The Essence of System-Level Design: Collaboration and Interdisciplinarity
A story of many research centers

Alberto Sangiovanni-Vincentelli,[1] Giovanni De Micheli,[2] Sharad Malik,[3],Naresh Shanbhag[4] and Jan Rabaey[1]

Abstract

Major technological innovation can arise either from breakthrough discoveries or from a visionary understanding of what becomes possible as technology evolves—combined with a methodology that leverages those breakthroughs. While breakthroughs often emerge from deep, focused exploration, visionary innovation typically relies on a meet-in-the-middle approach that integrates diverse disciplines to achieve an overarching goal. This is the essence of system-level design. By its very nature, it demands collaboration among researchers with varied expertise, usually brought together within organizations such as centers or institutes to pursue a common mission. To illustrate this point, we highlight several successful approaches in which Jan Rabaey has played a central role.

Introduction

Traditional research in the humanities, social sciences, and even many scientific fields is typically driven by a single investigator, sometimes supported by assistants, who defines and tackles specific research questions. Engineering research tends to be different. It often demands collaboration across areas of expertise to develop solutions that no individual researcher could achieve alone. This challenge is especially evident in electronic system design, where progress depends on connecting high-level application demands with the detailed realities of devices and interconnect technologies. Bridging these layers calls for a coordinated, meet-in-the-middle approach.

Academic environments are not always well-suited for this kind of work. Research structures typically prioritize individual achievement, and few institutions have leading experts spanning all the domains required for system-level design. Making progress in this space depends on strong, forward-looking leadership in both funding and research strategy. It also requires the creation of research centers with clear system-level objectives that unite researchers across disciplines and even across different organizations.

---

[1] University of California, Berkeley

[2] EPFL

[3] Princeton University

[4] University of Illinois at Urbana-Champaign

This paper will examine several such centers established beginning in 1999, highlighting Jan Rabaey's pivotal role in shaping their goals, strategies, and achievements.

The Gigascale Research Center (GSRC) – 1999-2013

GSRC was one of the two centers launched in 1999 by the visionary Focus Center Research Program (FCRP), a collaboration between the US Government and the US microelectronics industry through the MARCO Consortium (Microelectronics Electronic Research Consortium). The stated goal of the FCRP was to enable disruptive research in a multi-university setting. The center went through three phases led respectively by Richard Newton (1999-2003), who set the initial vision for the center, Jan Rabaey (2004-2009) and Sharad Malik (2010-2013).

Phase 1 (1999-2003): Addressing the Productivity Gap

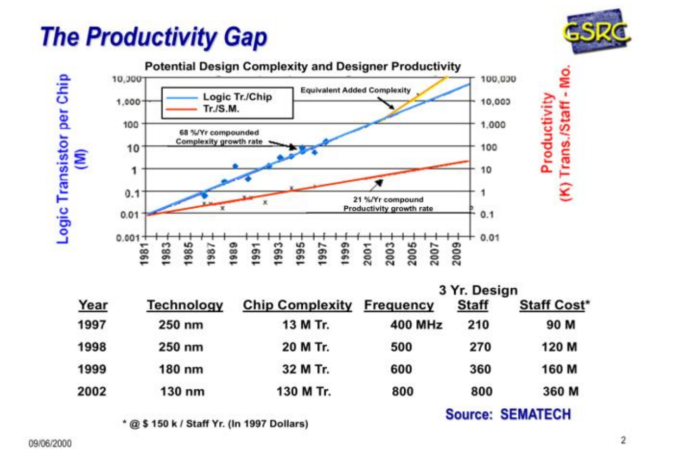

Figure 1. The productivity gap). Slide from an early GSRC Review Meeting (9/6/2000)

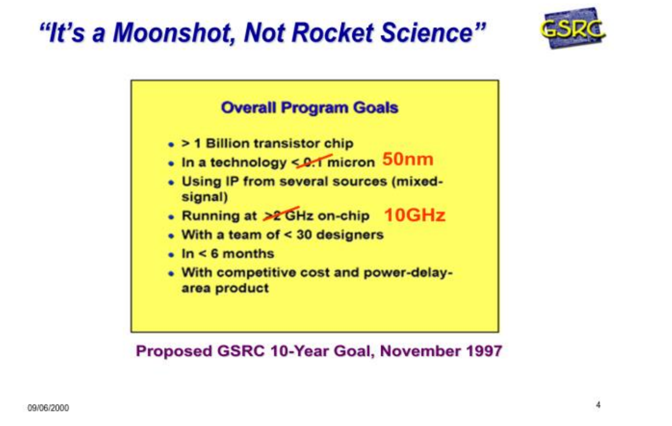

Figure 2. The GSRC moonshot vision

The GSRC center was launched with the primary goal of addressing the productivity gap challenge – that the number of available transistors per chip was growing much faster than designer productivity (51% CAGR vs. 21% GAGR, Figure 1). As formulated by its founding center director, the center aimed at a very specific 10-improvement in design productivity for cutting edge designs. Labeled as a **moonshot,** it contrasted to the common approach of just focusing on a set of topic areas (Figure 2). This moonshot view emphasized the need for collaborative research — it would take a tight-knit team effort to accomplish its ambitious goals. Also critical was the formulation of several of focus themes. The *Component/Communication theme* was led by Alberto Sangiovanni-Vincentelli and recognized the importance of communication – for data at rest (in memory) and data in motion (through on-chip communication) – in a departure from the classical focus on computation. *The Fully-Programmable Systems* theme led by Kurt Keutzer recognized the growing importance of systems supporting application domains through programmability and thus the importance of programming interfaces. At the lower levels of the stack, the *Constructive Fabrics* theme led by Larry Pileggi focused on connecting emerging technologies – both devices and interconnect – to the layers above for effective system-level design.

Figure 3. Some GSRC Phase 1 Accomplishments (Slides from 11/22/2000 and 9/19/2002 Review)

Several operational aspects contributed to the center's collaborative mission. It held in-person quarterly workshops that enabled all the center faculty and students to share results, connect their projects and interact. Also important to the center's operations was a close collaboration with the industrial partners to provide reality checks, that is "to obtain the data and insights essential to establishing the ground truths on which new perspectives must be built." This included inviting the industry partners to the quarterly workshops, an industrial advisory board, and active outreach through regular industry visits.

During this phase several important concepts in system-level design were established through GSRC research. The concept of platform-based design, as proposed by Alberto Sangiovanni Vincentelli: the platform being the System-on-a-Chip level analog of a processor; a design discipline for platform-based design; and the need to bring in the application domain as an integral part of the design flow [1]. These principles are now considered standard practice.

Phase 2 (2004-2009): Addressing System-Level Design in the Late-Silicon Era

Another important principle guiding the FCRP centers was the mandate for renewal. Centers are created with a life span of 5 years, after which the dissolve ("declare success") or go through a complete revision. The latter happened to the GSRC. With a change in leadership (from Richard Newton to Jan Rabaey) came also a renewed vision, as defined through intensive brainstorm meetings (Figure 4).

Figure 4. GSRC brainstorming session in Half Moon Bay (CA) in 2002. From left to right: Larry Pileggi, Jan Rabaey, Gary Baldwin, Tim Cheng, Alberto Sangiovanni-Vincentelli, and Andrew Kahng and Janie Irwin.

Phase II of GSRC addressed new challenges, especially those that were emerging in late-silicon design — that is, the end of the scaling era (Figure 5). One such challenge was how to map the concurrency inherent in the emerging applications to the concurrency offered by the next-generation platforms, especially triggered by the emergence of AI for applications such as computer vision. Kurt Keutzer led the effort considering both SOCs and highly concurrent platforms such as GPUs. This research was one of the first to recognize the importance of interfaces such as CUDA for GPUs.

Other efforts included: (1) *Communication-based design* - A complete design automation flow was envisioned for Networks on Chips (NoCs), including specification, synthesis, and verification [2] , was realized by Luca Benini and Giovanni De Micheli [3]; (2) *System resiliency* grew as a first-order concern due to shrinking geometries and device failures; and (3) *Alternative computational models* – statistical and distinct from the deterministic error-free von Neumann model – to enable ultra-low power computing (with Shanbhag, Verma and Rabaey as the main proponents).

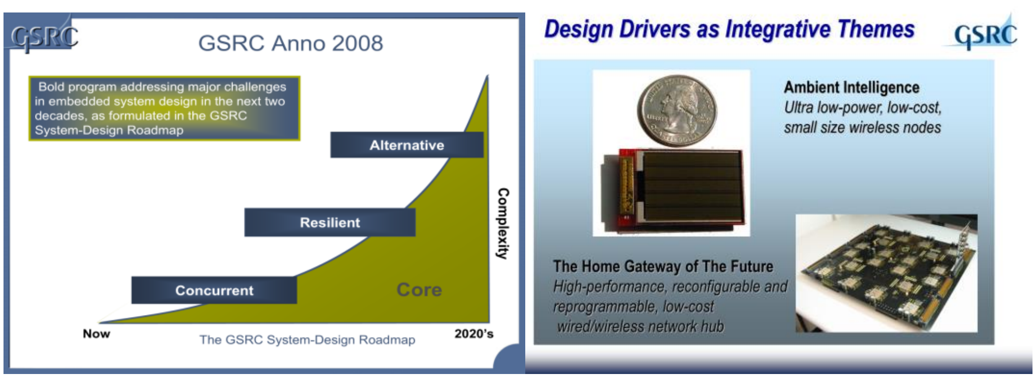

Figure 5. GSRC Phase II research themes (slides from 2004 and 2008).

Similar to the moonshot approach of Phase I, a couple of “design drivers” provided goal orientation, serving both as application-inspiration as well as reality check for the various deep-dive research efforts (Figure 5b).

Phase 3 (2010-2013): Application-Driven Platform Design

In its final incarnation, this time led by Sharad Malik, the focus of GSRC shifted to the deployment of *application-driven specialization* in the design and deployment of platforms to meet their power-performance-cost goals. This came with two distinct implications – the first was specialization in the compute, communication and memory architectures to match application needs, and the second was system-level algorithm-architecture co-design (Figure 6).

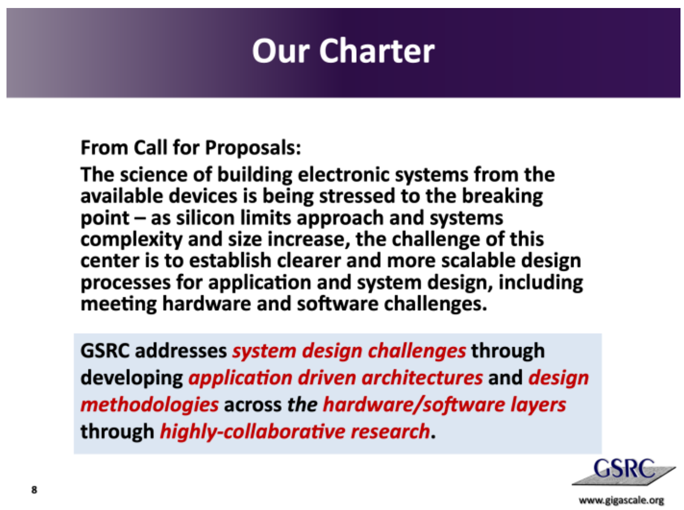

Figure 6. *Focus on Platform Design (Slide from 10/16/2012)*

The importance of *specialized accelerators* (as was demonstrated by David Brooks and his colleagues in the center) is well understood today. The unique contribution of the GSRC research was to approach accelerator design holistically, addressing not only computation but also the memory and communication subsystems needed to sustain it. Under the leadership of Wen-Mei Hwu and Demin Cheng, Programming models and tools were treated as essential components of the overall system design process. This distinguished it from the large body of compiler research in this area.

The applications- driven architecture setting of the center further enabled the exploration of radically new alternative computational models that take a statistical approach to information processing and are “non von Neumann” Instead of treating variations in the underlying circuit fabric as a problem to eliminate, this approach aligns those statistical properties with the inherent statistical requirements of the application. Naresh Shanbhag and his collaborators demonstrated several compelling successes using this methodology.

From PicoRadios over Cyber-Physical Systems to TerraSwarms

Simultaneous to the creation of the GSRC in 1999, another development with long-term impact was taking place in Berkeley. With the conclusion of the successful Infopad project (see “*Making Dreams Come True*”), Bob Brodersen and Jan Rabaey began considering the next steps on how to preserve the strong academia-industry collaboration that had been vital to its success. Wireless systems, particularly those relying on CMOS technology, were starting to emerge but were still at an early stage, creating a prime opportunity for bold, innovative research. Their answer was to establish a new type of academic research center: one that encouraged open collaboration, was supported primarily by industry, and focused on unconventional and forward-looking wireless challenges . A space in downtown Berkeley was leased and renovated ("a great environment enables great research)", and in January 1999 the Berkeley Wireless Research Center (BWRC) was born.

It grew into a magnet for innovation and rapidly built a reputation for leadership in advanced integrated circuit design. To avoid argumentation about intellectual property between the various partners, a no-IP public-domain policy was adopted. This served the center well and did not impede startup creation and technology adoption (as some had feared). The key to the success of the center were forward-looking visions combined with collaborating faculty and students from diverse backgrounds. Early center directors besides Brodersen and Rabaey included Paul Gray, Paul Wright, David Tse, Ali Niknejad, Bora Nikolic, Vladimir Stojanovic and Elad Alon). From the start, three main directions were identified (Figure 7) each of which went on to become major programs and success stories.

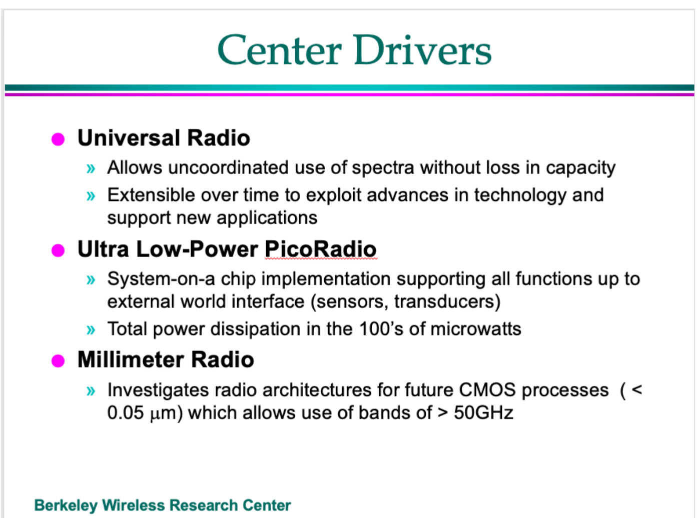

Figure 7. BWRC vision (1999).

To support the design process of these complex circuits, a design methodology called "Chip-in-a-Day" promoted complete design flows from high-level descriptions and automated hardware generation. The myriads of students graduating from the center went on to become leaders in the industry or create their own startup's (such as Atheros Communications, SiBeam, BeeCube, Blue Cheetah, Ayar Labs, Cortera Neurotechnologies and Neuralink, just to name a few) (Figure 8).

Today, BWRC is still going strong exploring novel directions, featuring a new generation of faculty and students. The vision has expanded going well beyond wireless including wired links, optical communications, biomedical and neurotech devices and advanced system integration. The soul of the center has not changed however and remains focused on providing leadership in advanced integrated circuit design.

Figure 8. BWRC students, faculty and staff at the center (2017)

The early pioneering efforts at BWRC (and Berkeley EECS) in the creation of low-power nodes for wireless sensor networks quickly morphed into a broader vision in which sensor and actuator nodes permeate the physical world and form a bridge between the physical and cyber world (appropriately called "cyberphysical systems"). The prospect of having "swarms" of distributed sensor and actuator nodes surrounding us to the count of 100's of nodes per person raised a broad set of questions and challenges: ad-hoc networking, locationing, programming paradigms, dynamic system management, application mapping, user interaction, reliability, etc. A new community came together almost overnight. In Berkeley, we responded by creating another center called the "SwarmLab" with start-up funding provided by Qualcomm in 2011. Rather than focusing on the components as BWRC did, the SwarmLab vision was to address the system challenges driven by innovative cyberphysical applications and user paradigms. It brought together faculty with very diverse backgrounds ranging from system modeling to innovative user interfaces. Just to name a few: Edward Lee, Kris Pister, Michel Maharbiz, John Kubiatowicz, Bjoern Hartmann and Eric Paulos. One of the returning tenets was that meaningful system design is only possible when it is accompanied by real prototyping. Here, the Berkeley InventionLab (a maker facility) played a crucial role (Figure 9).

Figure 9. Testing and demonstrating wearable devices at SwarmLab bootcamp (2014).

Given the all-encompassing nature of the mission, it was quickly realized that confinement to Berkeley was limiting the reach and impact. In a next step, a proposal was written and granted for a new multi-university research center under the SRC/DARPA StarNet program (2013).[5] The resulting TerraSwarm center, founded in 2013 with Edward Lee as director, brought together researchers from 9 universities across the United States, charting new grounds in collaborative interdisciplinary system research (Figure 10).

Figure 10. Terraswarm Vision 2025 (2013).

Rethinking Computing - The SONIC Starnet Center (2013-2018)

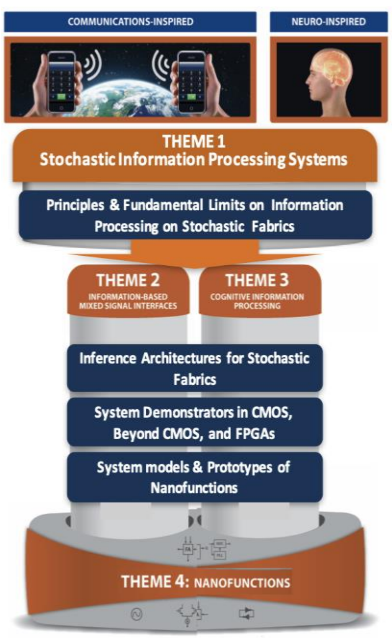

Figure 11. SONIC's vertically-integrated systems-driven research agenda to explore statistical models of computation on nanoscale circuit and device fabrics.

The success of GSRC's Alternative Models of Computation research theme from 2008-13, motivated the establishment in 2013 of yet another center under the StarNet program header, the **S**ystems **O**n **N**anoscale **I**nformation fabri**C**s (SONIC) center. Its mission was to explore Shannon and neuro-inspired statistical models to enable the design of computing systems at the limits of energy efficiency, throughput, and robustness. Led by Naresh Shanbhag, SONIC's research agenda (see Figure 11) was architected to be vertically integrated spanning:1) systems and algorithms (led by Andrew Singer); 2) information-based mixed-signal architectures (led by Boris Murmann); 3) neuro-inspired cognitive systems (led by Jan Rabaey and Naveen Verma); and 4) nanofunctions (led by David Blaauw and Philip Wong).

[5] StarNet was a follow-up to the FCRP program that funded GSRC.

The center brought together researchers from diverse backgrounds including information and communication theory, integrated circuit design and architectures, electronic design automation, and nanoscale devices. Tying these disparate domains together was the notion that it is the information content in the AI workloads that really matters and needs to be preserved across the compute stack in the presence of noise and other nanoscale non-idealities. The idea of allowing errors in the lower levels of the compute stack to be compensated at the upper levels was key to achieving efficiency, accuracy, and robustness, e.g., SONIC's device researchers focused on demonstrating nanofunctions rather than individual devices, while its systems researchers developed error compensating algorithms and architectures. Many system-level concepts were translated into circuit prototypes both in silicon and beyond (Figure 12).

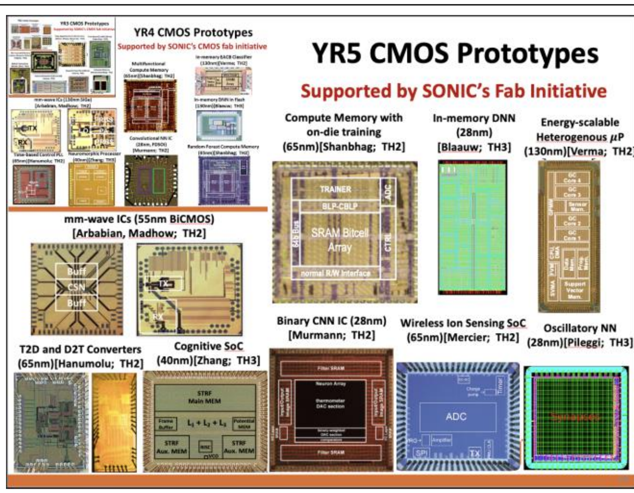

Figure 12. Prototypes in silicon (and beyond) demonstrating the richness of computational approaches.

Two of the key SONIC research outcomes were SRAM-based in-memory computing (IMC) and hyperdimensional computing (HDC) both of which are active areas of research to this day. In each of these outcomes, there was a strong alignment and cross-optimization across systems, architectures, circuits, and devices. SONIC's vertically integrated research agenda validated the importance of connecting systems to devices and became an exemplar for next generation semiconductor focused research centers.

With today's computational demands of AI workloads coupled with challenges in semiconductor scaling, the role of cross-stack system design has become even more important. Furthermore, AI systems, with their statistical metrics of functionality, are eminently suited for realization using Shannon and neuro-inspired models of computation. Jan's pioneering leadership through GSRC, SONIC, and other centers has charted a tangible path forward that positions us well to meet these challenges.

NanoTera - Exploring New Frontiers of Cyber-physical Systems

In parallel with the creation of Terraswarm, Giovanni De Micheli – now at EPFL in Lausanne – opened a new chapter in collaborative research as an evolution of the aforementioned projects and centers. The *Nano-Tera.ch* program [4] started in Switzerland around 2010, funded directly by the Swiss Federal Government, addressing health and environmental systems. It involved many local researchers (from EPFL, ETHZ, CSEM and the Universities of Geneva, Neuchatel, Basel and Lugano) and advisers from all over the world, including Jan Rabaey. A key aspect of the program was the formation of multi-institution and multi-discipline teams for collaborative research spanning from technology to system – a continuation of the spirit of GSRC broadened to the domain of biotechnology.

The broader context here is considering living beings that are biological systems that interface to the natural environment, via sensing physical, chemical and biological stimuli and reacting to them. The embedding of systems within a living environment requires them to incorporate new sensors and actuators. Two motivations for exploring this extended system vision are *health monitoring* and *environmental protection* (Figure 13). Systems that promote bettering health can reach the combined goals of achieving higher quality services and lower cost of operation Environmental protection includes, for example, monitoring air and water pollution at a fine grain and designing means to prevent disasters as triggered by rock or ice movement.

Looking beyond, one of the most challenging scientific tasks is to understand the brain, this wonderful natural computer that performs many complex functions with only 20 W of power. The brain has inspired new computational paradigms, some of which are extremely powerful as accelerators of computation. Jan Rabaey's book "Of Brains and Computers" [5] addresses the comparison of computing in the brain and in machines. It describes computing functions in nature and their evolution through centuries. It draws interesting comparisons with computing machines that humans have designed while leveraging inspirations from living systems to different extents. It shows the convergence of natural and artificial computing as measured by different parameters but highlights how natural computing is by far more energy efficient. Eventually it mentions research efforts that aim at combining living brain functions with electronic systems. Overall, it sheds light on new ways of conceiving computing systems.

Figure 13. Some outcomes of the Nano-Tera Program. (a) Autonomous swimming robot searching for pollutants in fresh water. (b) Probe of a portable 3D ultrasound system for remote diagnosis.

The ultimate question is whether living organisms and computing systems will merge. The partial, but ever increasing, *observability* and *controllability* of living organisms is a key aspect of the convergence of life and computing. At present, cyber-medical implants and systems are used primarily to address physical impairments. The ultimate question is whether living organisms and computing systems will merge (Figure 14). It is reasonable to imagine that future electronic devices, closely integrated with biological functions, will enhance human capabilities. In such a world, the distinction between a system embedded in a person and the person themselves may disappear, since decision making would be shared between biological cognition and artificial processing. The central challenge ahead is ensuring that every individual retains the ability to make informed decisions as these technologies continue to advance. Clearly a wonderful topic for forward-looking research centers …

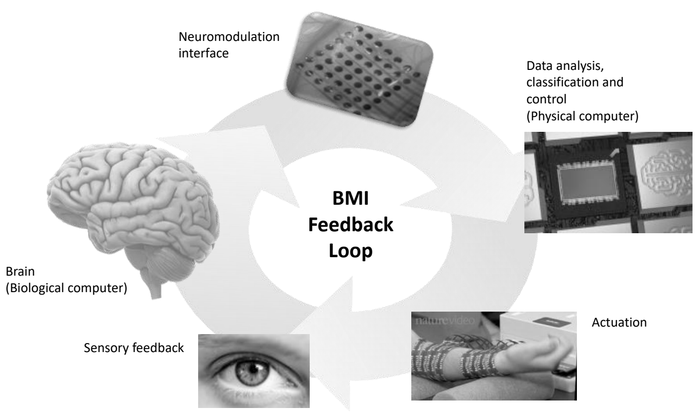

Figure 14. Brain-Machine Interface (BMI) feedback loop. Sensors acquire neural signals from the brain. Outcomes resulting from interpretation are directly or indirectly (e.g. through motor function) fed back into the brain. The feedback loop combines both a biological and a physical computer. (from "Of Brains and Computers") [5]

Looking forward

The evolution of electronic system-level design over the past decades has been nothing but fascinating and even spectacular. Today we are putting together complex systems at the warehouse as well as the microscale levels that are fully operational from the get-go. Integrated circuits containing 100s of billions or combining very diverse technologies are a reality today, and the design thereof does not take many 1000s of engineers as was feared in the late 1990s. The adoption of advanced design methodologies as originally envisioned by academia has surely contributed hugely to this success. Yet, moving forward will require even more radical innovation in system design. Possible showstoppers are complexity (again), limits of semiconductor scaling, power bounds and innovative applications with unforeseen needs (think brain digital twins, for example). Addressing those will need initiatives with a long-term perspective such as the FCRP (or StarNet) programs as well as industrial support to continue or even scale up. It also asks for taking risks and exploring possibly disruptive pathways. Without a doubt, AI will play a major role in ways unforeseen today. The future looks exciting for sure.

REFERENCES

[1] K. Keutzer, S. Malik, R. Newton, J. Rabaey and A. Sangiovanni-Vincentelli, System Level Design: Orthogonalization of Concerns and Platform-Based Design, IEEE Transactions on Computer-Aided Design of Integrated Circuits & Systems, vol.19, no.12, Dec. 2000, pp.1523-43.

[2] Sgroi M, Sheets M, Mihal A, Keutzer K, Malik S, Rabaey J, Sangiovanni-Vincentelli A. "Addressing the system-on-a-chip interconnect woes through communication-based design." Proceedings of the 38th Design Automation Conference, Las Vegas, June 2001, pp.667-72.

[3] L. Benini and G. De Micheli, “Networks on chips: a new SoC paradigm,” Computer, Volume: 35, Issue: 1, January 2002, pp 70-78.

[4] A. Leblebici, P. Mayor, M. Rajman and G, De Micheli, “Nano-Tera.ch: Engineering the Future of Systems for Health, Environment and Energy,” Springer, 2019.

[5] Jan M. Rabaey, “Of Brains and Computers”, Foundations and Trends® in Integrated Circuits and Systems: Vol. 2, No. 1–2, 2022.