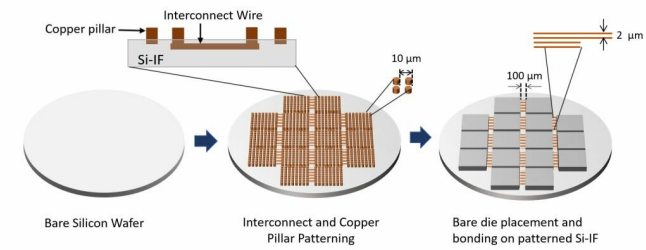
GPU news: Time for another go at waferscale computer

4 February 2019, by Nancy Cohen



The system assembly process flow is shown. Interconnect layers and copper pillars are made by processing the bare silicon wafer. Bare dies are then bonded on the wafer using TCB. Credit: Architecting Waferscale Processors - A GPU Case Study, HPCA 19.

Researchers at the University of Illinois at Urbana-Champaign and the University of California, Los Angeles, are behind the recent development for a wafer-scale computer that aims to be faster, more energy efficient, than contemporary counterparts.

Engineers aim to use something called "silicon interconnect fabric" to build a computer with 40 GPUs on a single silicon wafer. *TechSpot* and other sites reported on their work and their paper, to be presented this month.

Some background on Si-IF: "Over the past two decades, silicon chips have decreased in size by 1000x, while packages on circuit boards have only shrunk by 4x," said UCLA Technology Development Group. A solution is "silicon interconnect fabric (Si-IF)."

Samuel Moore at *IEEE Spectrum* has a much quoted article on the topic where he noted results: "Simulations of this multiprocessor monster sped calculations nearly 19-fold and cut the combination of energy consumption and signal delay more than 140-fold."

Namely, the research effort is among members of the department of electrical and computer engineering, University of California, Los Angeles, and department of electrical and computer engineering, University of Illinois at Urbana-Champaign. Their paper is titled "Architecting Waferscale Processors—A GPU Case Study."

Illinois computer engineering associate professor Rakesh Kumar and his colleagues already started work to build a prototype waferscale prototype processor system. The group will explore it further for insights as to any problems that may arise. They believed the time was ripe to revisit waferscale architectures.

Mark Tyson in *Hexus*: "Engineers at the University of Illinois Urbana-Champaign and at University of California Los Angeles think it is time to have another attempt at creating a wafer-scale computer."

The accent can be put on the word *revisit*. The team wrote in their paper, "Unsurprisingly, waferscale processors were studied heavily in the 80s. There were also several commercial attempts at building waferscale processors. Unfortunately, in spite of the promise, such processors could not find success in the mainstream due to yield concerns."

They said "the larger the size of the processor, the lower the yield—yield at waferscale in those days was debilitating. We argue that considerable advances in manufacturing and packaging technology have been made since then and that it may be time to revisit the feasibility of waferscale processors."

Illinois computer engineering associate professor Rakesh Kumar and his collaborators are set to make the case for a waferscale computer consisting of as many as 40 GPUs. The best headline to remind us why this is interesting can be found at *IEEE Spectrum*. "What's better than 40

GPU-based servers? A server with 40 GPUs."

What's special: They have standard GPU chips that passed quality testing—they are creating a technology they call silicon interconnect fabric (SiIF) to better connect them.

Shawn Knight in *TechSpot* wrote about this. "With such tight integration," said Knight, "from the perspective of the programmer, it would look like one giant GPU rather than 40 individual GPUs."

SiIF replaces the circuit board with silicon; there is no need for a chip package, said Moore. He reported that in one design they were able to squeeze in 41 GPUs. "They tested a simulation of this design and found it sped both computation and the movement of data while consuming less energy than 40 standard GPU servers would have."

Tyson wrote that "as many HEXUS readers will know, usually supercomputers spread applications over hundreds of GPUs on separate PCBs, communicating over long-haul links. Such links are slow and energy inefficient compared to interconnects within the chip architecture." He noted that Kumar spoke of getting data from one GPU to another as creating an incredible amount of overhead.

IEEE Spectrum's Moore explained their work in more detail.

"The SiIF wafer is patterned with one or more layers of 2-micrometer-wide copper interconnects spaced as little as 4 micrometers apart. That's comparable to the top level of interconnects on a chip. In the spots where the GPUs are meant to plug in, the silicon wafer is patterned with short copper pillars spaced about 5 micrometers apart. The GPU is aligned above these, pressed down, and heated. This well-established process, called thermal compression bonding, causes the copper pillars to fuse to the GPU's copper interconnects. "

Their work drew favorable comments. Tyson called it a brave but possibly timely move for the industry.

What's next? The team will present their findings at the IEEE International Symposium on High-Performance Computer Architecture. The event is from Feb 16 to 20 in Washington DC.

More information: Architecting Waferscale Processors - A GPU Case Study, passat.crhc.illinois.edu/hpca19_cam.pdf