2023 IEEE CICC Review

IEEE Custom Integrated Circuits Conference

포항공과대학교 전자전기공학과 박사과정 홍승우

Session 14 Heterogenous SoCs for Next-Gen Compute Applications

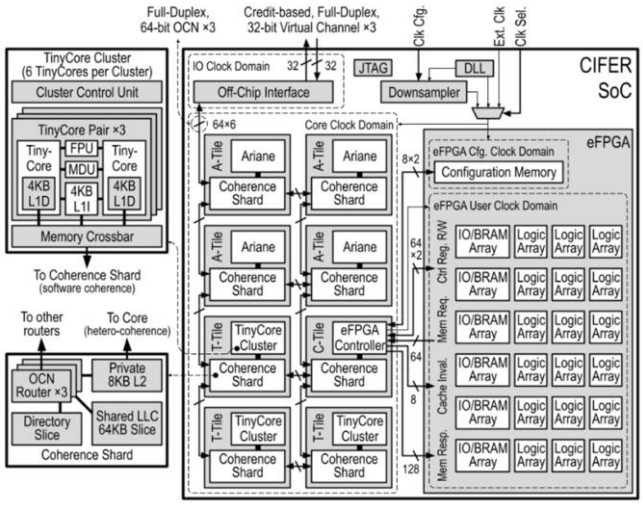
이번 2023 IEEE CICC의 Session 14은 Heterogeneous SoCs for Next-Gen Compute Applications 라는 주제로 총 4편의 논문이 발표되었다.

#14-1 논문에서는 ASIC이 아닌 FPGA 기반으로 클라우드 서버를 구성하는 방식에 대해 다룬다. 보통 FPGA를 cloud-scale 시스템에 사용하는 것은 power, performance, cost 측면에서 효율적이지 못하다고 여겨진다. 이 논문에서는 여러가지 application에 대한 ASIC, FPGA 구현 결과를 예로 들어 FPGA를 서버에 사용하는 것이 더 효율적일 수 있는 상황에 대해 설명한다. 특히, ASIC의 hard logic으로 구현이 불가능하여 core-based 구현을 하게 되는 경우 FPGA는 더 효율적일 수 있다고 주장한다.

#14-2 논문은 Princeton University, Columbia University, University of California에서 만든 Heterogeneous Manycore SoC (DECADES)이다. 108개의 heterogeneous tile이 12x9 2D mesh구조로 3-channel NoC로 연결되어 있으며, tile은 GeMM, Conv2D accelerator와 6-stage, in-order, single-issue 구조의 RISC-V CPU (Ariane), accelerator의 local memory나 communication buffer로 사용될 수 있는 scratchpad memory와 near-memory compute unit(Nibbler)을 지닌 intelligent storage (IS), 그리고 eFPGA로 구성되어 있다. 60개의 Arian tile은 초당 55 giga cache-coherent RV64 instruction을 issue할 수 있어 single chip 기준 최대 throughput을 달성하였으며, 1495개의 Nibbler lane과 eFPGA의 multiplier/adder 1172개를 사용하여 1.46 TOPS를 달성하였다.

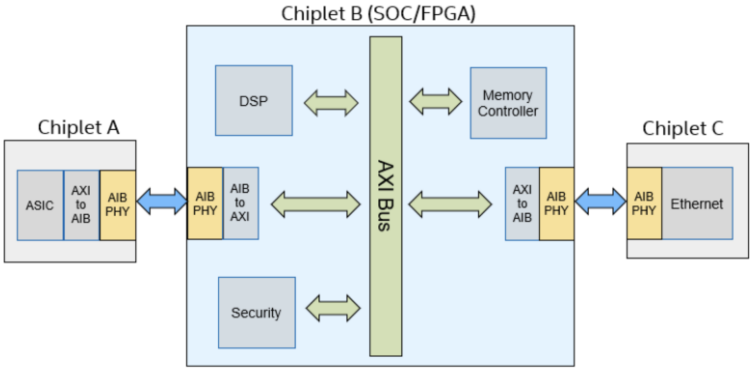
#14-3 은 14-2와 마찬가지로 Princeton University, Columbia University, University of California에서 발표한 Manycore-FPGA SoC (CIFER)로, DECADES와 달리 다양한 level의 parallelism을 구현하기 위한 hetero-granular architecture에 초점을 맞추고 있다. DECADES와 비슷하게 2x4 mesh 구조의 tile과 eFPGA로 구성되어 있으며, tile의 종류에는

Ariane core, TinyCore cluster, eFPGA controller가 있다. Ariane core는 Linux-capable, 64-bit RISC-V core로 hardware로 L1D, L1I와 L2의 cache coherence가 컨트롤되며 thread-level parallelism을 효율적으로 제공한다. 각각 6개의 32-bit RISC-V가 cluster로 구성된 TinyCore cluster의 경우 software coherence를 지원하며 Cluster를 통해 data-level parallelism을 향상시킨다. 마지막으로 eFPGA는 reconfigurable한 task를 담당하며, Configurable coherence를 통해 전체적인 SoC의 programmability를 향상시킨다.



[그림 1] CIFER SoC Architecture

#14-4 는 Chiplet을 위한 AXI4 Adapter에 관하여 Intel이 공개한 논문이다. Chiplet구조에서 PHY의 상위 layer에서의 control을 chip designer가 직접 control 하는 경우, PHY-to-application interface 개발에 많은 시간이 소모되기에, designer가 application 개발에 집중할 수 있도록 D2D communication에서 End-to-end로 AXI4를 사용하는 interface를 개발하였다. Chiplet PHY의 표준인 AIB와 AXI4 interface를 통해 heterogeneous integration을 가능하게 하였다.



[그림 2] AXI4 to AIB2.0 adapters in a multi-chiplet system

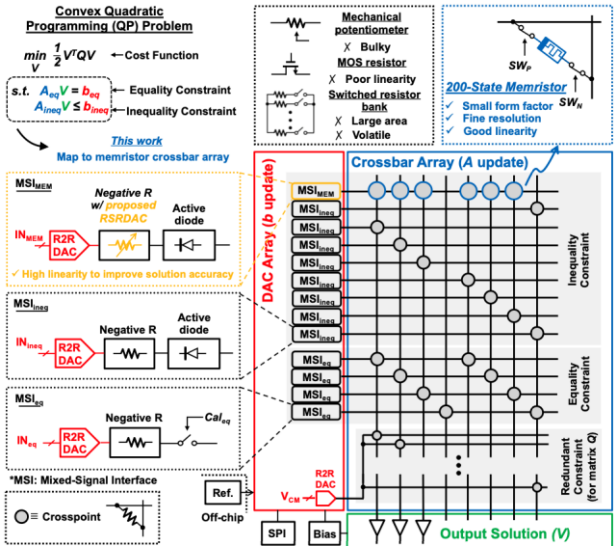
Session 21 Mixed-Signal Foundational IPs for Emerging Systems

이번 2023 IEEE CICC의 Session 21은 Mixed-Signal Foundational IPs for Emerging Systems 라는 주제로 총 4편의 논문이 발표되었다.

#21-1 은 Intel과 UC Berkley의 논문으로, high-throughput, low-power, energy-efficient 3DIC interconnect를 만드는 것에서의 challenge에 대해 다룬다. 해당 논문에서는 3D D2D interconnect의 theoretical bound를 분석하고, 표준 디자인에서의 영향에 대해 분석했다. 이에 앞서 interconnect 관점에서 device roadmap과 assembly scaling을 설명하고, vertical D2D interconnect 표준을 위한 design choice와 key assumption을 설명한다. 이를 검증하기 위해 routing analysis를 수행하여 on-die routing resource에 의한 최대 bandwidth, bump map 변화를 분석하고, I/O circuit에 대해 분석하고, clocking과 de-skewing strategy에 대한 tradeoff를 분석하였다. 본 논문은 Power delivery, thermal dissipation, testing for good die 등을 포함한 3D-IC와 interconnect design rule을 분석하여 추가적인 constraint를 확인했고, 디자인 옵션 등을 제공한 점에서 의미가 있었다.

#21-2 에서는 Adaptive body biasing (ABB)를 사용한 5D MIMO radar processing을 위한 low-power 25 processing element를 가진 multi-core DSP (RaDSP) architecture를 제안한다. 제안하는 DSP 아키텍처는 12개의 SAR ADC로 구성돼 있으며, FFT나 ML inference를 위한 RISC-V 기반 PE (rPE)와 distributed SRAM, raw data extraction을 위한 LVDS lane, 그리고 standard interface들로 구성되어있다. ABB와 low clock frequency를 통해 RaDSP는 SoTA의 90분의 1 수준인 52.6 mW를 달성하였다.

#21-3 에서는 conventional digital/analog quadratic programming (QP) solver의 문제점을 해결한 새로운 QP solver 구조를 제안한다. Weight-changing element로 programmable memristor로 구성된 crossbar를 이용하여 compact size, high linearity, non-volatality, Programmability에 이점을 갖도록 하여 정확도 및 성능을 향상시켰으며, 추가로 replica-switch resister DAC (RSRDAC)를 사용하여 tunable negative resistor를 linearize하였다. 제안된 QP solver 구조는 평균 1.23 us의 computing latency, relative error 1.1%를 나타내었으며, 기존의 analog QP solver 대비 output 성능을 유지하며 50배 이상의 latency 향상을 이끌어내었다.



[그림 3] Heterogeneous QP solver 구조

#21-4는 IBM의 논문으로, quantum computing system에서 room temperature에서 control과 qubit readout operation에 사용되는 전자기기들을 qubit이 실제 동작하는 cryogenic temperature에서 동작하는 cryogenic electronic device, cryogenic CMOS component에 대해 다룬다. 특히 qubit control을 위한 waveform을 저장하기 위한 SRAM와 qubit readout 과정에서 qubit state 판단을 위해 amplification, post processing을 수행하는 LNA에 대해 집중적으로 다룬다. room temperature와 동일하게 reliable한 Cryo-CMOS SRAM 디자인을 통해, power consumption을 33% 줄일 수 있었으며, Cryo-CMOS LNA는 peak gain 13.3dB, 3dB bandwidth 2.7 GHz, noise frequency 0.34dB to 0.44 dB를 달성하여 목표치를 도달하였으며, SoTA 대비 15배 이상의 전력 이득을 보였다.

저자정보

명예기자 홍승우

- 소　　속 : 포항공과대학교 전기및전자공학과 박사과정
- 연구분야 : DSP Architecture, ASIC/FPGA design
- 이 메 일 : seungwoohong@postech.ac.kr
- 홈페이지 : https://sites.google.com/view/epiclab