A Vision for using Co-design to enable a "Cambrian" Era in Computing

A. Aiken (Stanford), J. Bokor (Berkeley), V. Churavy (MIT), A. Edelman (MIT), D. Ernst (Cray), S. Hearne (Oak Ridge), J. Hennessy (Stanford), H. Mabuchi (Stanford), P. McIntyre (Stanford & SLAC), N.A. Melosh (Stanford), Sh. Misra (Sandia), S. Mitra (Stanford), J. Neaton (Berkeley & LBNL), T. Ogitsu (LLNL), O. Ovchinnikova (Oak Ridge), J. Rabaey (Berkeley), A. Salleo (Stanford), E. Schwegler, (LLNL), M. Shulaker (MIT), R. Stevens (U of Chicago & Argonne), T.P. Straatsma, (Oak Ridge), D. Turek (IBM), N. Wichmann (Cray), R. N. Zare (Stanford), and S. Shankar (Harvard)

Abstract

The following vision document on Co-design has been prepared based on our learnings from research and applications in microelectronics and Moore's law across academia, national laboratories, and industry. The Co-design refers to the methodology in which architecture of the computing platform, system hardware, computing and communication devices, algorithms and software are designed for an application.

The revolution in computing over the last five decades has been driven by an era in which Moore's law and related advances have led to enormous advances across many areas in society. Currently, we are at an inflection point driven by a variety of *positive* factors and *challenges*. The nexus of these two forces provides an enormous opportunity for us to re-think computing in the next few decades to address grand challenges in energy, materials, health, and societies. Our vision is to enable a new era in personalized computing ("Cambrian" era) that bridges information theory, computing and communication abstractions with materials, devices, hardware, systems, architecture, algorithms and software for enabling new applications.

We propose a founding of an "innovation hub" on Co-design that will leverage the strengths of the US in its research and teaching institutions, national laboratories, and entrepreneurial energy. We propose this initiative as a hybrid model in which academics, national labs/other federal agencies, and industry can come together to build an evolving, flexible, and long term self-sustaining effort. The federal government has always played a pivotal role in the enabling leadership of the U.S. semiconductor industry by directly funding and also in being one of the largest customers of the electronics industry. Given that the Department of Energy (DOE) is one of the largest users of high performance computing, we believe that this strategic effort outlined in this document will benefit from being seeded by the government. This will bring attention to all aspects of microelectronics industry including manufacturing. We see our vision as consistent with the DOE Executive Summary on Co-design (2018).

In the appendix, we are also attaching a summary of two meetings (in 2017 and 2019), in which many aspects of Co-design were discussed which cut across a wide swath of areas with participants who are computer architects and designers, biologists, chemists, mathematicians, physicists, engineers, practitioners, and medical doctors. We believe that the new era in computing requires a highly interdisciplinary expertise of scientists, engineers, technologists, and application practitioners as this is beyond the traditional scaling paradigm into which the community has matured.

1. Summary

Since Moore's statement from nearly half-a-century ago (Moore, 1965), advances in computing have crossed three eras: The first focused on Dennard's scaling (Dennard, 1974); the second era included breakthroughs in materials, devices, and lithography pattering; the third era was based on innovations in machine learning algorithms, and in architectures. The revolution in computing over the last five decades has been driven by the scaling law and the related advances, which in turn have led to enormous progress across many areas in society. Currently, we are at an inflection point driven by a variety of *positive* factors: ease of connectivity across the world, availability of large quantities of data, machine learning techniques including deep learning methods cutting across several disciplines, new computing architectures driven by applications, open source software and platforms entering mainstream, use of clouds for economically viable large-scale computing, and nanotechnology for precise engineering of systems. On the other side, we face *challenges* including slowing down of Moore's law, the increasing complexity and the cost of technology manufacturing leading to consolidations in the industry, and losing of the national leadership in the areas of computing in many of the areas pioneered in the United States. The slowdown of scaling in itself indicates the maturity of scaling (Moore, 2003). The nexus of these positive forces and challenges provides an enormous opportunity for us to re-think computing in the next few decades to address challenges facing national, human, and social aspects as applications. We are proposing that the fourth era in which Co-design of multiple building blocks can help address almost any application. Such a framework for Co-design could enable systematic optimization across algorithms, software, architecture, materials, hardware and system components. This in turn could enable computing solutions for any application, less constrained by the local design parameters. The following figure illustrates this evolution.

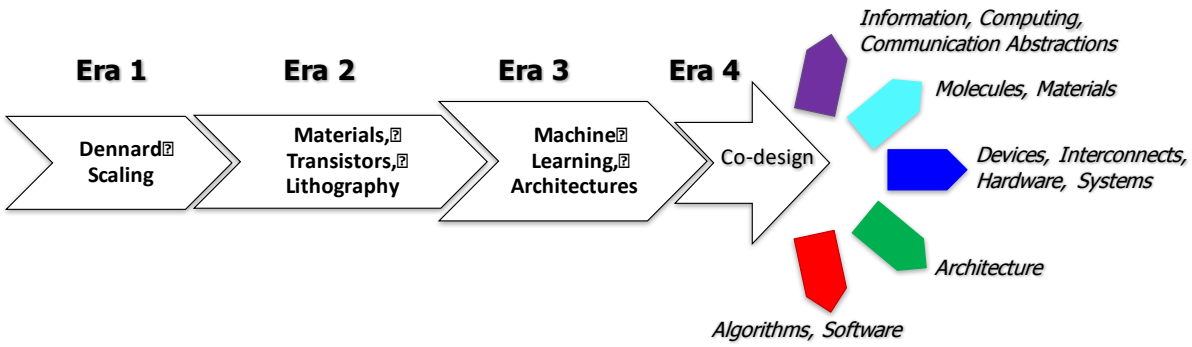

Figure 1: A high-level view of Co-design, as the next era in computing

There are several early examples of Co-design including the Anton family of machines for protein folding, Exa-scale computers for high performance scientific computing applications undertaken by the Department of Energy, Deep Blue and Watson designed by IBM, tensor processing units for machine learning developed by Google, etc. By providing the building blocks for Co-design, it would expand the ecosystem for the community to address a wider repertoire of applications than are currently addressed. Exploring several paths from conceptualization to physical prototyping of computing systems, by systematically using different building blocks should lead to a Cambrian explosion in computing. In order to

facilitate this, we are proposing an "innovation hub" that will enable the community to use the building blocks for developing physical computing prototypes for discovery and rapid exploration. The innovation hub in collaboration with the Department of Energy's national labs will provide unique expertise combining academia, career scientists. The hub will explore beyond industrial R&D outcomes using academic research and the capabilities in the national laboratories. The engagement of both established and newer industrial partners with the hub will provide connections with the semiconductor industry. The role of academic institutions in fundamental research and in education of the future workforce will be complemented by the long-term research in the national laboratories. This, in turn, will leverage the strengths of the US in academic and national labs research and hence provide a long-term viable pathway for innovations in the US similar to what the different electronics industry consortia accomplished in since 1980s.

2. Lessons Learned from Previous Efforts

As the authors of this paper have been in the industry, national labs, and academia, our perspectives cover all aspects of the relevant areas of research, development, and manufacturing in the electronics, semiconductor, and computing industries. In this section, we will discuss some of the lessons learned that will be useful to any similar collaborative effort on microelectronics or computing.

a. Consortia

Although there are several consortia and centers established in the US universities and national labs, we will focus on two specific entities related to the microelectronics industry. Over the years, the electronics industry has established and sustained several consortia including SEMATECH and Semiconductor Research Corporation (SRC). Each of these had different goals and were set-up to address specific challenges that the US semiconductor industry was facing in the 1980s.

SEMATECH (*SEmiconductor MAnufacturing TECHnology*) was founded to solve common manufacturing problems in a pre-competitive manner and regain competitiveness for the US that was challenged by Japanese industry in the mid-1980s. The consortium was a partnership between the United States government and 14 U.S.-based semiconductor manufacturers established in 1987 in Austin, Texas as an independent entity. SEMATECH was funded over five years by public subsidies coming from the U.S. Department of Defense/Defense Advanced Research Projects Agency (DARPA) for a total of $500 million and by the semiconductor companies for $500 million. The first CEO was Robert Noyce (co-founder of Intel with Gordon Moore). The focus of this effort was to focus on precompetitive research and development, set up a pilot plant for prototyping and manufacturing, use of modeling, simulation, and computer-integrated manufacturing and act as a technology catalyst. It is estimated that that the cost of research and development for every generation of new technology was reduced from 30% to 12.5% by this effort by mid-1990s (MIT, 2011). SEMATECH was credited with regaining leadership by acceleration of problem identification and down selecting possible solutions. In addition, the entity was able to be financially self-sustaining in a decade since its founding. According to the Semiconductor Industry Association, the U.S. industry rebounded over the next 10 years, and by 1997, it had regained its leadership position close to 50% global market share.

SRC (*Semiconductor Research Corporation*) was established in 1981 when the Semiconductor Industry Association to address university research to deliver early research results which also enables relevantly educated technical talent. The key purpose was to accelerate basic research in semiconductor research disciplines in collaboration with the universities. In 1982, during the first half of SRC's first full year of operation, research contracts were awarded to over twenty-five universities. SRC was credited with establishing and implementing an ambitious research agenda have been key to enabling the exponential growth that Gordon Moore envisioned and articulated in "Moore's Law" (SRC). The university research enabled by the consortia has provided both the breakthroughs and the workforce that has sustained the technology growth since the 1980s. In addition, the research in the national laboratories has also catalyzed the translation of ideas to manufacturing.

As mentioned above, both these entities contributed to sustaining the US leadership in the electronics design and manufacturing at least for one to two decades. However, in 2019, the semiconductor industry is in a transition as they think beyond traditional scaling (Arden et.al, 2010). Many U.S. firms are building semiconductor fabrication plants (fabs) outside the country, primarily in Asia. In addition, most of the semiconductor companies are becoming "fab less," by internally focusing on chip design and relying on contract fabs abroad to manufacture their products. At yearend 2015, there were 94 advanced fabs in operation worldwide, of which 17 were in the United States, 71 in Asia (including 9 in China), and 6 in Europe (Platzer, 2016). This semiconductor manufacturing highlights national security concerns, as electronics components are critical to defense and economy. Clearly the technological leadership won by the US industry is under challenge. Some of the key lessons to be learned include need for continuing involvement from the federal government for long term technology leadership, which in turn is determined by fundamental research in the universities and national laboratories. In addition, the consortia were dependent on contributions from its partners and state funding for its financial viability. Any new entity should address the strengths demonstrated by the above consortia while addressing the limitations: 1) Maintaining a manufacturing line is expensive unless the long term financial viability is taken into account; 2) Dependence on short term funding can be limiting.

b. Moore's Law and Scaling

Moore's law, related to the reduction of the critical dimension of the switching device, has been driven by innovations in geometrical scaling (Dennard's scaling), process and device engineering, lithography, materials, and manufacturing at scale. The digital and computing revolution enabled by the progress of Moore's law has catalyzed several other innovations in the modern society. In addition to several innovations in architecture, design, algorithms, software, there have been breakthroughs in chemistry, materials, processing, and manufacturing (Arden et.al, 2010).

As the dimensions of integrated circuits scale into tens of nanometers and smaller, the related complexity of sustaining Moore's law scaling is increasing leading to noticeable

delays in bringing newer generations of technology into production[1]. With the slowing of Moore's law, a few related questions arise: what it will take to architect and design a computer that operates at the limits of thermodynamics and nature, yet computationally efficient for practical applications? Even as the semiconductor technology scales to 5 nm or 3.5 nm or beyond, is computing as it exists now economically viable to sustain or relevant for new applications? We think that there is a need to re-evaluate the model of information processing from a fundamental perspective, as elaborated further below. In summary, it is clear that the old thinking of making incremental scientific improvements in isolation will not be enough to meet the challenges exploring computing technology in accordance with the traditional scaling framework of Moore's law. There is a need for both researchers in academia and national laboratories to be able to explore the options as well as industrial partners.

c. Other Co-design Efforts

There are several Co-design themed efforts including Anton, Department of Energy's Exa-scale Project, Application-centric activities on bioinformatics initiatives between DOE and NCI, and other efforts (Dally, 2018). In all these projects, several aspects of Co-design have demonstrated successful design linked to applications. The Anton project has shown the possibilities of how computing related performance gains from Co-design can lead to application-centric advances in context of studying protein folding at longer time scales than ever before (Shaw et.al, 2008). Similarly, the DOE efforts on new computer technologies have shown how Co-design thinking can better support high performance computational needs. As part of these evaluations, extensive discussions take place with hardware vendors, who share their technology roadmaps under non-disclosure agreements. These interactions inform both the vendors of the computational requirements and the facilities staff of the capabilities and opportunities provided by technologies that may be years in the future. This effort has proven to be very effective in both the design of new supercomputer hardware and providing the lead time for software developers in preparing for the labor-intensive process of refactoring and optimizing their codes. Over time, the tracking of code performance efficacy has shown that "just waiting for faster hardware" is no longer a usable strategy and that dedicating time to understanding platform performance is becoming required to advance computational abilities. Some of the current studies in this effort have shown that in many cases there is far more code performance upside available from software optimization than new hardware.

The process of architecting a system using Co-design techniques involves a significant amount of effort to be spent in bridging the knowledge, language, and abstraction barriers between the participating parties. One example of this level of effort is the time spent by vendors collaborating with the various participants in Department of Energy-funded programs in Exa-scale computing in understanding and optimizing for target applications with the premise of Co-design. One of the mechanisms for expressing requirements for these programs was through mini-apps – skeleton codes that were built to be representative of full-scale applications while retaining enough simplicity to be tractable. There is ample

[1] https://www.technologyreview.com/s/601102/intel-puts-the-brakes-on-moores-law/

opportunity for studying and improving the practice of cross-layer collaboration and developing tools to support those processes.

3. Goals:

The framework envisaged in this report should address all aspects of computing including both the research and translations to scale-up and manufacturing. We suggest a specific set of goals for the Co-design effort:

1. Develop a systematic framework for Co-design based on the building blocks for efficient computing of both existing and new applications;
2. Develop and build a rapid-turnaround prototyping platform for physical realization of new and novel computing systems consisting of translation from concepts to prototypes;
3. Develop a theoretical basis for computational and communication models and abstractions, for understanding how newly designed computing systems can process information more efficiently for faster and larger-scale computations for widely different applications;
4. Design and develop modular and reusable software framework, methodologies, and tools for integration, testing and validation of different underlying hardware systems;
5. Develop and facilitate an "innovation hub" that will enable the community to use the building blocks for developing physical computing prototypes for visualizing and testing new designs. This hub/infrastructure will allow systematic exploration of linkages between various architectures and newer information processing devices by designing and testing physical prototypes;
6. Develop and implement a framework which enables early adopters to have access to the necessary emulation and physical resources to exploit the full potential of the machine as a prototype and in early development. In addition, use an interactive design and testing environment to enable feedback of the lessons from early adopters to the developers of both the system software and hardware layers of Co-design stack.

To achieve these goals, we define the computing paradigm in terms of the following parameters: (1) Energy efficiency across the scales from materials to systems that will help design systems that maximize information processing per unit of energy; (2) Increases in information processing throughput through novel architectures; (3) Hardware and Software security for ensuring system integrity and trust (Hill, 2018); and (4) Personalization of computing in which systems are designed for optimal computing in each application. With these computing parameters, this framework should encompass **two dimensions** as illustrated below.

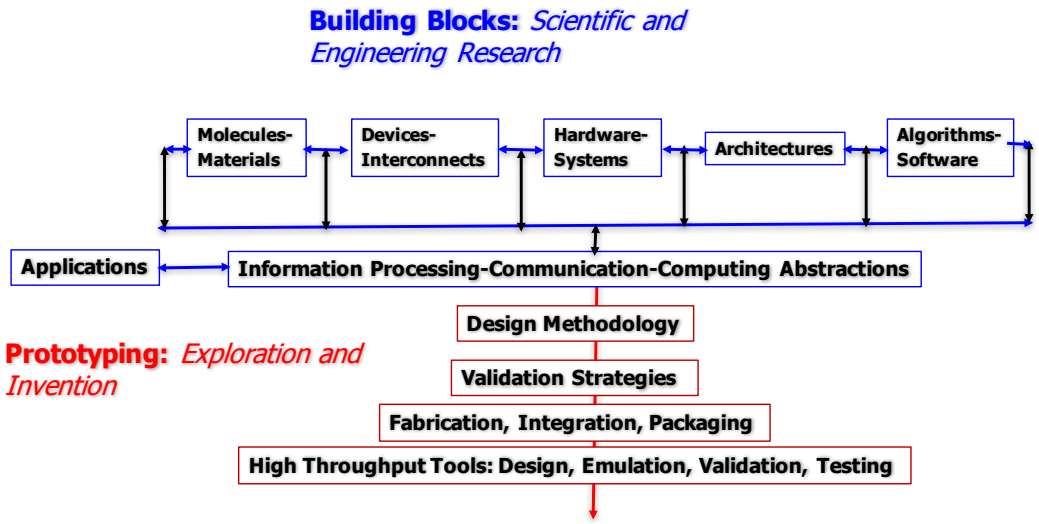

Figure 2: Two-dimensional Co-design framework for conceptualization and physical realization

The current methodology used in the electronics industry is generally sequential as follows: The design starts from system specifications, followed by architecture design through logic design, circuit design, circuit and physical layout design which is then transferred for fabrication and manufacturing. From the technology side, materials, devices and interconnects determine the design building blocks. Although there are collaborations between the two basic arms, the process is sequential in transfer of information. The new proposed framework should enable a modular, "legos-type" approach where researchers could put together the building blocks to design and prototype the various computing options. In addition to the top-down or sequential approaches, this could also help explore bottom-up (from information processing and computing abstractions) using atoms and materials to system level and to physical prototypes. We briefly touch upon the two-dimensional approaches below.

a. Scientific and Engineering Research

The vision of Co-design here should be to address multiple aspects including practical realization of new and novel architecture concepts for designing new computing systems. The existing methods through the eras 1-3 have been discussed in literature (ITRS, 2009; Hennessey-Patterson, 2019). These methods develop architecture and optimize design for an existing process technology. In era 4, we are proposing extending this to include architectures, hardware, software, and devices, basic computing/communication abstractions, and atoms (in the choice of molecules and materials). This in turn expands the scope of computing and bridges abstractions with a system. This framework should be able to use these building blocks to design a computing system for existing or new applications. These components and their interfaces, which are driven by the applications, are summarized below.

1. *Applications:*
   Applications determine the domain for which computing itself is formulated and can span the spectrum from High Performance Computing to low power including

applications that have not been realized yet. Given the criticality of the applications, they are addressed in a different section (below). There should be close collaborations with application writers to ensure their needs are understood by the rest of the stack. Working across the stack in applications, systems software and hardware creates opportunity for innovative Co-design, but also requires a disciplined process to manage the very different kinds of people and knowledge across the breadth of the effort.

2. *Information processing and communication models and computing abstractions*:
   High-level abstractions are helpful to target diverse hardware and extract performances efficiently. At present, we lack any broad theory or methodology that addresses the need for such rethinking of an entire design frame (set of abstractions in a stack). A fixed set of inter-layer abstractions in a protocol stack defines an engineering paradigm that promotes efficient incremental refinement of systems and devices, but revolutionary Co-design may require the invention of radical new abstractions. These should address several questions: 1) Is there a robust, portable, and efficiently-enforceable abstraction for a unit of secure information? 2) Can we formulate new protocol stacks that enable network applications to predict the energy cost of distributed operations? 3) When we discover new materials with unprecedented properties of a functional but unconventional nature, can we systematically analyze a range of signal formats, input-output behaviors, and associated composition (circuit) rules that devices based on them would support?

3. *Underlying molecules, materials, and chemistry, including synthesis and processing methods*:
   As this approach should also include a bottom-up approach, materials and molecules would be significant building blocks for computing. The rise of multifunctional atomically engineered materials opens up unprecedented opportunities for bottom-up engineering of building blocks for computing, starting indeed from single atoms. Due to its engineered functionality, the material or the molecule thus becomes the device, which is to be reproduced and integrated in a circuit and, later, a system. As such, Co-design must reach all the way down to the atomic level to integrate materials engineering and advanced characterization early in the process. Similar to newer architecture, the bottom-up efforts would help use properties of new classes of materials in specific applications beyond the traditional computing areas (e.g. Perovskites, 2D materials, soft materials, nanomaterials).

4. *Physics and engineering of devices, interconnects*:
   The semiconductor industry continues to face difficult challenges related to extending integrated circuit technology beyond the end of conventional CMOS dimensional scaling: 1) extending CMOS beyond its ultimately scaled density and functionality by integrating, for example, a new high speed, dense, and low power memory technology on the CMOS platform; 2) developing and scaling new information processing devices substantially beyond that are attainable by CMOS alone using an innovative combination of new materials, devices, and architectural approaches for extending CMOS. Similarly, there are needs for new memory

technologies that combine the best features of current memories in a fabrication technology compatible with existing technology platform for both stand-alone and embedded memory applications. Interconnects are critical to the communication at all levels of computing, both within a microprocessor and also with external components. In addition, they are critical to power delivery and thermal management systems. Design, fabrication and metrology methods for ultra-dense, monolithic integration of logic and memory devices, guided by a molecular-level understanding of the materials and chemistries, are sorely needed. The challenges for interconnects are related to materials and their interfaces, processing, measurements, integration and control, and reliability (ITRS-Interconnects, 2015). The current paradigm of copper conductors with liners and low-k dielectrics must be replaced to achieve denser interconnection and lower power operation. New approaches and new physics, using topological and low-dimensional materials for example, are needed to shift interconnect performance onto a growth curve.

5. *System-level analysis including Circuits for hardware integration*:
   This system-level work can progress in several phases. At the onset of the program, the system-level work can focus on exploration – both experimentally and theoretically. This will enable exploration of new nanomaterials and nanodevices, which themselves could enable new architectures. This should include system-level modeling and analysis with calibrated compact models of heterogeneous technologies, as well as experimental small-level demonstrations to learn the challenges – and potential benefits - associated with integrating each technology. After this early exploration identifies the most promising future systems, subsequent work can focus on both detailed analysis to understand the key benefits afforded by these new systems, while in parallel fabricate large-scale hardware prototypes. These hardware prototypes will both demonstrate feasibility while simultaneously providing experimental calibration for the on-going analysis work within the center.

6. *Architectures* (both von Neumann and other architectures):
   Most of the studied architectures that have been considered to date in the context of new devices utilize binary logic to implement von Neumann computing structures (Hennessy-Patterson, 1989; ITRS, 2009). As research in materials helps invent new devices and identify new properties, architectures that leverage these features will begin to appear (Kloss, 2016). Slowing down of Moore's law and the limitations of scaling, efficient improvements in each generation are limited to a few percent (Hennessy-Patterson, 2019). As indicated in this work and others, higher rates of improvement can be achieved mainly by new domain specific architectures, where the processing units which are programmable and are tailored to application domains (N. Jouppi et.al, 2018).

7. *Algorithms and Software*:
   There should be exploration of programming systems/operating systems that use the basic services to provide a holistic view of a collection of hardware resources, allowing applications to take the approach of programming the machine rather than

its components. A key piece will be providing programming abstractions that provide sufficient insulation from the details of the hardware that programs are portable to other systems or to the same system where components have been upgraded or replaced. A common low-level software interface, analogous to a compiler intermediate form that provides a minimum of abstraction above the level of the hardware, will decouple hardware and software efforts and allow parallel work on the layers above and below. There is an active research community working in this are that should be expanded within a bigger Co-design effort.

The components in these layers or building blocks themselves need both fundamental research and translational aspects to enable successful interfacing between the different layers for effective Co-design.

b. Prototyping

Historically, understanding the opportunities and limitations presented by new ideas in materials, devices and circuits has been limited by the practical limitation of testing those ideas at the system level. In this effort, there should be focus on delivering physical prototypes for exploring various computing options. The prototyping can be facilitated by advances in nanotechnology and fabrication, high throughput methods, and machine learning for translating ideas and concepts to physics systems. This will consist of developing the following components:

1. Design methodologies
2. Validation strategies
3. Tool sets for design, emulation, and validation
4. Fabrication, Integration, and Packaging

The explosion of tooling cost for characterization and fabrication in microelectronics now threatens the entire ecosystem because it presents an impediment for researchers to evaluate new ideas, especially in the US. Hence multiscale design, emulation, and validation tools incorporating machine learning must be developed to discover and computationally prototype new materials integrated into existing or envisioned future devices. In parallel, new platforms are needed for multi-modal and non-destructive high-throughput characterization of materials in various stages of patterning into device structures. Connecting to the new tools for developed for design and validation of materials and integration approaches, novel high-throughput synthesis methods for materials and their interfaces (e.g. semiconductor/insulator, semiconductor/metal, organic/inorganic etc.) must be reduced to practice in order to test simulation predictions and build richer materials informatics databases. Area-selective growth and controlled placement of nano- and micro-scale devices are needed to rapidly create physical prototypes of new circuits and systems as a realization of the premise of nanotechnology (Feynman, 1959).

Fast combinatorial exploration using advanced machine learning methods can help understand absolute limits imposed by physics and materials. At the physical limit, non-scalable manufacturing and characterization techniques now reach to the atomic scale through taking advantage of pathways such as surface chemistry and beam-based

manipulation in the national laboratories. These create the opportunity to characterize and manipulate, from the ground up, the parameters that underlie the principles of operation of current-day devices (e.g. electrostatics, strain, and thermal conductivity). Thus prototyping in Co-design can make it possible to intelligently simulate and test the system-level impact of a change at any level, thereby relaxing the requirement of achieving scalable manufacturing to evaluate innovations. This new paradigm will rely on developing different non-scalable 'manufacturing' tools - ones that accelerate discovery of new innovations, and ones that accelerate evaluation of ideas against basic science limits. At the lower levels of the framework, this requires physical experimentation, as a pure modeling and simulation approach is often computationally untenable for new materials and devices. Existing non-scalable fabrication projects on atomic precision advanced manufacturing and directed matter show the power of such an integrated multi-disciplinary approach at these lower levels, but also the limitations of pursuing innovations without an even larger team that is capable of assessing impact by developing a multi-scale Co-design framework. Hence, we think that prototyping is needed to demonstrate the scalability of the computing systems to different applications in technology areas. The design methodologies will demonstrate top-down (from application over system and architecture to information processing platforms) and bottom-up (from information processing to computing systems) pathways.

4. Infrastructure:

In order to facilitate the two-dimensional pathways, the innovation hub could be a multi-disciplinary, multi-university entity (e.g. Institute of Computing Technology), that is established in the US. This could collaborate with the Department of Energy (DOE) national laboratories and industrial partners to take advantage of the investments in advanced fabrication and characterization capabilities. The combined strength of the universities and the national laboratories within the innovation hub would enable translation from conceptualization to systems based on the lessons learned from the successful aspects of industrial consortia (e.g. SEMATECH, SRC), university-led research centers (e.g. Stanford BioX), and DOE hubs. The institute-enabled consortium could help transition of research by demonstrating prototypes, which can be transferred to industrial partners for scale-up and manufacturing. In addition, there should be efforts to organize annual meetings and semester-long workshops (similar to the Kavli Institute series) for engagement between academic researchers, national laboratory scientists, and industrial personnel in use of the different building blocks.

5. Applications:

As mentioned, the applications can be diverse, spanning the spectrum from High Performance Computing to low power including applications that have not been realized yet. As requested in the DOE RFI, we would like to list several applications. Many of them were discussed in the several meetings that we participated (please see appendices for a summary of the two meetings relevant to this discussion). A few examples of applications are listed below.

a. High Performance Computing including Edge Computing addressing data collections for materials characterization, bioinformatics (e.g Genomic-, Proteomic-based computing), and medicinal applications (e.g. image processing, diagnosis)

b. Low power computing and communication devices for sensing, mobile applications, and IoTs
c. Reconfigurable and heterogeneous computing (e.g., interfacing between different architectures, devices, etc.)
d. Probabilistic Computing for applications in which noise is integral to operation (e.g., biological systems process information in the presence of noise)
e. Advanced Machine Learning Algorithms and Software for Design, Automation, and Control

We illustrate a dozen potential applications in the following figure, that would be accelerated by the premise of Co-design: cognitive computing, edge computing for large-scale measurements, wearable sensing, distributed sensing, high performance computing for scientific applications, probabilistic computing, machine learning, artificial vision, medical diagnostics, bio informatics, high throughput chemical and physical measurements.

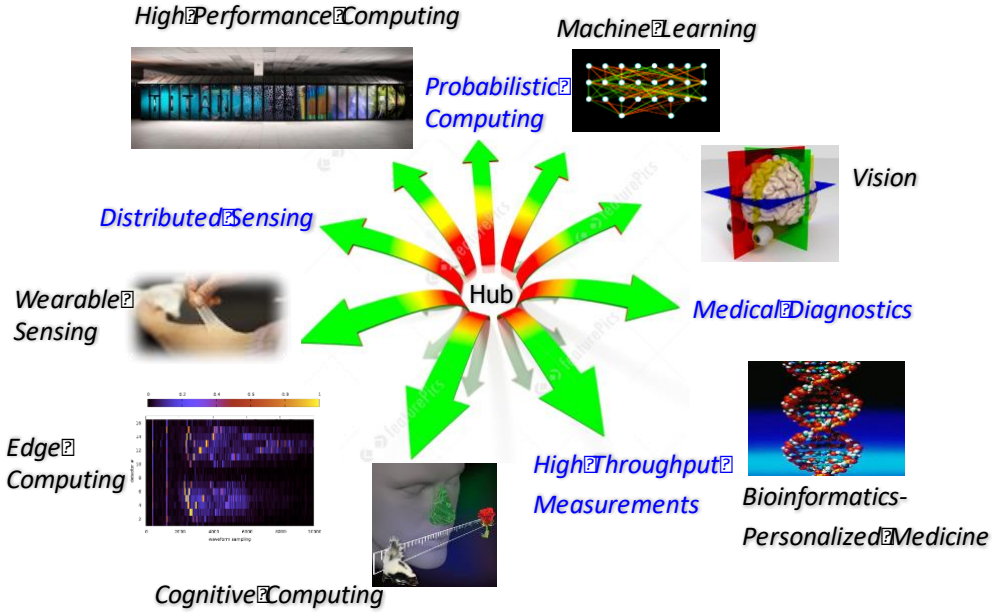

Figure 3: Examples of computing applications that could be enabled by an Innovation hub in Co-design for a "Cambrian" Era

6. Motivations and Notable Advantages:

As mentioned before, we are addressing multiple aspects of computing from basic research to prototyping. A comprehensive effort of Co-design would help in developing concepts, methods, tools, enabling rapid exploration by physical prototypes. These can be scaled to manufacturing either in collaboration with industrial partners or start-up spin-offs. These prototypes, would accelerate exploration of new computing systems and also reduces the risks during scaling up. Although not all concepts become products, the lessons learned in prototyping is useful in understanding the intricate relationships between the building blocks. The intent of this effort is to reduce the risks in translating concepts to products and to provide an ecosystem for design and realization of distinct computing systems. This is illustrated in the following figure.

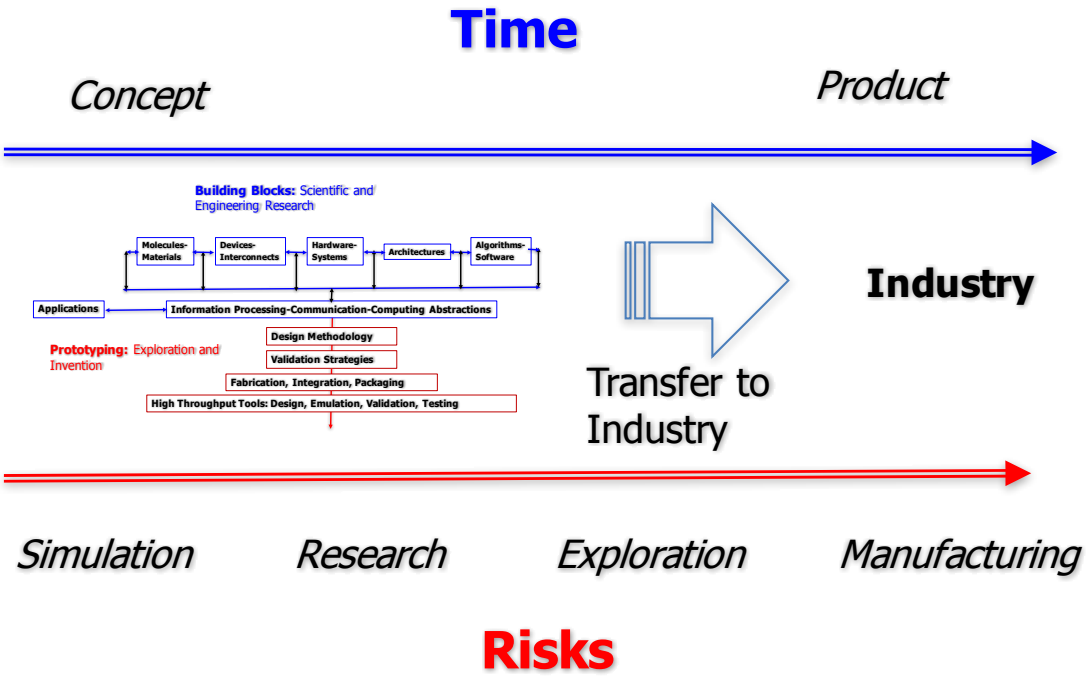

Figure 4: Use of different building blocks to develop prototypes for subsequent transfer to industry for scale-up and manufacturing (cross-check with Figure 2).

In the status quo, Co-design is often done at the local level, where hardware and software are co-optimized for a given application. Also, this information is rarely publicly available, which renders it difficult for researchers to understand how decisions that are made at a particular level (e.g., devices) can have ramifications at the systems level. In addition, as discussed before, the microelectronics manufacturing is decreasing in the US. The semiconductor companies have become slower in their ability to adhere to Moore's law (i.e., doubling the power of microchips every ~2 yrs), while abiding by the constraints of the manufacturing process and maintaining their profits. Even as the semiconductor technology scales beyond 5 nm, it is not clear that computing, as it exists now is efficient, economically viable, or relevant to all the applications.

Therefore, we think that there is a need to reevaluate the model of information processing and computing abstractions from a fundamental perspective and connect it to realization of practical computing systems. We aim to go beyond the incremental scientific improvements in isolation and bring a systematic and open approach to design across multiple levels in computing. This effort will enable exploring multiple layers across the stack and creating an ecosystem for which both successful and unsuccessful links can be shared and understood through fundamental research.

This in turn should enable this effort to be compelling as not a single computing solution is optimal for every application as summarized below.

a. Recasting computing in terms of open building blocks will help bridge the gaps between computing abstractions and physical porotypes. This should enable communities of scientists and engineers from academia, national labs, and industries to design and test prototypes for their applications. Teams interested in expanding the designs of existing systems or in conceptualizing new computing systems can work with the hub to scale the prototypes to manufacturing.
b. Methodologies, platforms, and toolsets for further scaling and manufacturing should enable a multiplicative effort, of a "Cambrian" era in computing.
c. A building block-based approach will complement More than Moore's efforts law (Arden et.al, 2010) as it will advance beyond the traditional scaling into smaller and smaller technologies as well as accelerating build-out in the 3rd dimension in both monolithic as well as advanced packaging technologies.
d. The efforts would need a team from multiple disciplines (e.g., chemists, physicists, informaticians, mathematicians, engineers, and computer scientists) working together to enable research breakthroughs and subsequent inventions and the resulting innovations.
e. The effort should leverage the strengths of the institutions in the US, the academic research/teaching universities and the national laboratories. Our hope is that this in turn will bring back to the US innovation in computing beyond the software.
f. As these efforts are led by institutions of higher learning, courses in multiple areas could be offered to train the future workforce.
g. The explosion of data and knowledge, along with the advent of new computer architectures enabling commercial deployment of machine and deep learning, has created the perfect storm motivating a reconsideration of how R&D is conducted. The community of players involved in R&D activity can now take advantage of knowledge at scale that was simply impossible before. And, as a consequence, insights and discoveries will be deeper and more efficient than anyone could have previously imagined in presence of this Co-design effort.

References

1. W. Arden, M. Brilloue¨t, P. Cogez, M. Graef, B. Huizing, and R. Mahnkopf, More-than-Moore, White Paper, 2010
2. Dally, W. et al. Hardware-enabled artificial intelligence. In *Proceedings of the Symposia on VLSI Technology and Circuits* (Honolulu, HI, June 18–22). IEEE Press, 2018, 3–6.
3. Dennard, R. et al. *Design of ion-implanted MOSFETs with very small physical dimensions*. *IEEE Journal of Solid State Circuits 9*, 5 (Oct. 1974), 256–268.
4. Department of Energy (DOE) Basic Research Needs for Microelectronics, October 2018.
5. Feynman, R. P. 1959, *There's plenty of room at the bottom*, Eng. Sci. 23, 22–36
6. Hennessy, J. and Patterson, D. *A New Golden Age for Computer Architecture*. Communications of the ACM, Vol. 62, No. 2, February 2019.
7. Hennessy, J. and Patterson, D. *Computer Architecture: A Quantitative Approach.* Morgan Kauffman, San Francisco, CA, 1989.
8. Hill, M., "*A primer on the meltdown and Spectre hardware security design flaws and their important implications*", *Computer Architecture Today* blog (Feb. 15, 2018); https://www.sigarch.org/a-primer-on-the-meltdown-spectre-hardware-security-design-flaws-and-their-important-implications/
9. ITRS: International Technology Roadmap for Semiconductors, 2009
10. ITRS: International Technology Roadmap for Semiconductors Interconnects, 2015
11. ITRS: International Roadmap for Devices and Systems, 2017 Edition
12. Integrated Circuit Design: https://en.wikipedia.org/wiki/Integrated_circuit_design
13. N. Jouppi, C. Young, N. Patil, and D. Patterson, *A domain-specific architecture for deep neural networks*. Commun. ACM 61, 9 (Sept. 2018), 50–58.
14. Kloss, C. *Nervana Engine Delivers Deep Learning at Ludicrous Speed.* Intel blog, May 18, 2016; https://ai.intel.com/nervana-engine-delivers-deep-learning-at-ludicrous-speed/
15. Moore, G. Cramming more components onto integrated circuits. *Electronics 38*, 8 (Apr. 19, 1965), 56–59.
16. Moore, G. No exponential is forever: But 'forever' can be delayed! [semiconductor industry]. In Proceedings of the IEEE International Solid-State Circuits Conference Digest of Technical Papers (San Francisco, CA, Feb. 13). IEEE, 2003, 20–23.
17. M. D. Platzer, J. F. Sargent Jr., "*U.S. Semiconductor Manufacturing: Industry Trends, Global Competition, Federal Policy",* 2016
18. Shaw DE, Deneroff MM, Dror RO, Kuskin JS, Larson RH, Salmon JK, Young C, Batson B, Bowers KJ, Chao JC, et al. (2008); "Anton, a special-purpose machine for molecular dynamics simulation". Commun ACM 51:91–97
19. SRC: https://www.src.org/about/: J. Hennessy Quotes

Appendix A: Meeting on Co-design, Harvard University, April 13-14, 2017

Organizers: David Brookes (Harvard), Alan Edelman (MIT), Roscoe Giles (Boston University), Efthimios Kaxiras (Harvard), Paul Messina (Argonne National Lab), and Sadasivan Shankar (Harvard)

A two-day workshop that brought together interdisciplinary expertise, to assess the state of the applications of computing for bio, chemistry, materials, medical, and technology levels, and articulate a research vision for co-designing hardware, software, and algorithms for different applications in 2017. The meeting highlighted a need for describing a research agenda to inform new public and privately-funded research on a new way of thinking about co-design. The goal is for this to initiate new research programs to evaluate the use of these technologies for designing computing architecture, algorithms, and software tailored for applications. The workshop was held April 13-14, 2017 at Harvard University, Cambridge, Massachusetts. The workshop included participation from the office of the Department Energy, National Federal Laboratories ten universities, and six companies. About 45 presentations spanning a breadth of topics that touched on all aspects of computing. The specific topics included the following:

1. Algorithms, Software, and Programming
2. Architecture
3. Bio and medical applications
4. Exa-scale Project
5. Hardware
6. HPC Applications
7. Materials/Chemistry Applications
8. Big Data
9. Cross-cut

Appendix B: Meeting on Co-design, Stanford University/SLAC, March 5-6, 2019

Organizers: Hideo Mabuchi (Stanford), Paul McIntyre (Stanford/SLAC), Subhasish Mitra (Stanford), and Sadasivan Shankar (Harvard)

The intent of the meeting was to assess the state of the scientific applications of computing and articulate a research vision for Co-design from the level of materials and fundamental physical phenomena through hardware, software, and algorithms for different applications. This was in line with the Department of Energy effort on Microelectronics, but also was looking beyond electronics into computing itself. About 45 people attended from several universities, National Labs, and industry. The meeting had several theme-based talks and six interactive panels covering areas from materials and devices to architectures and applications, including the physics of computing as listed below.

1. Algorithms, Software, and Programming
2. Bio-medical Computing
3. Co-design: Lessons from nascent quantum engineering
4. Cognitive Computing
5. Computing in Harsh Environments
6. Devices and Systems
7. HPC for Chemistry, Materials
8. Materials for wearable and flexible electronics
9. Materials and Devices
10. Memory and Computing
11. Neuromorphic Computing
12. New Computing Applications, Materials
13. Precision Manufacturing
14. Sensing
15. Physics of Computing