Decadal Plan for Semiconductors: New Compute Trajectories for Energy Efficiency

Question: Victor, in your three technical priority bubble diagram, you have Memory, Logic and Interconnect. If energy efficiency (and thermal design) becomes a limiting factor, the issue goes beyond interconnection of data. It becomes interconnection of energetic in semiconductors. Does this not portend a new need for thermal design/packaging innovations as a priority to address future physics-based limitations to computation?

Answer: Certainly, thermal management is a critical issue, especially as we go 3D, but is not really a new issue. The circles on the bubble diagram represent three main physical components of computing, I would not describe them as 'priorities' though. Heat is byproduct of computation and we expect to reduce heat production by utilizing new compute trajectories. One hypothesis is that new organization of the Logic-Memory-Interconnect triad could get us on a new trajectory. However, we do expect that thermal management will be a key component of system co-design going forward- and new techniques for thermal management appropriate to 3D will need to be developed.

Question: Can speakers highlight which projects/questions were impacted or driven by USG Federal Funding?

Answer: After a competitive procurement award process, the DOE Exascale Computing Project made architecture investments with industry in order to close gaps in companies' technology roadmaps to help them meet the DOE ECP high level performance goals. The total DOE investment was over $250M for the total portfolio of 3 year projects. Since the industry partners provided a cost share match at the 40% level, these R&D projects had a cumulative total value of over $440M and were granted DOE advance waiver to IP generation. The DOE used this approach to encourage industry partners to commercialize these co-designed advanced architecture technologies via separately competed DOE exascale system contract awards. DARPA and NSF, in partnership with SRC through programs like JUMP and nCORE, also establish pre-competitive university R&D projects that benefit all SRC member companies and the ecosystem. DOE and NSF have also awarded quantum computing R&D projects with several companies, as well as a large number of DOE National Laboratories and University researchers.

Question: Where do emerging devices such as spin-based and ferro-electric fets and integration with silicon play into the bigger bits/s to MIPs transformation ?

Answer: Spin based, ferro-electric based, and other novel devices may provide energy efficient compute capability when they are matured to manufacturing readiness. Some of these solutions may not be stand alone / platform capable and might be integrated as complimentary along with traditional CMOS. These heterogeneous solutions might become tenable provided they attain significant system-level cost-effective energy-efficient improvements otherwise not possible while also delivering required computing performance. Computing architectures that take advantage of novel device functionalities may also be critical knobs for overall performance and energy-efficiency at the system-level.

Question: Quantum computing seems to be promoted as a technology for solving really hard problems. Does this make quantum computing a niche technology in the best case scenario? i.e., it is likely that large federal agencies will be its primary beneficiaries and not the average consumer of computing technology.

Answer: The highest impact of Quantum Computers will be where they can bring the power of quantum physics to full bear, i.e., in applications where best use of superposition, entanglement and interference is made. It is an active field of research to explore and develop new quantum algorithms and proof the advantage of QC compared to classical computers. There are many difficult problems in science and business where quantum computing can make a difference in the future, e.g., in simulating quantum systems, machine learning, database search, Quantum Monte Carlo simulations, optimization, and more. These are not niche applications but at the core of many business problems in industries such as transport and mobility, oil & gas, finance and insurance, any industry where materials innovation is important etc.

Question: My question is for Victor. For p=2/3, it seems to be for general purpose processors. If we look at domain-specific processors, would p value change 9for the better)?

Answer: Yes, the p=2/3 is a measure of energy efficiency progress for converting energy to MIPS. We see the trajectory changes when GPUs replace General Purpose CPUs, because for the same performance level (as defined by equivalent MIPS) GPUs consume less energy. This is because General Purpose CPUs spend a large portion of the consumed energy to optimize instruction issue and extract parallelism through speculative operations from scalar programming code. GPUs rely on explicit parallelism in the application which improves efficiency. Domain-specific processors extend this concept to optimize the micro-architecture to a smaller class of algorithms, allowing greater power efficiency.

Question: With the prospect of billions of new federal spending on semiconductor research, where do you see the greatest potential for filling in the road map?

Answer: A more detailed discussion and examples of specific research directions, can be found in the Decadal Plan for Semiconductors document - https://www.src.org/about/decadal-plan/

Question: AI is very energy intensive, yet it is being thrown willy-nilly at all manner of problems. It is important not to neglect those avenues of research that attempt to solve the same problems without the need for AI.

Answer: There are new concerns about Green AI that are calling attention to the carbon footprint of compute intensive AI/ML methods that rely on big data. The DOE is examining how AI/ML methods can be applied to scientific discovery that supports the ability to integrate deep learning methods with scientific simulations described by the solution to a system of governing equations. While there are certainly cases where large volumes of science experiment data are collected, often data collection from experiments are expensive, or are rare events, so test data may be limited. For these scientific discovery challenge problems, the DOE is working on the definition of converged applications that integrate

traditional scientific simulations with AI/ML, uncertainty quantification, graph analytics, and other forms of data analytics. For the commercial world, AI is still in the infancy as the industry explores the best use cases, relying mainly on programmable HW approaches for fastest time-to-market. As AI applications mature, domain specific architectures will likely emerge that can greatly improve the energy efficiency by orders of magnitude.

Question: Quantum computing will only be possible in a data center. What about edge computing? Won't most of the computation be done at the edge? Will I ever wear a quantum computer on my wrist?

Answer: The early classical computers in the 1940s and 1950s were huge and sometimes occupied entire buildings. They were probably bigger than today's quantum computers. The "wrist computer" was unthinkable in 1950. As technology advanced and application space grew, we saw transition from huge mainframes to desktops, notebooks, and Apple watches. One can imagine the same may happen with the quantum computers.

Question: The 2/3 scaling law is a classic chemical engineering scaling law when volume to surface area ratios are a limit. Scaling the volume of a chemical reaction follows 2/3 when the mass and heat transfer limited reactions are in continuous processes through a pipe. The area (ability to dissipate heat) goes with the square of the pipe surface area, where the volume (the ability to transfer mass) goes with the cube of the radius. Since we are heat limited by the transfer of heat and electrons (mass and charge), the 2/3 law is likely the same. (Friendly amendment to prior comment: we are ultimately limited but the transfer of heat, spin and charge, not just heat and charge.)

Answer: This is a good and stimulating question. The surface/volume scaling considerations are basis of the famous Rent's rule which states that by optimized circuit design with minimization of interconnect length, there is a relation between the number of external terminals T and number of 'cells' (e.g. transistors), N, in the circuit. T=tN^r, where r is close to 2/3. The Rent's rule hypothesis was mentioned at the webinar as one of the research directions. Certainly, the analogy with general heat, charge etc transfer law is illustrative.

Question: The R&D typically comes with both research and manufacturability. For the universities the funding has become less and hence hard to get post-docs or students in the microelectronics. There needs to be excitement on the research (academia), R&D (e.g. industry), and Manufacturing as in the 1990s. So the question is how does this fit in the bigger picture. The reason for this question is if there is not enough thought into all aspects, in 20 years, we could be having the same discussion about the US losing leadership….

Answer: Workforce development is a key part of the discussion around new infrastructure, which aligns with your point about creating excitement across academic research, industry development, and into manufacturing. Standing up infrastructure to serves the needs across the technology transfer gap which can engage students and industry technologists in a shared development facility could go a long way towards a strong pipeline of workforce development.

This question makes an excellent point. Both graduate and undergraduate engineering students choose career directions in areas that have a good prospect for a future well paying job. Without an accompanying U.S. industry driving demand for experts in the field, state-of-the-art R&D in the U.S. may not be sustainable.

###

About SIA
The Semiconductor Industry Association (SIA) is the voice of the semiconductor industry, one of America's top export industries and a key driver of America's economic strength, national security, and global competitiveness. Semiconductors – the tiny chips that enable modern technologies – power incredible products and services that have transformed our lives and our economy. The semiconductor industry directly employs nearly a quarter of a million workers in the United States, and U.S. semiconductor company sales totaled $208 billion in 2020. SIA represents 98 percent of the U.S. semiconductor industry by revenue and nearly two-thirds of non-U.S. chip firms. Through this coalition, SIA seeks to strengthen leadership of semiconductor manufacturing, design, and research by working with Congress, the Administration, and key industry stakeholders around the world to encourage policies that fuel innovation, propel business, and drive international competition. Learn more at www.semiconductors.org.

About SRC
A not-for-profit research and development (R&D) company committed to redefining possible® in the areas of defense, environment and intelligence. Our mission is to help keep America and its allies safe and strong by protecting its people, environment and way of life. We do this by focusing on our customers' needs through the innovative application of science, technology and information to solve problems of national significance.