EDACAFÉ

New fabs; Tesla training system; GaN CMOS; CUDA on Risc-V; Y Combinator startups - September 10, 2021

EDACafe Editorial

Roberto Frazzoli

Roberto Frazzoli is a contributing editor to EDACafe and a seasoned freelance journalist specialized in electronics. His weekly contribution to EDACafe – since early 2019 - includes news updates, coverage of major industry events, special reports, and interviews. Prior to joining EDACafe, Roberto … More »

No shortage of interesting news this week – both from the industry and from academia – still it's worth devoting some space to a Tesla event that took place last month, showing how system-level assembly technologies can make a difference for supercomputers.

Fab and foundry updates: Intel, SMIC, X-Fab, Samsung

Intel's CEO Pat Gelsinger has reportedly said the company will invest up to €80 billion over the next decade to build new chip fabs in Europe.

According to the Wall Street Journal, **SMIC** is teaming up with the Shanghai government to build an $8.87 billion chip production line in the city.

Germany-headquartered **X-Fab Silicon Foundries** is now able to support volume heterogeneous integration via Micro-Transfer Printing (MTP), thanks to a licensing agreement with X-Celeprint. Technologies that may be combined include SOI, GaN, GaAs and InP, as well as MEMS. X-Celeprint's pick-and-place MTP technology stacks and fans-out ultra-thin dies.

Samsung Electronics has reportedly chosen the city of **Taylor, Texas**, as the site for its planned $17 billion new chip plant – which will be four times larger than Samsung's existing fab in Austin. To attract the investment, the city of Taylor has reportedly offered Samsung a $314 million worth of tax breaks for the next ten years.

Tesla's training chip and system

As widely reported by the media, at the recent Tesla AI Day the carmaker described its new D1 machine learning chip and the training system built around it. To briefly summarize the presentation given by Ganesh Venkataramanan – fully available in this CNET video – the D1 is a 50 billion transistors chip built in seven nanometer technology, based on an architecture that uses a number of optimizations such as four-way multi-threading and a custom instruction set. According to Tesla, the D1 achieves 362 teraflops in BF16 and more than two times the bandwidth coming out of the state-of-the-art networking switch chips. Venkataramanan said that the D1 was entirely designed by Tesla team internally. Tesla then built a 'training tile' consisting of a large multichip module that integrates twenty-five known good D1 dies onto a fan-out wafer. The innovative solutions specifically developed by Tesla to build the 9-petaflops training tile include “a new way of feeding power vertically: we created **a custom voltage regulator module that could be reflowed directly onto this fan-out wafer** (…) We brought PCB level technology of reflow on to this fan-out wafer technology,” Venkataramanan pointed out. Assembling multiple training tiles, Tesla is planning to build a **1-exaflops (10^{18})** machine learning training supercomputer.

Tesla’s Ganesh Venkataramanan showing a ‘training tile’. Credit: Tesla

Integrating GaN CMOS logic with GaN power devices

Researchers at the Hong Kong University of Science and Technology have recently developed a technology to build gallium nitride complementary MOS logic circuits, paving the way to the integration of power switching devices and peripheral circuits on a single chip. The team used a commercially available GaN-on-Si power HEMT (high-electron-mobility transistor) platform, commonly employed for power electronics. Key achievement was the development of a GaN p-channel FET technology – obtained through a new oxygen plasma treatment technique – as all current GaN HEMTs are n-FETs. The team demonstrated a series of GaN CMOS elements including NOT, NAND, NOR gates, a latch cell and ring oscillators – as well as the feasibility of their integration with power devices. Downscaling will be needed to reach higher operating frequencies.

Executing CUDA code on Risc-V GPUs

A research team from Georgia Institute of Technology and Seoul National University has proposed a pipeline that can execute CUDA source code on a Risc-V GPU architecture. The pipeline consists of several steps using intermediate representations (IR): translates CUDA source code into NVVM IR, converts NVVM IR into SPIR-V IR, forwards SPIR-V IR into POCL to get Risc-V binary file, and finally executes the binary file on an extended Risc-V GPU architecture. Except for the CUDA toolkit – an Nvidia product which is required to compile NVVM – all other components are open-source and can be easily found in Github. The team has also built a translator from NVVM into SPIR-V.

Y Combinator startups: liquid cooling for datacenters, moving transistors to BEOL

Y Combinator, the Mountain View-headquartered seed money startup accelerator, has recently held its Summer 2021 Demo Day. Although a large part of participating startups are app developers, at least two of them are directly related to the hardware and semiconductor world: Ferveret and Zinite. **Ferveret** develops liquid cooling technology for datacenters inspired by nuclear power plant cooling. The company claims that its solution saves datacenters 96% in cooling costs, 68% on capital costs and reduces their carbon footprint by 40% while increasing chip performance by 2X. **Zinite** is developing a transistor technology that allows to build transistors in the middle-end-of-line and back-end-of-line (MEOL/BEOL), in alternative to the usual front-end of line (FEOL) placement. According to the company, this solution provides simpler and better heterogeneous integration.

Applied Materials' innovations for chiplet packaging and 200mm SiC wafers

Applied Materials has introduced advanced software modeling and simulation tools accelerating Die-to-Wafer Hybrid Bonding. The company has also announced a joint development agreement with EV Group (EVG) to develop co-optimized solutions for wafer-to-wafer bonding. Other updates from Applied Materials include new products that help silicon carbide chipmakers transition from 150mm to 200mm wafer production: the Mirra Durum CMP system, that dramatically reduces finished wafer surface roughness; and the VIISta 900 3D hot ion implant system, that injects ions with minimal damage to the lattice structure, resulting in a dramatic reduction in resistivity. Details can be found in the presentation from the recent company's 2021 ICAPS and Packaging Master Class.

The VIISta 900 3D hot ion implant system. Credit: Applied Materials

Memory compression IP specialist gets new funding

Sweden-based **ZeroPoint Technologies** has recently closed a €2.5 million seed round, from Nordic VC Industrifonden and present shareholders, to finance the company into 2021 and beyond. The company claims to provide "the world's only available **real-time memory compression IP block** for System on Chips, effectively doubling a computer's main memory capacity and memory bandwidth." ZeroPoint Technologies is a spinout from Chalmers University of Technology in Gothenburg, Sweden.

Acquisitions

California-headquartered **Indie Semiconductor**, a provider of automotive solutions, has signed a definitive agreement to purchase Canada-based **TeraXion**, a provider of low noise lasers, Bragg gratings and integrated photonics. Key synergies are in the development of LiDAR systems.

Upcoming events

MIPI DevCon 2021 will be held virtually on 28-29 September.

https://www10.edacafe.com/blogs/editorial/2021/09/10/new-fabs-tesla-training-system-gan-cmos-cuda-on-risc-v-y-combinator-startups/