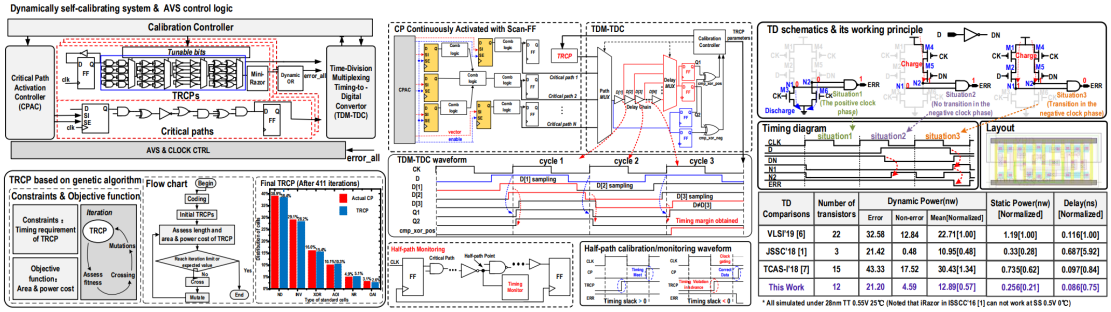
A-SSCC 2022
IEEE Asian Solid-State Circuits Conference

서울대학교 전기정보공학부 석박통합과정 박현준

Session 9 : Energy-Efficient Digital Circuit Techniques

이번 A-SSCC 2022의 Session 9은 Energy-Efficient Digital Circuit Techniques라는 주제로 총 5편의 논문이 발표되었다.

#9-1은 National ASIC System Engineering Research Center와 Southeast 대학에서 발표한 논문으로 near-threshold voltage에서 PVT variations에도 안정적인 동작을 지원하는 Adaptive Voltage Scaling(AVS)기법을 제시한다. 직접 모니터링 기반의 AVS에서는 Timing Margin Sensing에 모든 Critical Path(CP)가 반드시 고려되지는 않으므로 오판의 가능성이 존재한다. 반면 worst case Critical Path가 제대로 Sensing 되었다면, Violation 없이 최적의 Voltage를 정할 수 있어 AVS gain이 높은 장점이 있다. Worst case Tunable Replica Cricital Paths를 통한 간접적인 모니터링방식의 AVS는 worst case CP보다 긴 Delay를 갖는 TRCP를 사용해 timing margin을 측정하면서 필요한 Voltage를 적응적으로 제공하는 방식이 있다. 이 경우, Worst case에 TRCP가 맞춰지면서, Critical Path의 오판문제는 없지만, 불필요한 margin을 TRCP가 가질 수 있기 때문에 AVS gain이 낮은 문제가 있다. 해당 논문에서는 간접 모니터링 방식을 기반으로 Genetic Algorithm을 사용하여 실제 Critical Path를 최대한 모방하는 TRCP를 결정함으로써, Overhead를 줄여 두 방식의 장점을 취하는 방법론을 제시한다. 우선 Post-silicon calibration동안 순차적으로 critical path를 활성화시키면서 Path-Mux를 통해 TDC에 입력을 선택하도록 한다. TDC는 입력의 timing margin을 순차적으로 측정하여, TRCP를 튜닝한다. 이때, TRCP는 고려중인 모든 PVT 조건에서 여러 개의 Critical Path를 보정해야 하므로 최적화하기 어려운 문제가 있다. 해당 논문은 그림1의 Genetic Algorithms을 사용하여 실제 path와 최대한 비슷한 std-cell type과 분포를 갖는 TRCP구조를 결정한다. 또한, 그림1와 같이 TRCP의 transition이 CLK의 반주기 이후(Negative Phase)에서 발생하는 경우를 감지하는 TD를 사용하여, Violation을 감지하는 Half-path calibration/monitoring을 수행한다. 해당 아키텍쳐를 28nm 공정에서 제작된 NN 가속기에 적용하여 31MHz에서 VDD를 0.45V로 낮추어 58%의 전력 이득(232%의 frequency 이득)을 0.65%의 AVS area cost를 들여 얻었다.

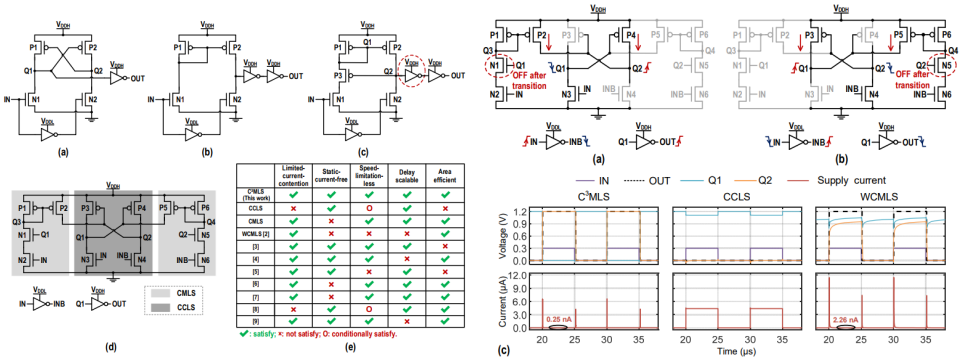


TD Comparisons	Number of transistors	Dynamic Power(nw)			Static Power(nw) [Normalized]	Delay(ns) [Normalized]
		Error	Non-error	Mean[Normalized]		
VLSI'19 [6]	22	32.58	12.84	22.71[1.00]	1.19[1.00]	0.116[1.00]
JSSC'18 [1]	3	21.42	0.48	10.95[0.48]	0.33[0.28]	0.687[5.92]
TCAS-I'18 [7]	15	43.33	17.52	30.43[1.34]	0.735[0.62]	0.097[0.84]
This Work	12	21.20	4.59	12.89[0.57]	0.256[0.21]	0.086[0.75]

* All simulated under 28nm TT 0.55V 25℃ (Noted that iRazor in ISSCC'16 [1] can not work at SS 0.5V 0℃)

[그림 1] (좌) 제안 시스템의 전체 구조 및 TRCP 결정을 위한 Genetic 알고리즘 순서도 (중) 시스템 동작에대한 상세 구조 (우) Transition detector schematic 및 동작 원리

#9-2은 Peking University Shenzhen Graduate School에서 발표한 논문으로 multiple supply voltage system에서 ultra wide range level shifter의 구조로 C3MLS를 제안하고 있다. cross coupled level shifter(CCLS) 구조는 제한된 conversion 범위와 transition에 큰 에너지를 사용하는 단점을 갖는다. Current mirror level shifter(CMLS)의 경우 CCLS의 이러한 단점을 일부 해결하지만, 입력 데이터에 따라 static current를 소모할 수 있어 큰 에너지를 소모한다. 해당 논문에서 제안하는 C3MLS는 그림2과 같이 주 변환부가 CCLS 스테이지와 2개의 보조 CMLS 스테이지로 구성된다. CMLS 스테이지의 보조 mirror 전류의 스위칭 동작으로 CCLS의 current contention 문제 및 static current 문제를 해결한다. 예를 들어, low-to-high transition의 경우, Q1 노드는 처음엔 VDDH의 전압을 갖는다. 이어서 IN이 high로 변화하면서 N1과 N2가 켜지고 Q3를 pull-down 하면서 P2를 통해 전류를 생성한다. 이에 따라 Q2의 전압이 올라가면서 P3의 driving 세기를 약화시킨다. Mirror 전류로 Q2의 전압을 올려 P3의 Driving 세기를 약화시키는 과정을 보조하므로, N3와 P3간 current contention을 상당부분 해결할 수 있다. Transition 이후 N1이 완전히 꺼지면서 static current path를 꺼버린다. 이를 통해 CCLS stage에서의 static current를 제거한다. 테스트 칩은 55nm low power CMOS 공정으로 제작되었고, 10kHz ~ 10MHz의 범위에서 300mV 미만(0.17 V)의 subthreshold Voltage를 1.2 V의 core Voltage로 변환할 수 있다. 또한 기존연구 대비 최소 31%의 EDP 성능 향상을 보인다.

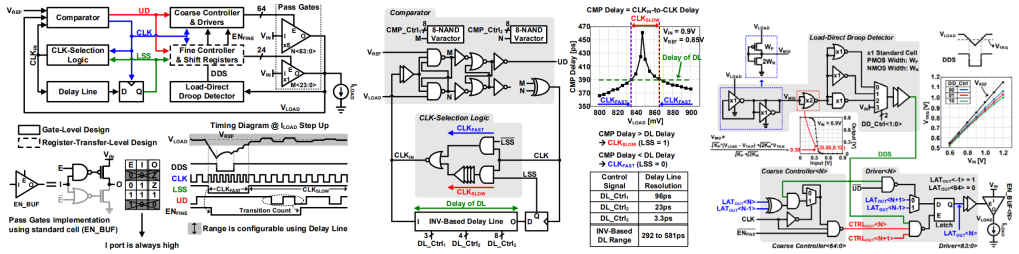


	Limited-current-contention	Static-current-free	Speed-limitation-loss	Delay scalable	Area efficient
C3MLS (This work)	✓	✓	✓	✓	✓
CCLS	×	✓	O	✓	×
CMLS	✓	×	✓	✓	✓
WCMLS [2]	✓	×	×	×	✓
[3]	✓	✓	✓	✓	×
[4]	✓	✓	✓	×	✓
[5]	✓	✓	×	✓	×
[6]	✓	×	✓	✓	✓
[7]	✓	×	✓	✓	✓
[8]	×	×	O	✓	✓
[9]	✓	✓	✓	×	✓

✓: satisfy; ×: not satisfy; O: conditionally satisfy.

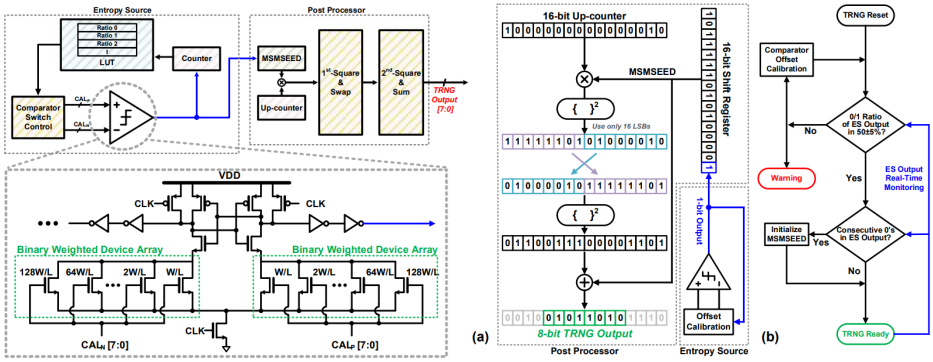
[그림 2] (좌) Level Shifter Schematic (우) (a) C3MLS의 low-to-high transition 동작 (b) C3MLS의 high-to-low transition 동작 (c) C3MLS, CCLS, WCMLS의 시뮬레이션 파형

#9-3은 Columbia University, 서울대학교, 숭실대학교, 충남대학교에서 발표한 논문으로 fully-synthesizable Digital Low-Dropout Regulator(DLDO)를 제안한다. Voltage-droop이 발생했을 때, 출력 전압의 빠른 복구를 위해서 Coarse 컨트롤러가 업데이트된 출력을 기반으로 드라이버를 제어한다. 그림 3에서 비교기는 로드 전압과 레퍼런스 전압의 차이를 비교하고, CLK-Selection Logic은 업데이트 속도를 선택하는 역할을 한다. LSS 신호는 CLK신호와 DL 신호의 선후관계에 따라, 비교기 delay가 DL보다 길면 1을 반대는 0을 출력한다. 비교기의 metastability로 인해, 로드 전압이 레퍼런스 전압과 가까울수록 delay가 증가하는 현상을 이용하여 에러가 작을수록 비교기 지연이 길어져 LSS가 1을 출력하도록 해서 느린 CLK을 선택하여 Fine tuning하고, 에러가 클수록 LSS가 0을 출력하도록 하여 Coarse tuning하도록 한다. LD-DD에서는 Inverter Cell과 NOR gate를 이용하여 steady state에서 low voltage이되, ac gain이 큰 상태에의 전압을 만들고, Voltage Drop을 Sensing하여 DDS를 출력한다. 테스트 칩은 40nm CMOS 공정에서 제작되었으며, voltage droop 98mV에 대해 response time 0.8ns, settling time 5ns로 측정되었다. 공급 전압의 범위가 0.6 ~ 1.2 V이고 dropout voltage가 50mV일 때, quiescent current는 59.3~ 724 uA로 측정되었다. Peak current efficiency는 Vin이 0.6 V 일 때 99.6 %로 측정되었다. 또한 기존 연구 대비 가장 좋은 current density(13.01 A/mm2)를 보였다.



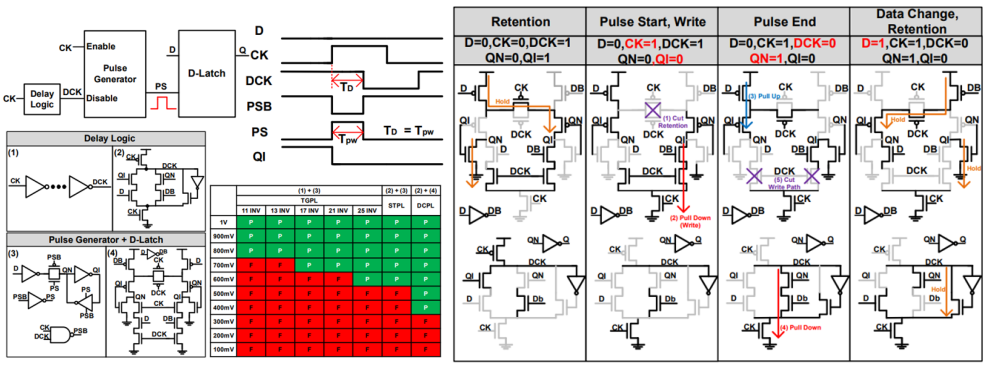
[그림 3] (좌) 제안하는 DLDO의 구조 및 timing diagram (중) 비교기, CLK-Selection Logic, Delay Line 구조 (우) Load-Direct Droop Detector 구조 및 Coarse Controller & Driver 구조

#9-4은 건국대학교에서 발표한 논문으로 All-Digital True Random-Number Generator를 제안한다. 일반적으로 True Random Number Generator(TRNG)는 무작위성을 내포한 Entropy Source(ES)와 Random Number의 퀄리티를 높이기 위한 Post processor로 구성된다. 해당 논문에서는 Power Supply Attack(PSA)를 피하기 위해 비교기 기반의 ES와 Middle Square Method post-processing을 사용하였다. 비교기 기반의 ES의 경우 Calibration Logic이 필수적이다. 본 논문에서는 전통적인 Cap-bank 대신 디지털 비교기를 사용함으로써 area를 줄였다. Calibration은 ES에서 출력되는 0 혹은 1의 값을 카운터가 연속적으로 모니터링하면서 LUT의 값을 통해 offset의 크기를 추정하고 비교기의 실효폭을 조정하는 방식이다. Post processing에 사용된 MSM 알고리즘은 암호학에서 의사난수를 생성하는데 사용되는 것으로, 몇 가지 기본 연산으로 구성되어 전력효율이 우수하다. 본 논문에서는 전력 소비를 줄이기 위해 제곱 연산의 수를 절반으로 줄인 그림 4에 나타난 수정된 알고리즘을 사용한다. ES 출력은 16비트 MSM 시드값이 되고 독립적으로 구성된 업카운터의 값에 곱해진 후 제곱된다. 이후 8개의 MSB와 8개의 LSB가 스왑 되고 이어서 다시 제곱이 이루어진다. 마지막으로 제곱 결과를 MSMSEED와 합산한 후 16비트 결과의 중간에서 8비트를 추출하여 TRNG 출력을 획득한다. 양질의 ES 출력을 얻어 필요한 제곱 연산의 수를 줄여 기존 연구 대비 총 전력 소비량을 40% 감소하고, 각 제곱 작업을 파이프라이닝하여 속도를 극대화하였다.



[그림 4] (좌) 제안 시스템의 전체 구조 및 TRCP 결정을 위한 Genetic 알고리즘 순서도 (우) 시스템 동작에 대한 상세 구조 및 Transition detector schematic 및 동작 원리

#9-5은 성균관대학교에서 발표한 논문으로 Differential Contention-Free Pulsed Latch(DCPL)를 제시했다. Near threshold voltage에서 순차회로는 Process variation에대한 sensitivity가 크게 증가하여 동작 속도가 감소하고 제대로 기능하지 못하게 된다. 본 논문은 문제 해결을 위해 Pulsed Latch 구조를 사용하였다. Pulsed latch는 펄스폭을 결정하기 위한 Delay Logic, 실제 펄스를 생성하는 Pulse Generator, D Latch로 구성된다. 그림 5의 (1)의 경우 Conventional한 Delay 라인인데, low-voltage 동작에서는 variation의 영향을 심하게 받아 적합하지 못하다. 또한 (3)의 경우 Conventional한 Pulse generator인데, (1)번과의 조합의 경우 많은 수의 inverter가 필요하여 power 및 area측면에서 불리하다. (2)번과의 조합의 경우 단순한 dynamic XOR 게이트를 사용하여 적응적으로 delay를 조절하지만, single-ended write 구조로 QN만 input D에의해서 쓰여진다. 이로 인해 최악의 경우 QN이 변함에 따라 Delay Logic이 DCK transition이 일어나 pulse가 종료될 수 있다. 이는 QI가 안정적으로 쓰여지기 전에 (3)의 TG를 꺼버릴 수 있으므로 불안정한 쓰기 동작을 유발한다. 본 논문에서는 (4)의 차동 래치 구조를 사용하여 DCK를 Pull down하여 펄스를 종료할 수 있는 QN/QI의 rising transition을 QN/QI의 fall transition보다 늦게 발생하도록 그림5와 같이 동작하여 안정적인 쓰기를 보장한다. 테스트 칩은 28nm 공정으로 제작되었다. 많은 Delay Line이 필요한 TGPL대비 적은 전력을 소모하고, STPL 및 TGFF와는 유사한 전력을 소모한다. 하지만 TGPL 및 STPL대비 NTV 영역에서 훨씬 높은 신뢰도를 보여준다.



[그림 5] (좌) Pulsed Latch 구조 및 몬테카를로(10k) 시뮬레이션 결과 (우) DCPL 동작

저자정보

명예기자 박현준

- 소　　속 : 서울대학교 전기정보공학부 석박통합과정
- 연구분야 : High Speed I/O, Information Theory
- 이 메 일 : spp098@snu.ac.kr
- 홈페이지 : https://sites.google.com/view/wschoi?pli=1

A-SSCC 2022
IEEE Asian Solid-State Circuits Conference

포항공과대학교 전자전기공학과 박사과정 변영훈

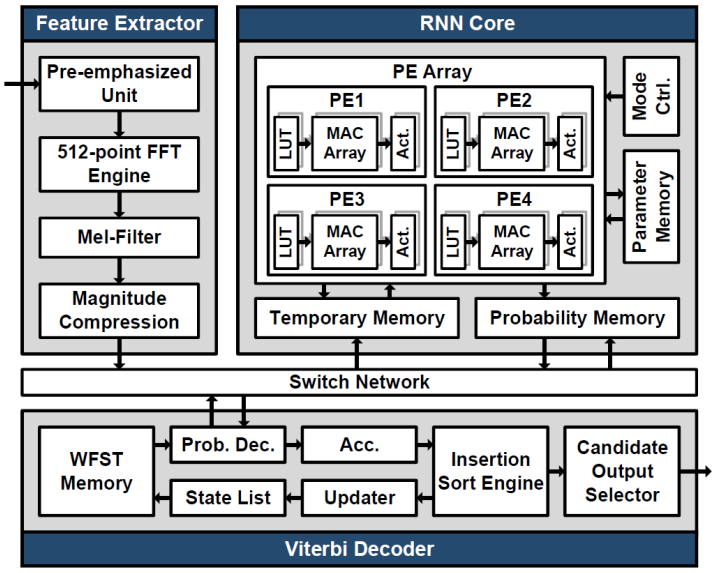
Session 15 Energy-Efficient Machine Learning Processors and High-Speed Interface

이번 A-SSCC 2022의 Session 15에서는 총 네 개의 논문이 발표되었다. 앞의 두 논문(15-1, 2)에서는 IIoT device에서의 심전도 기반 유저 인증을 위한 대한 아키텍처와 BLiGRU기반 speech-to-text accelerator를 다루고 있고, 뒤의 두 논문(15-3, 4)에서는 high-bandwidth data transfer를 위한 digital-to-analog encoding의 일종인 PAM transceiver와 3D-stack memory의 synchronous transceiver topology에서 encoding으로 인해 발생되는 limited data-rate를 해결하는 회로를 제안한다.

#15-1 본 논문에서는 산업용 사물 인터넷 (IIoT)에서 심전도 (ECG) 생체 인증을 활용할 수 있도록 하는 QR-decomposition 기반 extreme learning machine engine을 제안한다. 일반적으로 IIoT 생체 인증 시나리오에서는 모델의 retraining 없이 기존의 유저 정보를 잊어버리지 않으면서도 새로운 유저의 정보를 학습해야 한다. 일반적으로 새로운 유저의 인증을 위한 back-propagation 과정에서 기존 유저들의 인증 성능이 급격하게 떨어지는 현상(catastrophic forgetting)이 발생하는데, 이를 방지하기 위해 class incremental learning (CIL)이라는 방식이 사용된다. 이 때 classifier를 업데이트 하기 위해 recursive least square (RLS) 연산을 필요로 한다. 제안하는 칩은 QRD-ELM engine과 unified CORDIC (u-CORDIC) engine으로 구성되는데, 먼저 QRD-ELM engine의 경우 matrix inversion 없이 QRD를 하기 위해 2D systolic array를 대각선으로 folding하는 one dimensional diagonally-mapped linear array (1D-DMLA)을 포함한다. 다음으로 u-CORDIC engine에서는 circular, scaling, 그리고 linear operation들을 지원하는데, 기존 방식과 다르게 processing element (PE)를 통합하여 연산 효율을 올렸다. 최종적으로 model과 algorithm의 co-design을 통해 기존 AISC기반 inference 전용 ELM[1]과 비교해 4.04배 높은 inferencing throughput과 6.4배 높은 energy efficiency를 달성하였다.

#15-2 본 논문은 energy-efficient하면서도 높은 정확도를 가지는 speech-to-text accelerator를 제안한다. 제안된 칩은 28nm CMOS 공정으로 제작되었는데, 9.58mm^{2}의 코어 영역에 9.42M개의 logic gate를 사용하고 0.6V 전원에서 최소 1.25MHz, 최대 100MHz로 동작한다. TIMIT dataset에서 최대 15.2% phone error rate (PER)로 normalized energy 기준 기존 연구 대비 6.5~177배 적은 에너지만을 소모하는데, frame당으로 보면 약 1.3mW의 에너지를 소모한다. 이를 위해 본 논문에서는 bidirectional light gated recurrent Unit (BLiGRU), Fast Fourier Transform (FFT)을 사용한 filter bank (FBANK), Viterbi decoding을 사용한 beam search와 같은 기법들을 사용한다. BLiGRU은

Recurrent neural network (RNN)을 기반으로 한 모델인데, forward와 backward path에서 사용하는 weight를 공유하는 방식을 통해 전체 parameter의 개수를 절반으로 줄이고 scaling factor pruning (SFP), multi-bit clustering (MBC), linear quantization (LQ)와 같은 기법을 사용해 압축되었다. 해당 논문은 기존 JSSC나 ISSCC에서 제안된 speech-to-text accelerator와 비교할 때 낮은 normalized energy per frame이나 PER을 보여준다는 점에서 강점을 가지고 있지만 최대 100MHz에서 밖에 동작하지 못한다는 점이나 파워 측정이 1.25MHz에서 되었다는 점을 감안할 때 28nm의 최신 공정에도 불구하고 기존 65nm 공정을 사용한 칩들과 비교하여 갖는 강점들이 대부분 optimized model로부터 나온다는 점이 아쉬운 부분으로 남는다.



[그림 2] 제안하는 speech-to-text accelerator의 구조도.

#15-3 본 논문에서는 6Gbps data rate를 갖는 PAM-3 transceiver에 대한 offset compensation 기법을 제안한다. Pulse Amplitude modulation (PAM)이란 통신에서 대역폭을 확장시키기 위한 방법 중 하나로, digital signal을 amplitude가 다른 analog signal로 encoding 하는 방식을 말한다. Transmitter는 PAM-3 인코더 출력을 off-chip channel을 통해 전송하고 receiver는 signal을 differential 신호로 변환한 뒤 single-to-differential amplifier (S2D)와 decision feedback equalizer (DFE)를 사용하여 equalize한다. 논문에서 제안하는 핵심 아이디어는 offset compensation block인데, 이는 offset sensing 회로와 reference generator로 구성된다. Offset sensing 회로는 fault pattern을 감지하여 offset의 amount를 찾아내고 이것이 사라질 때까지 S2D 및 DFE에 대한 reference generator를 control해 reference voltage를 조절하는 역할을 한다. 전체 transceiver의 area는 0.1mm^{2}이고 1.15V 및 1.18V voltage에서 각각 6.45Gbps 와 6Gbps의 data rate로 동작한다. 제안하는 기법의 성능은 Rx eye diagram과 측정된 bit error rate (BER)를 통해 보여주는데, 3개의 chip에서 발생하는 worst-case eye-opening을 38%까지 개선하였고 낮은 supply level에서 voltage variation에 대해 강인함을 보여준다는 점에서 그 의의를 찾을 수 있다.

#15-4 본 논문에서는 3D-stack memory의 구현을 위한 synchronous transceiver topology에서 사용하던 기존 Manchester encoding에서 발생하는 limited data-rate를 해결하기 위해 encoding 없이 inductive bias를 이용해 데이터를 전송할 수 있는 receiver 회로를 제안한다. 기존 방식에서는 encoding으로 인해 data-rate가 8.5Gbps로 제한되었지만, 제안하는 회로는 수신기 코일에서 유도된 펄스를 감지하기 위해 대기 시간이 짧은 SR latch와 함께 clock hysteresis comparator를 사용하는 방식을 통해 12.8Gbps의 data-rate를 달성한다. 이 때 no-signal을 감지하기 위해 clock comparator는 의도적으로 imbalance하게 구성된다. 이전에 입력된 bit에 따라 clock hysteresis comparator는 no-signal이나 positive, 또는 negative pulse를 감지할 수 있다. DDR에서와 같이 F/2의 clock frequency에서 작동하기 위해 2개의 clock hysteresis comparator가 병렬로 사용되는데, 각 comparator의 이전 출력은 no-signal을 감지하기 위해 서로 다른 comparator로 feedback되는 구조를 이루고 있다. 제안하는 encoding-less wireless communication scheme은 bit당 에너지 소모가 0.5pJ로, 기존 Manchester encoding에 비해 36% 감소하였다. 이는 data-rate나 energy efficiency, IO area efficiency와 같은 지표들을 고려할 때 기존 wired and wireless integration 방식에 비해 월등히 좋은 성능을 보여주지만, 이를 구현하기 위해선 chip fabrication 이후 test를 통해 receiver circuit을 calibrate해야 한다는 한계점이 존재한다.

참고문헌

[1] Y. -C. Chuang et al., "An arbitrarily reconfigurable extreme learning machine inference engine for robust ECG anomaly detection," in IEEE OJCAS, vol. 2, pp. 196-209, Jan 2021.

저자정보

명예기자 변영훈

- 소　　속 : 포항공과대학교 전자전기공학과 박사과정
- 연구분야 : Deep Learning Model Compression
- 이 메 일 : byh1321@postech.ac.kr
- 홈페이지 : sites.google.com/view/epiclab/member/yhbyun