Executive Summary of the Visioning Workshop

on

# Nanotechnology-inspired Information Processing Systems of the Future

August 31-September 1, 2016

Sponsored by the Computing Community Consortium (CCC)

## Executive Summary

It is undeniable that computing and associated technologies have transformed our lives, and will continue to do so in the foreseeable future. Computing systems have become drivers for future growth in sectors such as health (continuous health monitoring), transportation (self-driving vehicles), energy (smart buildings), finance, education, to entertainment, leisure, and overall wellness. These systems will become so pervasive that they will define the human experience itself.

Nanoscale semiconductor technology has been a key enabler of the computing revolution. It has done so via advances in new materials and manufacturing processes that resulted in the size of the basic building block of computing systems – the logic switch and memory devices – to be reduced into the nanoscale regime. In this process, nanotechnology has provided increased computing functionality per unit volume, energy, and cost.

Today, this symbiotic relationship between semiconductor technology and computing is undergoing a major upheaval both at the device technology level and the application levels. At the device technology level, traditional scaling of device sizes has slowed down and the reduction of cost per transistor via pure geometric scaling of process technology is plateauing. Simultaneously, at the application level, new computing workloads have called for a migration from an "algorithmic" compute world dominated by Turing-inspired processes to a "learning-based" information processing paradigm. This shift is driven by the convergence of an abundance of data and the computing resources needed to process it in application spaces that span the Cloud, mobile, and the Internet of Thing (IoT).

In order for computing systems to continue to deliver substantial benefits for the foreseeable future to society at large, it is critical that the very notion of computing be examined in the light of nanoscale realities. In particular, one needs to ask what it means to compute when the very building block – the logic switch – no longer exhibits the level of determinism required by the von Neumann architecture. What does it mean to compute when information extraction dominates over raw data processing? What are the fundamental limits of computing in this new era? Indeed, given the reliance of major industry sectors on the continued growth in the capabilities of computing and information processing systems, the future economic growth, global competitiveness and national security all depend upon our ability to satisfactorily answer these questions.

There needs to be a sustained and heavy investment in a nation-wide Vertically Integrated Semiconductor Ecosystem (VISE). VISE is a program in which research and development is conducted seamlessly across the entire compute stack – from applications, systems and algorithms, architectures, circuits and nanodevices, and materials. A nation-wide VISE provides clear strategic advantages in ensuring the US's global superiority in semiconductors. First, a VISE provides the highest quality seed-corn for nurturing transformative ideas that are critically needed today in order for nanotechnology-inspired computing to flourish. It does so by dramatically opening up new areas of semiconductor research that are inspired and driven by new application needs. Second, a VISE creates a very high

barrier to entry from foreign competitors because it is extremely hard to establish, and even harder to duplicate. VISE changes not just the rules of the game but establishes a different game all together – one in which the highest levels of innovation across widely disparate domains need to come together cohesively.

Fundamental research is needed that explores alternative models of computation that acknowledge nanoscale realities by embracing their intrinsically statistical attributes. These include Shannon/brain-inspired models, probabilistic and stochastic models of computation. Fundamental limits on energy efficiency, latency, and accuracy need to be established via a combination of automata and information-theoretic approaches. New design principles and system theory based on such models need to be investigated in order to realize future computing systems. In fact, a rethinking of the design abstraction is required that effectively ties application needs to the nanoscale device technologies, and that enables the design of scalable computing platforms with reasonable design effort. This requires innovative heterogeneous integration tools, programming models, architectures, and heterogeneous 3D system structures. Abstraction layers need to be rearranged or vanish entirely.

Novel algorithms and platform architectures need to be explored that blur the traditional boundaries between storage, sensing, computing, and communication, in order to design computing systems that provide unique energy-delay-accuracy-functionality trade-offs. New platform concepts such as in-sensor computing, in-memory computing, and distributed systems, need to be developed by exploiting opportunities in removing the barriers between diverse modalities. In particular, there needs to be increased focus on memory-centric platforms covering the entire stack. These platforms need to be developed in the context of cloud-based, autonomous, and human-centric applications. Platform-aware learning algorithms and systems that comprehend resource-constraints, i.e., constraints on energy, storage, computation, communication, variability, and form-factor, need to be investigated. including, and high data rate communications.

Device technology research going forward must address the needs of emerging applications and new models of computation, and not focus solely on the logic switch. New device primitives/functions (nano-functions) need to be defined, both from a systems-driven (top-down) and device-driven (bottom-up) approach, to arrive at optimal solutions. Novel substrates such as DNA and memory technologies need to be investigated. Cost-effective monolithic 3D integration of logic and memory in a fine-grain fashion need to be explored. While nanotechnologies have plenty of diverse capabilities to offer, integration and integration methodologies are key.

There is a clear need for a national infrastructure as part of VISE that provides heterogeneous integration capabilities and scalable design methodologies in order to enable systems integration and scalable system demonstrations. Enabling the next revolution in information processing requires that we rise from the current technology stagnation to create scalable integrated systems that can be manufactured in an economic way. Making this happen will require an approach similar to the VLSI revolution at the end of the 1970s when scalable and reliable manufacturing was made available through design rules that limited the design space and options (“freedom from choice”). Standard interfaces are a first step, but are not sufficient. An accompanying design methodology and tool set

including modeling, design, operation and verification for nanoscale 3D systems is needed. A number of prototyping sites for 3D heterogeneous systems should be made available for access to the larger community. These could be housed at the National Labs, interested semiconductor partners, or independent research labs such as Albany Nanotech in the US, or IMEC and LETI in Europe. Creating a scalable heterogeneous 3D prototyping and design capability will undoubtedly require a sizable investment in all of the aspects of the ecosystem, in absence of which it is highly likely that many game-changing ideas and concepts in the next-generation of information processing will stagnate, or that other countries or continents may take the lead.

In summary, computing systems implemented using nanotechnology has the potential to deliver immense societal benefits as it has done in the past. Intelligent, energy efficient and trustworthy machines can dramatically enhance and transform the human experience, in how we interact with and perceive the world around us and ourselves. However, to realize this potential, it is critical to address the challenges facing the longstanding symbiotic relationship between nanoscale semiconductor technology and computing. Nanoscale technology challenges related to heterogeneous integration and scaling, need to be addressed jointly with challenges related to energy and latency costs of information extraction from abundant data in emerging applications. In order to address these challenges, it is recommended that there be a: 1) sustained and heavy investment in a nation-wide Vertically Integrated Semiconductor Ecosystem (VISE), including a shared national infrastructure for heterogeneous integration, 2) focus on fundamental research to explore alternative models of computation that acknowledge nanoscale realities by embracing their intrinsically statistical attributes, 3) investigation of novel algorithms for novel platform architectures such as in-memory, in-sensor, and distributed platforms, and a 4) refocusing of device technology research that goes beyond the logic switch in order to address the needs of emerging applications and new models of computation.

## Nanotechnology-inspired Information Processing Systems Workshop

The 1.5-day Nanotechnology-inspired Information Processing Systems visioning workshop brought together a broad community of leading researchers from the areas of computing, neuroscience, systems, architecture, integrated circuits, and nanoscience, to think broadly and deeply about ideas for designing information processing platforms of the future on beyond CMOS nanoscale process technologies in the context of three *application-driven platform-focused* topical areas – *cloud-based*, *autonomous*, and *human-centric* systems. The rest of this report provides a platform-centric view of the workshop summary described above. The workshop was organized by: Jan Rabaey (UC Berkeley), Naresh Shanbhag (UIUC), Hava Siegelmann (DARPA), Philip Wong (Stanford), Mark Hill (U Wisconsin), Randy Bryant (CMU), and Ann Drobnis (CCC), with support from Khari Douglas (CCC) and Helen Wright (CCC).

**Participants:**

- Sarita Adve, University of Illinois at Urbana-Champaign
- Sankar Basu, NSF
- Randy Bryant, Carnegie Mellon University/CCC
- Doug Burger, Microsoft
- Gert Cauwenberghs, University of California San Diego
- Luis Ceze, University of Washington
- Bill Chappell, DARPA
- Mei Chen, University of Albany
- Tom Conte, Georgia Tech
- Sandra Corbett, CRA
- Khari Douglas, CCC
- Ann Drobnis, CCC
- Yiftach Eisenberg, DARPA
- Ralph Etienne-Cummings, Johns Hopkins University
- Anne Fischer, DARPA
- Wilfried Haensch, IBM
- Bill Harrod, ASCR
- Mark Hill, University of Wisconsin
- Tom Kazior, Raytheon
- Samee Khan, NSF
- Amir Khosrowshahi, Nervana
- Carolyn Lauzon, Department of Energy
- Daniel Lee, University of Pennsylvania
- Jie Liu, Microsoft Research
- Sharad Malik, Princeton University
- Veena Misra, North Carolina State
- Subhasish Mitra, Stanford University
- Klara Nahrstedt, University of Illinois at Urbana-Champaign
- Vijay Narayanan, Pennsylvania State
- Jan Rabaey, University of California Berkeley
- Dan Radack, IDA
- Ed Rietman, University of Massachusetts
- Sayeef Salahuddin, University of California Berkeley
- Linton Salmon, DARPA
- Alan Seabaugh, Notre Dame University
- Naresh Shanbhag, University of Illinois at Urbana-Champaign
- Hava Siegelmann, DARPA
- CY Sung, Lockheed Martin
- Josep Torrellas, University of Illinois at Urbana-Champaign
- Lav Varshney, University of Illinois at Urbana-Champaign
- Naveen Verma, Princeton University
- Lloyd Whitman, OSTP
- Philip Wong, Stanford University
- Helen Wright, CCC
- Katherine Yelick, University of California Berkeley
- Todd Younkin, Semiconductor Research Corporation