DAY ✓ ONE PROJECT

An Institute for Scalable Heterogeneous Computing

Eric Breckenfeld

July 2021

FAS

Summary

The future of computing innovation is becoming more uncertain as the 2020s have brought about a pivot point in the global semiconductor industry. We owe this uncertainty to several factors, including the looming end of Moore's Law, disruptions in semiconductor supply chains, international competition in innovation investment, a growing demand for more specialized computer chips, and the continued development of alternate computing paradigms, such as quantum computing.

In order to address the next generation of computing needs, architectures are beginning to emphasize the integration of multiple, specialized computing components. Within this framework, the U.S. is well poised to emerge as a leader in the future of next-generation computing, and more broadly advanced semiconductor manufacturing. However, there remains a missing link in the United States' computing innovation strategy: a coordinating organization which will down-select and integrate the wide variety of promising, next-generation computing materials, architectures, and approaches so that they can form the building blocks of advanced, high-performance, *heterogeneous* systems.

Armed with these facts, and using the existing authorization language in the 2021 National Defense Authorization Act (NDAA), the Biden Administration and Congress have a unique opportunity to establish a Manufacturing USA Institute under the National Institute of Standards and Technology (NIST) with the goal of pursuing **advanced packaging for scalable heterogeneous computing**. This Institute will leverage the enormous body of previous work in post-Moore computing funded by the federal government (Semiconductor Technology Advanced Research Network (STARnet), Nanoelectronics Computing Research (nCORE), Joint University Microelectronics Program (JUMP), Energy-Efficient Computing: From Devices to Architectures (E2CDA), Electronics Resurgence Initiative (ERI)) and will bridge a key gap in bringing these R&D efforts from the laboratory to real world applications. By doing this, the U.S. will be well positioned to continue its dominance in semiconductor design and potentially regain advanced semiconductor manufacturing activity over the coming decades.

Challenge and Opportunity

Nearly every economic sector is increasing connectivity, adding sensing and data collecting capabilities, and integrating intelligence and automation to the maximum extent. The potential breakthroughs are immense and include:

- Health – Personalized, targeted, cost-effective and remote healthcare and therapies

- Education/workforce – Effective remote digital learning and work environments through augmented and virtual realities

- Transportation – Autonomous and energy-efficient vehicles and corresponding infrastructure

- Manufacturing – Safe, energy-efficient factories with robotics and automation

- Utilities – Highly efficient and affordable smart grids and sensors to track harmful pollutants and diseases

- Communications – All people seamlessly connected with each other and the infrastructure using 5G+ and sustainable data centers

- Security – Born-secure cyber systems, autonomous intelligence and weapon systems, digital twins, and technologies to augment warfighter effectiveness

The net result of this trend is a deluge of data. As it stands today, these sectors are predominantly leveraging and expanding upon cloud computing capabilities. This approach is proving to be unsustainable as the rate of data is increasing far beyond our ability to move that data from user or edge devices to processing centers. In fact, the global rate of data acquisition from sensors has surpassed the cumulative ability of humans to consume that data, meaning we have more data than we know what to do with. In order to address this data crisis, more data must be processed locally at the user device or edge computing levels. However, in order for processing to occur in small, practical consumer devices, they must become vastly more energy efficient. The primary approach for improving the efficiency and performance of computation over the past half-century has been down-scaling of transistor sizes, often associated with Moore's Law. As a result of the limitations of physics, this approach will not continue to yield the same improvements over the next decade as it has over previous decades.

The semiconductor industry and the U.S. government have prepared for this eventuality through decades of R&D efforts. While these approaches have yielded promising results, it is clear that there will not be a silver bullet solution. There will be no *single* material or architecture which can be adopted *en masse* to replace our previous devices and improve performance for all computational problems. Fully addressing these demands requires the use of many different chips, each specialized for certain tasks, and perhaps operating based on completely different computing paradigms, integrated together. This is what is called *heterogeneous computing*. Discrete circuit blocks which are designed to be integrated together into a larger system are called *chiplets*. For example, beyond the well-known consumer examples of the flexible Central Processing Unit (CPU) and massively parallel Graphics Processing Unit (GPU), there are more highly specialized chips focusing on niche and often computationally-intensive applications like AI or cryptographic accelerators. Beyond these examples are more ambitious approaches that leverage entirely different computing paradigms, such as quantum computing which leverages the uniqueness of quantum physics, or neuromorphic computing which uses novel materials to mimic biological neural systems.

Developing new chips is one challenge of heterogeneous computing. Perhaps a greater challenge, however, is the lack of well-established methods for integrating, validating, testing, manufacturing, and programming these systems. Optimizing the development and standardization of the underlying system software and data interconnects could make an impact to address these key challenges. In those cases where current firms have begun to leverage heterogeneous approaches in their devices, these methods have been developed in-house for use with a specific series of products. Although heterogeneous computing progress in large firms is promising, positive results are often proprietary to the firm in question and do not necessarily improve the broader innovation ecosystem for small and medium firms in the United States.

Although these challenges are daunting, the U.S. semiconductor industry sets the global pace for innovation and dedicates a substantial portion of its annual revenue into R&D efforts.[1] Scientific funding from agencies such as the National Science Foundation (NSF), the Department of Energy (DOE), and the Department of Defense (DoD), in partnership with the Semiconductor Research Corporation has created a research ecosystem among universities, federally funded research and development centers, and national laboratories that is second to none. Finally, Administration-level efforts such as the National Strategic Computing Initiative, the National Nanotechnology Initiative, and the National Quantum Initiative have been highly successful in coordinating many stakeholders towards the development of advanced, post-Moore computing. With the support for advanced packaging in the 2021 Endless Frontiers Act, and the authorization of a microelectronics-focused Manufacturing USA Institute in the 2021 National Defense Authorization Act, it is also clear that there is congressional support behind additional efforts in this area.

There is an opportunity for the Biden Administration to leverage the convening authority of the Federal Government and stand up a Manufacturing USA Institute which will help integrate, scale up, and standardize heterogeneous chiplets and the associated software into usable, consumer and edge devices. There are several advantages of doing this at a pre-competitive Manufacturing Institute. First, if these approaches are developed predominantly within large companies, where there is already ongoing work in this area, small and medium US firms will be unable to leverage such innovation. Second, the Manufacturing USA Institutes have done a very good job at bringing together private companies along with university and Federal Government stakeholders. The inclusion of private companies will help focus the downstream transition efforts into consumer-facing device applications. This is critical because the DoD and DOE already have in-house heterogeneous computing efforts which target military platforms and high-performance computing (HPC), respectively. Although the recommended Institute will certainly leverage ongoing basic research from these other efforts, the targeted applications should be consumer and municipal facing and not uniquely intended for military platforms or HPC systems.

Plan of Action

The Biden Administration has the opportunity to partner with Congress to allocate funding to a Manufacturing USA Institute focused on advancing packaging for scalable heterogeneous computing. This Institute would perform several related functions, none of which are currently performed by any active Manufacturing USA Institute. The Institute would aim to:

1. Develop advanced packaging technology: The need to integrate specialized chiplets or IP blocks in energy efficient or small-form-factor ways will require the Institute to develop superior packaging methods.

2. Develop software: As the toolbox of chiplets grows, new coding language practices and underlying system software will need to be developed to handle the novel computing paradigms and architectures while still allowing these new chiplets to efficiently communicate with previously developed components. Close integration of hardware and software design provides an enormous functional advantage and generalizing such capabilities to U.S. companies would provide a valuable competitive advantage.

3. Develop physical design and verification tools: Once novel materials or architectures can be manufactured with sufficient yield, the Institute should assist with the development of physical design and verification tools in order to reliably produce working chiplets of the given technology, thus creating a growing toolbox of specialized chiplets that can be leveraged by Institute members.

4. Scale-up manufacturing: During the scale-up process, the Institute should help advance manufacturing processes and process design kits (PDKs) relevant to identified computing technologies.

5. Co-design and down-select technologies: At the earliest stage of development, the Institute would help with the design and down-selection of potential computing technologies. In this way, novel computing paradigms can benefit from a comprehensive end-to-end optimization to provide the greatest benefit once a technology is ready to be scaled up. The Institute should collaborate with the Semiconductor Research Corporation, and the various government research efforts that seek to identify novel computing materials and architectures. Examples of these efforts are the DoD's STARnet and JUMP efforts as well as the NSF's nCORE and E2CDA efforts, as well as the Electronics Research Initiative at DARPA.

Many technologies are already at a sufficiently advanced stage that they will not need assistance on items 4 or 5, or possibly even 3. Examples of this are machine learning accelerators or cryptographic accelerators. The Institute should strive for early successes by integrating these chiplets into a standardized approach while the slower processes of down-selecting and scaling-up emerging technologies are done in

parallel. Ultimately, the Institute will be performing all of the identified activities at all times because technologies will advance at different rates.

The 2021 National Defense Authorization Act contains language authorizing several actions which could strengthen such an Institute, and it could be argued that this language authorizes the Institute described in this work. Section 9906 of the 2021 NDAA authorizes, among other items, a National Semiconductor Technology Center, a National Advanced Packaging Manufacturing Program, and the creation of a Manufacturing USA Institute under the National Institute for Standards and Technology (NIST). According to the language contained therein, the approximate function of these entities is to:

- Conduct advanced semiconductor manufacturing, design, and packaging research and prototyping;

- Advance testing and assembly, materials characterization, and production automation for next generation microelectronics;

- Develop and deploy educational and skills training to support the future U.S. semiconductor industry.

An advanced packaging Institute for Scalable Heterogeneous Computing fits entirely with these requirements. It is quite sensible to place such an Institute under the authority of NIST, and therefore the Department of Commerce, given the emphasis on consumer electronic innovation.

There is also significant support from industry stakeholders for increased government activity in this area. For example, the Semiconductor Research Corporation (SRC) released their Decadal Plan in 2021, with recommendations for where the government should focus efforts for the future of semiconducting technologies. As a part of this report, they identified what they call "five major seismic shifts" for information and communication technologies, stressing that we will need:

1. Fundamental breakthroughs in analog hardware in order to generate smarter devices that can sense, perceive, and reason;

2. Radically new memory and storage;

3. New research directions to address the imbalance of data transmission and generation;

4. Breakthroughs in security hardware research;

5. Novel computing paradigms with vastly improved energy efficiency.

The Institute described in this work would make contributions to scaling up technologies that would impact all five seismic shifts identified by the SRC. The SRC recommended that Congress assign $3.4B per year over the next 10 years towards

these efforts. As a component of that recommended amount, the Heterogeneous Computing Manufacturing Institute should receive roughly the same level of funding as the previous Institutes, roughly $100M in government funding to be matched by funding from private industry and other non-federal sources.

Implementation and Precedents

There are numerous precedents for a Scalable Heterogeneous Computing Institute, which will not only serve as models for how the Institute should be structured, but which will serve as critical partners, advisors, or stakeholders in the implementation of the Institute.

Other manufacturing Institutes: There are other Manufacturing USA Institutes which can serve as a guide for structuring this Institute, and which will provide valuable collaboration opportunities based on their specific expertise. The Advanced Functional Fabrics of America and NextFlex Institutes hold key insights into advanced packaging challenges which tie into their specific areas of expertise. The American Institute for Manufacturing Integrated Photonics and Power America Institutes each advance alternate computing materials and paradigms which will be valuable parts of the innovation pipeline through which novel approaches can be matured and integrated into broad, heterogeneous systems. Finally, the Advanced Robotics for Manufacturing, Clean Energy Smart Manufacturing Innovation Institute, Cybersecurity Manufacturing Institute, and Manufacturing times Digital Institutes can help identify the industry value-add and system level needs of advanced heterogeneous computing modules.

Prior research efforts: Government funded research efforts set a precedent for the long-established need for developing and maturing advanced computing materials, paradigms, and architectures. STARnet, nCORE, JUMP, and E2CDA are some examples of such programs. Beyond establishing a precedent for investment in these areas of research, they have produced a broad portfolio of advanced computing materials, devices, and architectures, which will need to be down-selected and scaled up.

Government initiatives: Many federal initiatives and programs have been implemented which will either contribute towards the success of or derive value from a Scalable Heterogeneous Computing Institute. The Electronics Resurgence Initiative, National Nanotechnology Initiative, Networking and Information Technology Research and Development Program, National Strategic Computing Initiative, and Brain Research Through Advancing Innovative Neurotechnologies are examples of such efforts. These initiatives will provide value for a Scalable Heterogeneous Computing Institute because they can convene key stakeholders across their relevant federal agencies and research communities to help identify organizational priorities.

Other coordinating organizations: Related, non-government organizations tied to the semiconductor industry will also serve in a critical role in providing the expertise needed to scale key technologies beyond the research stage, and engaging in other valuable information sharing. SEMATECH, the Semiconductor Research Corporation,

and Metal-Oxide-Semiconductor Implementation Service (MOSIS) are examples of such organizations.

Funding agencies: Although this Institute would be organized within the Department of Commerce via NIST, partnerships with other federal agencies would be critical, most notably the DOE, NSF, DoD, and NASA. Such partnerships would leverage existing and future research programs, help identify domain-specific use cases and opportunities, improve network effects of existing embedded entrepreneurship programs, and marshal current research and development infrastructure, such as that at the DOE National Laboratories or at certain FFRDCs. One excellent example of valuable existing infrastructure is the facility at the Sandia Microsystems Engineering, Science and Applications (MESA) complex.

By leveraging the organizations and efforts described above, the Scalable Heterogeneous Computing Institute should be able to leverage the existing U.S. research infrastructure to the fullest possible extent.

Conclusion

By making the investments recommended above, the Biden Administration can help ensure that the United States leads the future of computing innovation. A successful Scalable Heterogeneous Computing Institute will leverage decades of wise investments that the U.S. government and semiconductor industry has already made. Beyond this, a successful Institute will serve a critical role in solving the largest gaps in current computing technologies. By doing so, such an Institute will also ensure that the associated advances will broadly benefit the US innovation ecosystem, thus helping unleash the inventive potential of small and medium groups, firms, and companies within the United States.

Frequently Asked Questions

1. What could the computing industry look like in 5-10 years with the implementation of this Institute?

On the current economic trajectory, the U.S. faces a critically high risk of continued expansion of offshore IC and package fabrication relative to domestic capabilities. International adversaries and business competitors take the development of these industries very seriously and even with the right institutional support, strong and consistent leadership beyond the scope of any single agency or Institute will be key to empowering domestic economic competitiveness and innovation leadership. Within that framework, such an Institute would be one component, albeit an important one, of a much broader competitiveness strategy. Ideally, this Institute would serve a key role in helping achieve a broad industry goals such as:

- Ensuring that the current trends of the industry towards integration of heterogeneous components are leveraged effectively in order to eliminate possible speed bumps to innovation. Put more simply, an important central goal is to get everybody on the same page more quickly. This could mean that new products or technologies come to the consumer market several years earlier than they would in a world without such an effort.

- Enabling the critical technology developments in the area of heterogeneous integration to occur within the United States. The global microelectronics industry is at a pivotal point and decisive action in the present could mean the difference between the U.S. losing its edge in semiconductor innovation, or becoming the global center for electronics manufacturing. Implementation of this Institute will help ensure that the latter case is realized.

- Establishing that the processes, technologies, and intellectual property generated in this area are accessible to small and medium firms, rather than existing solely within the largest companies. By coordinating this work at a Manufacturing USA Institute, the early and mid-stage research can be performed in a pre-competitive setting, thus allowing broader access to a wider range of U.S. firms, and more fully unlocking the innovative potential of technology development. This is not to say that the efforts of this Institute will be at odds with the interests of larger U.S. companies, as they will benefit greatly from an enhanced innovation ecosystem and a reduction of the R&D burden on those firms.

2. What kinds of private partners might the Institute collaborate with?

It will be critical for this Institute to collaborate with private partners at all levels. It is difficult to provide a definitive list, but some examples of important partnerships are:

- Large entities such as Intel, IBM, NVIDIA, and others will be critical partners, as they already have a long-established interest in the core technologies;

- Industry associations like SEMI or the Semiconductor Industry Associate (SIA) will be critical in helping identify how specific industry needs evolve over time;

- Industry consortiums like SEMATECH, or the relatively new Quantum Economic Development Consortium will be important for similar reasons;

- Design tool vendors, such as Cadence, Synopsys, and Mentor, will be some of the most important potential collaborators as they possess the greatest expertise in chip design and integration rules.

3. What are more examples of possible heterogeneous chiplets?

Some examples of chips which are currently or might eventually be relevant are:

- Central processing unit (CPU) – A CPU is versatile, general purpose computing engine, sometimes described as the "brain" of the computer. The CPU is able to handle a very wide variety of tasks, and typically runs on a small number of high-performance computing cores.

- Graphics processing unit (GPU) – A GPU is a more specialized computing engine than a CPU, and is typically designed to have a very large number of lower-performance computing cores. While it cannot perform the same variety of tasks that a CPU can, it is able to perform extremely repetitive tasks (such as graphics rendering) in a massively parallel fashion by leveraging its large number of computing cores. By doing so, it is able to perform a certain set of tasks much more quickly than a CPU.

- Hardware accelerators – Sometimes there are tasks that can be run via software on a CPU or GPU, but which require an enormous amount of resources to be run effectively. In these cases, it is sometimes more effective to design a chiplet to perform a certain, narrow set of tasks in a very efficient way as defined within the hardware. This makes the chiplet relatively inflexible in the tasks that it can perform, but extremely resource-efficient. A computing engine which has its function defined by its hardware is sometimes called an application-specific integrated circuit (ASIC).

   - Artificial intelligence (AI) accelerator – A hardware accelerator dedicated to accelerating machine learning or AI applications.

   - Cryptographic accelerator – A hardware accelerator dedicated to performing cryptographic functions, such as the encryption of a data stream.

- Non von Neumann chiplets – In a typical computing architecture, the processing cores are physically separated from the memory. When the computing cores must load a program or access other data, they will send instructions to receive that data from memory. This architecture is called the "von Neumann architecture", named after John von Neumann. The electronic

highway that carries that data is called a "bus" and it has a limited data bandwidth. This is referred to as the "von Neumann bottleneck" and in many modern applications, such as deep neural networks and machine learning, the majority of computing energy is spent on transferring data across this bus. There are many proposed architectures which circumvent the von Neumann bottleneck. Two examples are:

- o Monolithic 3D integration of compute/memory – Monolithic 3D integration is a fabrication process where tiers of circuits, such as logic and memory, are fabricated on top of one-another. This can allow an extremely high density of interconnections between different layers, and thus a much higher data bandwidth than a typical von Neumann architecture.

- o Neuromorphic computing – An architecture loosely inspired by a biological brain, in which logic elements perform computation and memory by iteratively adjusting the strength of connection to nearby elements, much like biological neurons. In this way, they perform both a logic and memory function. There are many other proposed technologies that use the same physical element to perform both memory and logic operations, which inherently overcomes the von Neumann bottleneck because there is no physical separation between the material elements that perform such operations. This is one example of such a technology.

- Alternate computing approaches – Some approaches to computing are so different from current technology, that they will require completely different hardware than current silicon-based electronics. Two examples are:

- o Quantum computer – A quantum computer is a computing element which leverages the properties of quantum states, such as quantum entanglement and superposition to perform specialized computations. Although this is still an emerging field, it is expected that quantum computers will be able to perform certain important tasks far more quickly than conventional computers.

- o Optical computer – Optical or photonic computing uses waves of light to perform computational tasks, similar to how conventional computers use electrons. Advantages of photonic computing potentially include higher speeds, greater scalability, and greater information storage when compared to current technology.

About the Author

Eric Breckenfeld is a Lead Scientist at Booz Allen Hamilton where he serves as a consultant for predominantly U.S. defense clients, including DARPA and the U.S. Naval Sea Systems Command in the areas of electronic materials/components and hardware supply chain security. Prior to his work at Booz Allen Hamilton, Eric was an AAAS Science and Technology Policy Fellow at the National Nanotechnology Coordination Office where he managed the Sustainable Nanomanufacturing and Nanoelectronics for 2020 and Beyond Signature Initiatives. Previously, Eric was a National Research Council Fellow at the Naval Research Laboratory where he performed basic and applied research on electronic materials. Eric received his BS in Engineering Physics from the University of Wisconsin-Madison and his Ph.D. in Materials Science and Engineering from the University of Illinois Urbana-Champaign.

About the Day One Project

The Federation of American Scientists' Day One Project is dedicated to democratizing the policymaking process by working with new and expert voices across the science and technology community, helping to develop actionable policies that can improve the lives of all Americans, and readying them for Day One of the next presidential term. For more about the Day One Project, visit **dayoneproject.org**.

The Day One Project offers a platform for ideas that represent a broad range of perspectives across S&T disciplines. The views and opinions expressed in this proposal are those of the author(s) and do not reflect the views and opinions of the Day One Project or its S&T Leadership Council.