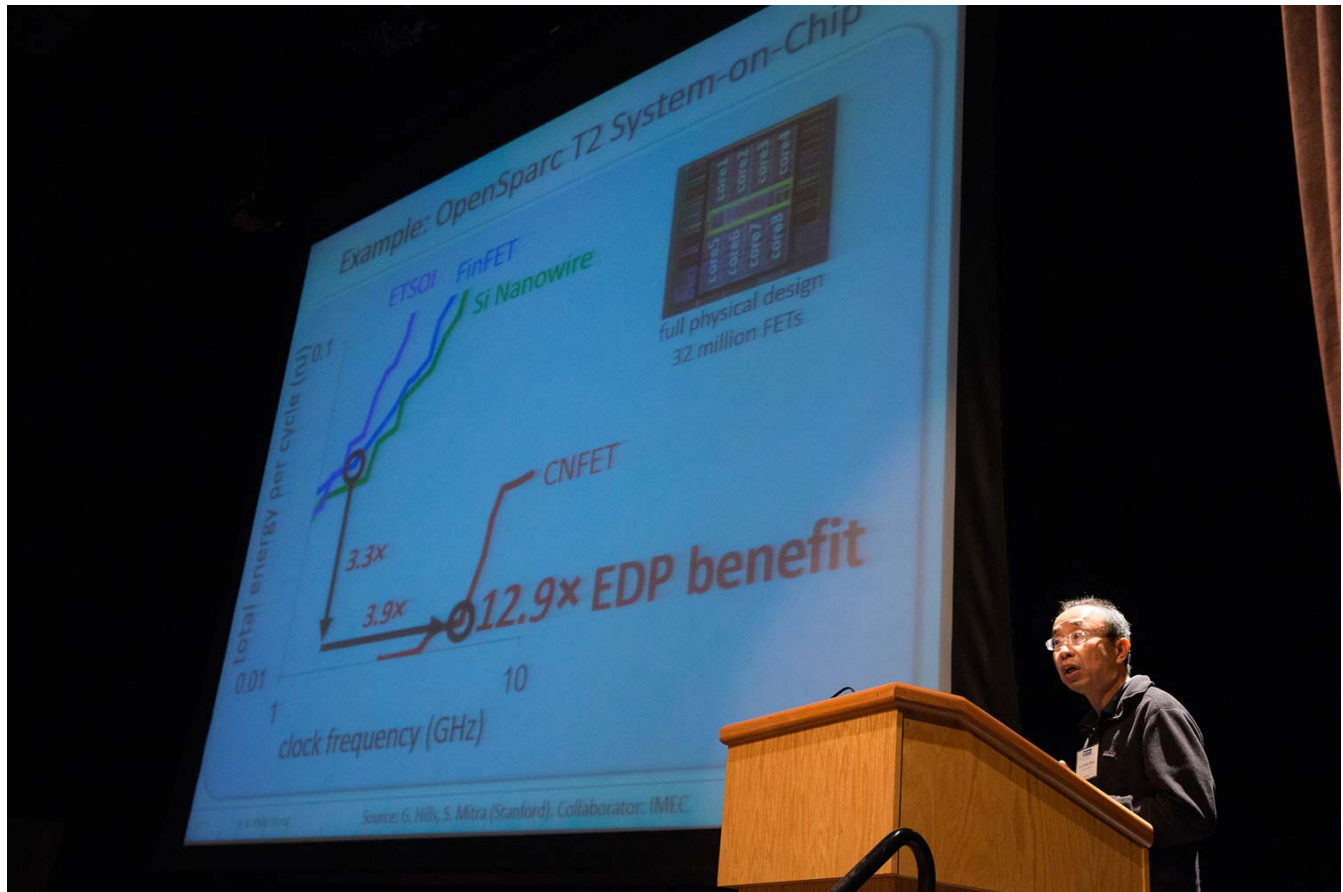May 16, 2017 | By Tiffany Fox

NVMW 2017 — Brain-Inspired Memory "At the Center of the Universe"

In his keynote presentation at the 2017 Non-Volatile Memories Workshop, Stanford Electrical Engineering Prof. H.S.-Philip Wong made the case for a new type of 'brain-inspired' silicon-compatible 3D chip that provides massive storage and improves efficiency by reducing computational 'sprawl.' See more photos of NVMW. Photos by Alex Matthews for the Qualcomm Institute

It's been decades in the making—with hundreds of academic papers and new technologies to show for it—but non-volatile memory (NVM) has now become firmly established as the memory of choice for long-term persistent data storage. It was no doubt a refreshing change of pace, then, for attendees at this year's annual Non-Volatile Memories Workshop to spend time discussing not if NVM will succeed, but instead how to make NVM even more powerful (or even how to make it, in one presenter's bold vision, the center of the computing universe).

NVM, in its most basic sense, is what makes it possible for you to turn off your computer (or unplug your flash drive, a form of NVM) and have it still 'remember' the last draft of the book you're working on. It's also relatively low-power, boasts fast random-access speed and is known for being rugged since it doesn't require a spinning disk or other moving mechanical parts.

The latest Non-Volatile Memories Workshop was held March 12-14 at the University of California San Diego and brought together both academic researchers and industry representatives from Intel, Facebook, IBM, Microsoft and other major corporations to discuss, for the eighth year running, how to best store increasingly large amounts of data in smaller spaces. Ideally, this would also be done at faster data transfer speeds and at lower cost to the consumer. While those three aspects will likely always be at the forefront of NVM development, how NVM will evolve—and how computing itself will change as the result of advances in NVM—is anyone's guess.

Stanford Electrical Engineering Professor H.S.-Philip Wong has a hunch, however, and he and others at NVMW think the answer can be found in one of the most powerful and complicated computational tools in the known universe: The human brain.

In its 2015 grand challenge to develop transformational computing capabilities using nanotechnologies, the White House called the human brain "a remarkable, fault-tolerant system that consumes less power than an incandescent light bulb." Could the IT industry, which has long struggled to make computers more energy-efficient, rise to the occasion and create a scalable, fast and 'green' computation platform that could rival the gray matter inside our skulls?

Wong his colleagues (including UC San Diego Bioengineering Professor Gert Cauwenberghs) were up to the challenge, but they knew that for brain-inspired computing to succeed, the energy-demand of memory and accesses would have to be addressed. In his keynote presentation at NVMW, Wong described the development of his team's end-product: "The Nano-Engineered Computing Systems, or N3XT Technology for Brain-Inspired Computing," a new type of silicon-compatible 3D chip that provides massive storage and improves efficiency by reducing "computational sprawl."

Building 'Up' Instead of 'Out'

Take a look inside a modern computer processor and it's not too difficult to imagine it as a suburban neighborhood filled with single-story ranch-style homes, where wires connect chips much the way roads connect structures.

But suburban neighborhoods are also notorious for "sprawl" and all the problems that go along with it: wasted space, long commutes and traffic jams. Similarly, inefficiently designed chips lead to data bottlenecks and longer processing times—something that won't pass muster among consumers increasingly accustomed to ultra-fast computing for gaming, video-streaming and other high-demand applications.

Paul Siegel, professor at UCSD's Center for Memory and Recording Research and NVMW co-organizer, said that brain-inspired computing will "impact the ways that we – and a new generation of intelligent robots – communicate, make things, travel, and even express ourselves artistically."

Just as city planners have long called for building "up" instead of "out," Wong and his team architected the N3XT (pronounced 'next') microchips in the same way, building super-dense skyscraper-style chips that integrate processors with memories and increase efficiency by a factor of a thousand.

"In two-dimensional systems, energy is spent on memory access and being idle," explained Wong at the workshop. "But with 3D memory-on-chip and high bandwidth, you are only limited by compute computation, and that's what we want."

Wong further explained that this type of "hyperdimensional computing" could ultimately combine biologically-inspired algorithms and machine-learning algorithms with neuromorphic hardware and conventional hardware (CPUs, GPUs, supercomputers, etc) to create neuromorphic chips—chips that can simulate synaptic function in the brain to perform "one-shot learning" that wouldn't require training over time.

"N3XT takes to the next level the concept of low-power, volume-efficient, '3D' circuitry that is already revolutionizing NVM design," said Paul Siegel, professor at UCSD's Center for Memory and Recording Research, a NVMW co-organizer and an affiliate of the UC San Diego Qualcomm Institute. "By tightly integrating memory and logic, it paves the way for future applications of machine learning and artificial intelligence, technologies that have transformed the fields of natural language processing, e-commerce, self-driving vehicles, and personalized health."

Added Siegel: "As a platform for the next generation of brain-inspired computing, N3XT opens the door to a flood of novel real-time computing applications that will profoundly affect how we reason about and interact with the world. These advances will impact the ways that we - and a new generation of intelligent robots - communicate, make things, travel, and even express ourselves artistically."

Memory-centric Computing: "Music to Our Ears"

It's one thing to rethink chip architecture—it's another thing altogether to suggest a complete overhaul of how computers are built, from top to bottom. But that's essentially what Western Digital Chief Technology Officer Martin Fink proposed in his keynote talk, delivered on the second day of NVMW before technical sessions on database and file systems, coding for structured data and other specializations. (The previous day's technical sessions included talks on operating systems and runtimes for NVM, error modeling and error-mitigation codes as well as architecture and management policies).

In his keynote presentation on day two of NVMW, Western Digital Chief Technology Officer Martin Fink proposed that engineers rethink the entire computing stack "from the ground up" by architecting chips to be completely memory-centric, where computer memory rather than the CPU "is at the center of the universe."

Fink began by discussing one of the stark realities of computing: Networks are increasingly unable to handle the high rate of data growth and are simultaneously becoming more expensive to operate. Storage, meanwhile, continues to diminish in cost while increasing in density, which suggests the next logical step is to move toward memory-centric computing.

To achieve this, Fink—the former director of HP Labs—proposed that engineers must rethink the entire computing stack "from the ground up" by architecting chips to be completely memory-centric, where computer memory rather than the CPU "is at the center of the universe" (something Siegel remarked "should be music to our ears"). Petabyte-scale memory, in other words, would be its own independent system (likely a type of memory known as resistance RAM or ReRAM), with CPUs attached to the memory as needed. Instead of moving the data to the CPUS, in other words, the CPUs would move to the data.

Fink admitted that "rethinking how we've done things since 1945 is really, really hard. Rather than the CPU being the center of the universe," he continued, "we need to flip the model to minimize the travel of data through the storage memory hierarchy and instead distribute the

CPUs and have 'near' and 'far' data with memory storage at the center."

Fink likened this change in paradigm to the disruption that arose when scientists believed that the universe revolved around the Earth "and then Copernicus came in and turned that on its head."

"There is a lot of work across the ecosystem that needs to happen, a lot of invention required," Fink added, encouraging those in attendance at NVMW to write systems and applications software that rethinks the current hierarchy.

NVMW featured talks and presentations by representatives from various industry partners, including Facebook, IBM, Microsoft and Intel, which was represented by Stephanie Jones, a storage engineer in Intel's NVM group.

"Martin has the right idea – the changes that NVMs require are dramatic and we have found that once you start unraveling them, they reach almost every aspect of computer system design," said Steve Swanson, director of UCSD's Non-volatile Systems Laboratory a co-organizer of the workshop and a QI affiliate.

Fink acknowledged that modern computing is built on mathematician John Von Neumann's concept that a computer's program and the data it processes can be stored in the computer's memory, but he also noted that Von Neumann proposed this approach in 1945 when computer memory was a scarce resource. Added Fink: "If you had access to all of the existing compute technology, but were building the world's first computer, would you build it the same way? I think the answer is no. It's time to let go of all the conventional thinking."

MEDIA CONTACT

Tiffany Fox, 858-246-0353, tfox@ucsd.edu

UC San Diego's Studio Ten 300 offers radio and television connections for media interviews with our faculty, which can be coordinated via studio@ucsd.edu. To connect with a UC San Diego faculty expert on relevant issues and trending news stories, visit https://ucsdnews.ucsd.edu/media-resources/faculty-experts.