ultimately lies in the gap between the applications' demand and the underlying computing architecture: the closer the computing system matches the structure of the application, the most efficiently the available computing power is exploited. Consequently, enabling a deeper customization of architectures to applications is the main pathway towards computation power efficiency. Theoretically, customization can enable improvements in power efficiency as high as two orders of magnitude, since it enables the computing platform to approximate the ideal intrinsic computational efficiency (ICE), defined as the energy consumption per operation achieved by purely computation circuits, e.g. FP adders.

The current uncertainty regarding on-chip HPC solutions and the essentially open nature of current architecture-level research will be regarded by MANGO as an opportunity, rather than a limitation. The fundamental intuition behind the project is that effective techniques for both performance/power efficiency and predictability ultimately share a common underlying mechanism, i.e., some form of fine-grained adaptation, or customization, used to tailor and/or reserve computing resources only driven by the application requirements. Along this path, the project will involve many different, and deeply interrelated, mechanisms at various architectural levels, from the heterogeneous computing cores, up to the memory architecture, the interconnect, the runtime resource management, power monitoring and cooling, also evaluating the implications on programming models and compilation techniques. In particular, to explore a new positioning across the PPP space, MANGO will investigate system-wide, holistic proactive thermal and power management aimed at extreme-scale energy efficiency by creating a hitherto inexistent link between hardware and software effects, which will involve all layers modeling in HPC server, rack, and datacenter conception. The combined interplay of the multi-level innovative solutions brought by MANGO will result in a new positioning in the PPP space, ensuring sustainable performance as high as 100 PFLOPS for the realistic levels of power consumption (< 15MWatt) delivered to QoS-sensitive applications in large-scale capacity computing scenarios. MANGO will provide essential building blocks at the architectural level enabling the full realization of the long-term objectives foreseen by the ETP4HPC strategic research agenda [1].

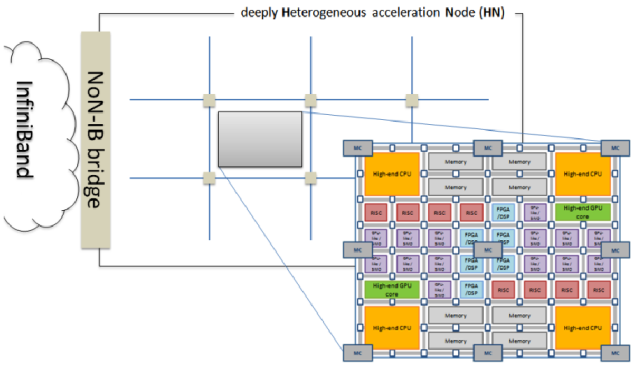

Fig. 1. MANGO Hardware Architecture

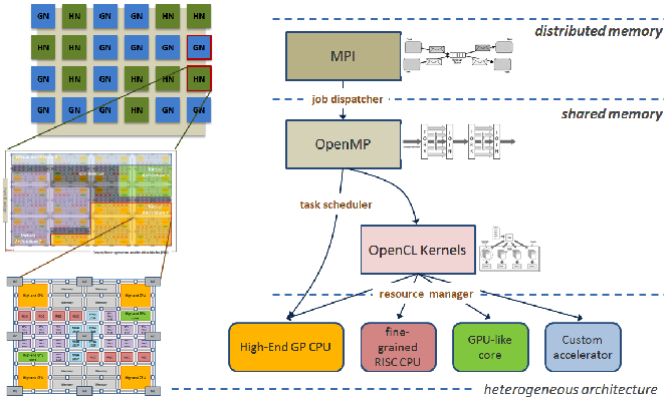

Fig. 2. MANGO Software Stack

B. Organization of the paper

The rest of the paper is organized as follows. In Section II we describe the targeted MANGO architecture. Then, in Section III we describe the programming models and runtime management resources to be used in MANGO and in Section IV the thermal and cooling innovations proposed in the project. Section V shows the prototyping roadmap for MANGO whereas Section VI briefly describes the application scenarios. Finally, the paper finishes in Section VII with some conclusions.

II. Hardware Architecture Concept

At the architecture level, the MANGO project foresees a scenario where General-purpose compute Nodes (GNs), hosting commercial-off-the-shelf solutions (e.g. Intel Xeon Phi processors or high-end NVIDIA GPU accelerators), coexist with *Heterogeneous Nodes* (HNs), forming a common HPC infrastructure. HNs, as depicted in Figure 1, will essentially be on-node clusters of next-generation manycore chips coupled with deeply customized heterogeneous computing resources. The manycore architecture will be open, it will not rely on COTS solutions available today, but rather it will enable broad-spectrum, ground-breaking research in the area of on-/off-chip architecture. Building on recent trends in HPC research, in fact, HNs will allow borrowing solutions from the embedded/System-on-Chip domain, which is now recognized as a promising pathway to extreme-scale low-power HPC. HNs will contain a multi-chip mesh of power-efficient RISC cores augmented with custom vector resources (SIMD and lightweight GPU-like cores) as well as a dedicated memory architecture and a custom Network-on-Chip providing advanced support for partitionability and time-predictability. The cores in the multi-chip manycore architecture will be connected through a *Network-on-Node* (NoN), forming a continuum at the off-chip (on-node) level from the on-chip interconnect.

Since the first stages of the project, the architecture exploration will be extensively supported by a purposely developed emulation platform. HNs will not be prototyped in a final ASIC form, but a mixed approach will be taken. In fact, RISC processors will be instantiated as ASIC cores tightly coupled with a large-scale reconfigurable hardware fabric used to emulate in near real-time the customized acceleration units, the advanced memory management architecture and the NoN, as well as the NoN bridge to the external interconnect. The platform will support fast design space exploration and validation of the solutions at both the software- and thermal/power-level. These techniques will inherently involve multiple aspects within the system, from programming down to the architecture definition, deeply intertwined with chip- and system-wide control mechanisms of physical parameters, primarily power consumption and temperature. To gain a holistic understanding of their impact on performance/power/predictability (PPP) and quantitative information about their effectiveness, MANGO will also develop a comprehensive toolset for PPP and thermal models, which will operate in close relation with the PPP run-time information collected from the platform.

The MANGO experimental platform will include 16 GN nodes with standard high-end processors, i.e. Intel Xeon E5, as well as NVIDIA Kepler GPUs, along with 64 HN nodes. GNs and HNs will be connected through InfiniBand. HNs will contain ASIC ARM cores and a high-capacity cluster of FPGAs used to emulate the rest of the HN system. The final HN infrastructure will contain dozens of manycore chips, and thus thousands of cores. The prototypical board will enable components to be easily plugged and removed and will allow different resource mixes, e.g. nodes highly populated with ARM cores and few high-end FPGAs (e.g. 192 + 64) or vice versa (e.g. 64 + 192), plus memory modules.

III. Programming Model and Runtime Management

To reach exascale parallelism, the programming model needs to be hierarchical, much like the runtime management system. Traditionally, the programming model for homogeneous HPC systems is based on a combination of MPI and OpenMP. When heterogeneity comes into the game, the programming model needs to be extended to allow the exploitation of hardware resources. OpenCL is an open standard for the development of parallel applications on a variety of heterogeneous multi-core architectures [2]. In MANGO, we aim at integrating the expression of new architectural features as well as QoS concerns and parameters within the existing stack of languages and libraries for extreme-scale HPC systems, by augmenting the runtime library APIs with new functions, as well as by introducing new pragmas or keywords to the language.

Figure 2 shows the MANGO programming model stack and its interaction with the underlying architecture and runtime software components. The programming models employ MPI to express inter-node computation, while at the node level OpenMP will be used to allow the expression of irregular applications, and OpenCL will serve as an intermediate language behind OpenMP, allowing the construction of virtual accelerators on the fly, collecting compatible cores not already allocated. The experience of the 2PARMA project [3] will help in this regard. The programming model will be integrated with the runtime management facilities, allowing job

dispatcher, task manager, and virtual device manager to interact with the corresponding three levels of the programming model.

The fine-grained virtualization mechanisms exposed by the MANGO deeply heterogeneous architecture will however pose new challenges for optimizing the performance and time-predictability of accelerated kernels. Building on previous results related to custom hardware-accelerated systems in the context of the syMParallel experimental toolchain and the HtComp project [4], [5], [6], MANGO will address the optimization of statically-predictable kernels based on the polyhedral model [7]. The activity will follow two different paths: a) characterizing kernels in isolation, inferring and controlling through suitable code-level transformations the patterns within the application that are relevant to power consumption and/or time predictability (e.g. memory access patterns and related choices in terms of memory partitioning and allocation); and b) exploring innovative techniques for analysing the interference between *multiple* applications running concurrently under QoS constraints [8].

Finally, taking steps from previous power-perfomance investigations considering accurate estimates for both the architacture and the actuators [9], [10], MANGO addresses in a holistic manner the concept of energy reduction considering both computing energy and cooling efficiency as primary goal, thus combining fine-grained monitoring of energy, temperature and power in servers and racks, but also optimization of the mechanical cooling part to use two-phase cooling at rack level.

First of all, MANGO proposes to extend for HPC servers the latest state-of-the-art works on semi-analytical thermal modeling approaches to enable the fast calculation of power/thermal figures of servers under dynamic workload behaviours [11]. The control/optimization policies in the MANGO resource management will be able to evaluate the QoS and non-functional requirements of the applications, in a hierarchical, system-wide multi-objective optimization going beyond electronics to mechanical aspects (e.g., liquid cooling pump control) [12]. The collected information from monitors enables the prediction of temperatures in the different parts of the servers and racks, which will be passed to the hierarchic runtime manager, structured in a hierarchical architecture exploiting both OS and hypervisor levels, which will be able to tune the system knobs (P-states, fan control, tasks assignments, etc.), to mitigate performance variability [13]. Overall, we will target to exploit the run-time thermal-power predictions to mitigate performance variability due to thermal emergencies under highly dynamic workload variations [14], as it is one of the key challenges in the highly heterogeneous MANGO computing server architecture.

IV. Thermal and Cooling Innovations in MANGO

MANGO will extend the experience acquired in the latest research on advanced compact modeling for liquid-cooling monitoring [15] to explore the time constants of thermal and energy control knobs to develop next-generation cooling technologies for HPC systems. In particular, we will explore the use of a novel passive thermosyphon (gravity-driven) cooling technology that will attempt to include multiple *parallel heat sources at multiple elevations* to eliminate energy consumption. Thus, in MANGO we will carry out preliminary evaluations for the first time in HPC system to re-convert generated heat into electricity at chip level by exploiting the use microfluidic fuels cells combined with the liquid cooling technology [16]. The preliminary testbeds to evaluate both thermosyphon and energy-recovery process through micro-fluidic fuel cells will be developed and measured in the facilities of the EPFL partner. The objective is to evaluate the creation of HP servers and rack cooling technologies that can re-use part of their waste heat to generate electricity that reduces external power supply needs.

V. The MANGO Platform Roadmap

The MANGO strategy for building an effective largescale emulation platform will be articulated in three phases.

a) Phase 1 – Stand-alone single-board emulator: The research activities involving architecture exploration will initially rely on current available hardware made of a standalone emulation platform based on FPGA devices and a general purpose node. The standalone emulator will be based on a modular and scalable approach, with several FPGAs being assembled on dedicated daughter modules plugged on a common motherboard. The motherboard will give complete access to all available I/Os of the FPGA, leaving maximum freedom regarding the FPGA interconnection structure, which will allow to define the HN interconnect. The proFPGA quad V7 system provided by ProDesign as a standalone emulation platform will be used. The board is equipped with three Xilinx Virtex 7 XCV2000T FPGA modules and one Zynq module, containing a dual core ARM processor as well as reconfigurable hardware fabric to prototype external subsystems, handling up to 48 M ASIC gates alone in one board. Several proFPGA systems will be interconnected enabling the full HN infrastructure to be implemented. Due to the fact that multiple proFPGA quad or duo systems can be stacked or connected together, scalability is ensured. The highspeed boards together with the specific high speed connectors allow a maximum point to point speed of up to 1.8 Gbps over the standard FPGA I/O and up to 12.5 Gbps over the MGT of the FPGA.

b) Phase 2 – From FPGA stand-alone board to a dedicated chassis: A new board for HPC will be implemented compling with the physical constraints of HPC and datacenter racks, considering as well requirements for cooling and power supply researched within the project. The board will be extended to deliver further number of daughter boards and will provide proper connectivity through optical links to other boards. Pin-to-pin connectivity between FPGAs (either at the same board or at different boards) will allow expandability and scalability. This enables MANGO to explore future chip configurations in a predictable and accurate manner. Daughter boards will be extensible and open to new developments, particularly to new 64-bit ARM cores or even more advanced solutions like the hybrid Xeon E5+FPGA chip recently announced by Intel. In this phase, the HN interconnect will be applied to the set of HN nodes (the board) developed. It will embrace connectivity at the board level, between ARM and FPGA modules, inside the FPGA modules (within the accelerators and RISC processors implemented), and between the boards. This means a single and unified interconnect will be designed for the overall HN infrastructure (made of 64 nodes).

c) Phase 3 – Rack assembly: As a final phase, the complete rack will be implemented and populated of GNs and HNs. The system will enable a large-scale platform used to reproduce in near real-time the behavior of the MANGO manycore architecture. The full platform will consist of a rack collecting up to 16 blades equipped with high-end CPUs, e.g. Intel Xeon chips, and GPUs, mounted on the motherboard, as well as 64 HN nodes. A custom backplane will provide connectivity across the blades, both through standard bridges and using pin-to-pin connections across the FPGA chips, effectively providing a single large-scale reconfigurable hardware fabric used to emulate the fine-grained accelerator tiles envisioned in the MANGO architecture. The inter-FPGA pin-to-pin backplane interconnection will be reconfigurable on-field, providing a large degree of flexibility for the emulation of the on-chip network interconnect.

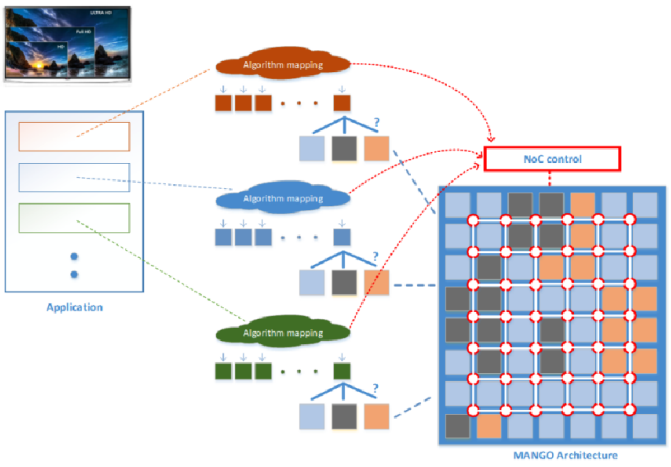

Fig. 3. Mapping Applications on the MANGO Platform

VI. Application Scenarios

The combined interplay of the above techniques will have a decisive impact on the positioning of MANGO in the PPP space. To show this impact, the MANGO HPC architecture will be validated through application platforms with stringent QoS and high-performance requirements. This will require an exploration of mapping algorithm parts to various heterogeneous tiles (CPUs, GPUs, RISC processors, accelerators), and the exploitation of novel algorithms and architectures for the interconnect to enable more efficient data flow between processing elements. Network configuration and data flow are extremely important since the communication overhead can significantly reduce efficiency of processing tiles. Only combination of both can result in the efficient implementation needed for real time operation as shown in Figure 3. The target application platforms are online video transcoding, medical imaging, real-time signal/video processing, and security services for cloud-based systems.

Video transcoding refers to the problem of adapting on-the-fly Internet video contents based on user's device features or specific operational conditions. The importance of the transcoding application and its efficient implementation of future exascale HPC systems can be seen from following reports. By 2018, global IP traffic will reach 1.6 zettabytes per year, or 131.6 exabytes per month. Globally, IP video traffic will be 79 percent of all consumer Internet traffic in 2018, up from 66 percent in 2013. This percentage does not include video exchanged through peer-to-peer (P2P) file sharing. The sum of all forms of video (TV, video on demand [VoD], Internet, and P2P) will be in the range of 80 to 90 percent of global consumer traffic by 2018 so this volume of data will not be possible to process without exascale, MANGO like HPC systems. Adaptation involves video properties, such as spatial, temporal and amplitude resolution or even video codec.

Medical imaging applications increasingly involve 3D images, the rendering of which is highly memory intensive and sensitive to optimization through specialized processing units [17]. Latency requirements are varied, with critical operations having strict real-time constraints, and background tasks having more relaxed requirements.

Sensor data processing is often required to deliver timely information from sensor array – e.g., in ground radar systems or in surveillance or critical infrastructure monitoring systems. Algorithms in this field exhibit both data-dependent and data-independent computation.

Security services for cloud-based systems are needed to ensure isolation of virtual machines, protection of data communication, and intrusion detection. Such services will be offered in MANGO through an HPC system acting as a front-end for virtualized systems.

VII. Conclusions

The MANGO project starts in October 2015 and is funded by the European Commission under the Horizon 2020 FET-HPC program. The project will last for three years, with the goal of addressing power, performance and predictability in HPC.

References

[1] European Technology Platform For HPC, "ETP4HPC Strategic Research Agenda: Achieving HPC leadership in Europe," http://www.etp4hpc.eu/strategy/strategic-research-agenda/, 2013.

[2] Khronos Group, "The Open Standard for Parallel Programming of Heterogeneous Systems," https://www.khronos.org/opencl/, (retr. Jul 2015).

[3] C. Silvano, W. Fornaciari, S. Reghizzi, G. Agosta, G. Palermo, V. Zaccaria, P. Bellasi, F. Castro, S. Corbetta, A. Di Biagio, E. Speziale, M. Tartara, D. Melpignano, J.-M. Zins, D. Siorpaes, H. Huebert, B. Stabernack, J. Brandenburg, M. Palkovic, P. Raghavan, C. Ykman-Couvreur, A. Bartzas, S. Xydis, D. Soudris, T. Kempf, G. Ascheid, R. Leupers, H. Meyr, J. Ansari, P. Mahonen, and B. Vanthournout, "2PARMA: Parallel Paradigms and Run-time Management Techniques for Many-Core Architectures," in *VLSI 2010 Annual Symposium*, ser. LNEE. Springer NL, 2011, vol. 105, pp. 65–79.

[4] A. Cilardo and L. Gallo, "Improving multibank memory access parallelism with lattice-based partitioning," *ACM Trans. Archit. Code Optim.*, vol. 11, no. 4, pp. 45:1–45:25, Jan. 2015.

[5] A. Cilardo, L. Gallo, and N. Mazzocca, "Design space exploration for high-level synthesis of multi-threaded applications," *Journal of Systems Architecture*, vol. 59, no. 10, pp. 1171–1183, 2013.

[6] A. Cilardo and L. Gallo, "Interplay of loop unrolling and multidimensional memory partitioning in hls," in *Design, Automation Test in Europe Conference Exhibition (DATE), 2015*, March 2015, pp. 163–168.

[7] M.-W. Benabderrahmane, L.-N. Pouchet, A. Cohen, and C. Bastoul, "The Polyhedral Model Is More Widely Applicable Than You Think," in *Compiler Construction*, ser. LNCS, R. Gupta, Ed. Springer Berlin Heidelberg, 2010, vol. 6011, pp. 283–303.

[8] D. Zoni, S. Corbetta, and W. Fornaciari, "Thermal/performance trade-off in network-on-chip architectures," in *System on Chip (SoC), 2012 International Symposium on*, Oct 2012, pp. 1–8.

[9] D. Zoni, F. Terraneo, and W. Fornaciari, "A dvfs cycle accurate simulation framework with asynchronous noc design for power-performance optimizations," *Journal of Signal Processing Systems*, pp. 1–15, 2015.

[10] D. Zoni, F. Terraneo, and W. Fornaciari, "A control-based methodology for power-performance optimization in nocs exploiting dvfs," *Journal of Systems Architecture*, vol. 61, no. 5-6, pp. 197 – 209, 2015.

[11] A. Sridhar, M. Sabry, and D. Atienza, "A Semi-Analytical Thermal Modeling Framework for Liquid-Cooled ICs," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 33, no. 8, pp. 1145–1158, 2014.

[12] A. K. Coskun, D. Atienza, M. Sabry, and J. Meng, "Attaining Single-Chip, High-Performance Computing Through 3D Systems with Active Cooling," *IEEE Micro Magazine*, vol. 31, no. 4, pp. 63–75, 2011.

[13] J. Kim, M. Sabry, D. Atienza, K. Vaidyanathan, and K. Gross, "Global fan speed control considering non-ideal temperature measurements in enterprise servers," in *Proc. IEEE/ACM Design, Automation and Test in Europe (DATE '14)*, Dresden, DE, 2014, pp. 303–308.

[14] J. Kim, M. Ruggiero, D. Atienza, and M. Ledergerber, "Correlation-Aware Virtual Machine Allocation for Energy-Efficient Datacenters," in *Proc. IEEE/ACM Design, Automation and Test in Europe (DATE '13)*, Grenoble, FR, 2013, pp. 216–221.

[15] A. Sridhar, A. Vincenzi, M. Ruggiero, and D. Atienza, "Neural network-based thermal simulation of integrated circuits on gpus," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 31, no. 1, pp. 23–36, 2012.

[16] M. Sabry, A. Sridhar, P. Ruch, D. Atienza, and B. Michel, "Integrated Microfluidic Power Generation and Cooling for Bright Silicon MPSoCs," in *Proceedings of the IEEE/ACM Design, Automation and Test in Europe (DATE '13)*. Dresden, Germany: IEEE-ACM Press, 2014, pp. 10–15.

[17] M. Kovač, "E-Health Demystified: An E-Government Showcase," *Computer*, vol. 47, no. 10, pp. 34–42, Oct 2014.