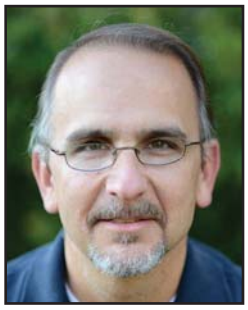
Session 3 Overview:
Highlighted Chip Releases: Modern Digital SoCs

INVITED PAPERS



Session Chair:
Thomas Burd
AMD, Santa Clara, CA

Session Co-Chair:
Rangharajan Venkatesan
Nvidia, Santa Clara, CA

Session Moderator:
Dennis Sylvester
University of Michigan, Ann Arbor, MI

This session highlights three major new Systems-on-Chip released recently, spanning several application areas. The invited product papers reside at the bleeding edge within the exciting fields of gaming, machine-learning accelerators, and data-center GPUs. The papers delve into practical system-related topics, mass-production related challenges and solutions (e.g., system interconnection design decisions, thermal/voltage/acoustic issues, packaging, etc.) in addition to circuit content, software interaction, and silicon measurement results.

2021 IEEE International Solid- State Circuits Conference (ISSCC) | 978-1-7281-9549-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISSCC42613.2021.9365755

8:30 AM

3.1 XBOX Series X: A Next-Generation Gaming Console SoC

Paul Paternoster, Microsoft, Sunnyvale, CA

In Paper 3.1, Microsoft and AMD jointly introduce their new XBOX Series X System-on-Chip, with emphasis on power-reduction techniques such as fine-grained power management and supply monitoring, thermal/acoustic constraints, and yield/performance/power tradeoffs using compute-unit redundancy. The 7nm chip improves CPU/GPU performance by 3×/2× over Microsoft's prior-generation gaming SoC.

8:35 AM

3.2 The A100 Datacenter GPU and Ampere Architecture

Jack Choquette, Nvidia, Santa Clara, CA

In Paper 3.2, Nvidia highlights their new A100 datacenter GPU and Ampere architecture, focusing on a next-generation Tensor core for efficient matrix multiplies. The 826mm^2 54B transistor A100 die includes a large number of new features including support of new data types, the streamlining of data movement reflecting recent advances in deep-learning algorithms, and advanced hardware and software support for multi-GPU systems including improved high-speed I/O.

8:40 AM

3.3 Kunlun: A 14nm High-Performance AI Processor for Diversified Workloads

Jian Ouyang, Baidu, Beijing, China

In Paper 3.3, Baidu introduces Kunlun, their first in-house design targeting artificial intelligence. The chip seeks to combine programmability in its XPU-cluster compute unit with high energy efficiency in deep learning via its XPU software-defined neural network compute unit. This hybrid architecture combined with a unified programming model allows Kunlun to be readily applied to a range of applications; diverse examples are shown in industrial defect detection and conventional search engines.

8:45 AM

Session 3 Authors – 30 Minute Live Q&A