A 7-nm FinFET 1.2-TB/s/mm^2 3D-Stacked SRAM with an Inductive Coupling Interface Using Over-SRAM Coils and Manchester-Encoded Synchronous Transceivers

Kota Shiba[1], Mitsuji Okada[2], Atsutake Kosuge[2], Mototsugu Hamada[2], and Tadahiro Kuroda[2]

[1]The University of Tokyo, [2]Research Association for Advanced Systems (RaaS)

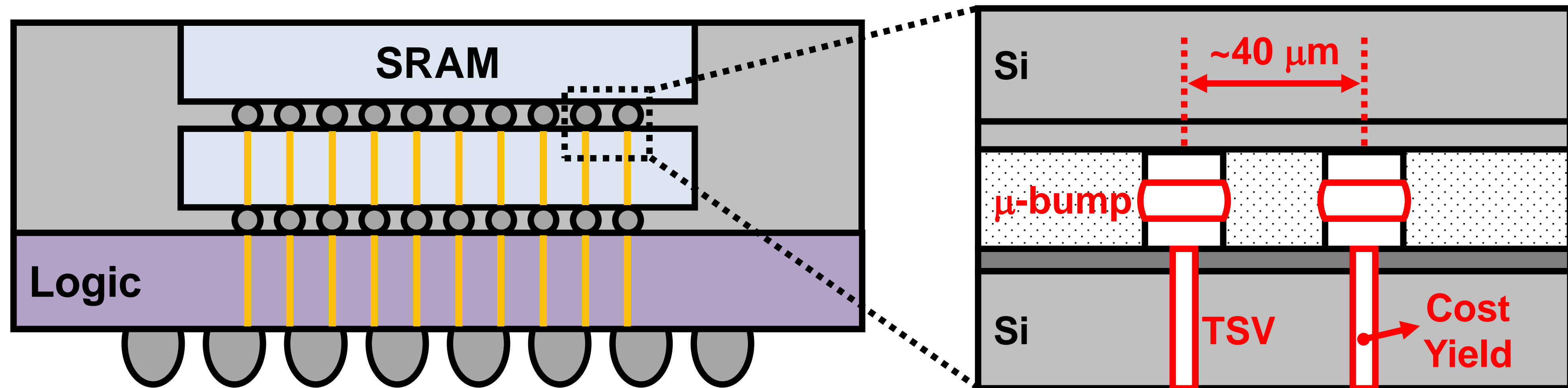
2022 Hot Chips 34 Symposium
Aug. 21 – 23, Virtual Conference

Abstract

A 0.7-pJ/bit, 8.5-Gbps/link inductive coupling inter-chip wireless communication interface for a 3D-stacked SRAM has been developed in a 7-nm FinFET process. A new physical placement method that allows coils to be placed over off-the-shelf SRAM macros with small magnetic field attenuation, together with the use of synchronous communication using Manchester encoding and a clocked comparator to enable the detection of small-swing signals, achieve a 26% reduction in SRAM die area compared to TSV-based stacking. Inter-chip communication at 0.7-pJ/bit, 8.5-Gbps/link was confirmed using test chips. A 4-hi 3D-stacked SRAM module using the proposed interface is estimated to achieve a 1.2-TB/s/mm^2 area efficiency, representing a two-orders-of-magnitude improvement over state-of-the-art 3D-stacked SRAM.

Introduction

- Mobile AI devices need high-bandwidth, low-latency memory with small form factor
- 3D-stacked SRAM (3D-SRAM) can meet these demands
- But current 3D-SRAM using TSV and μ-bump has issues with cost, yield and area efficiency [1][2]



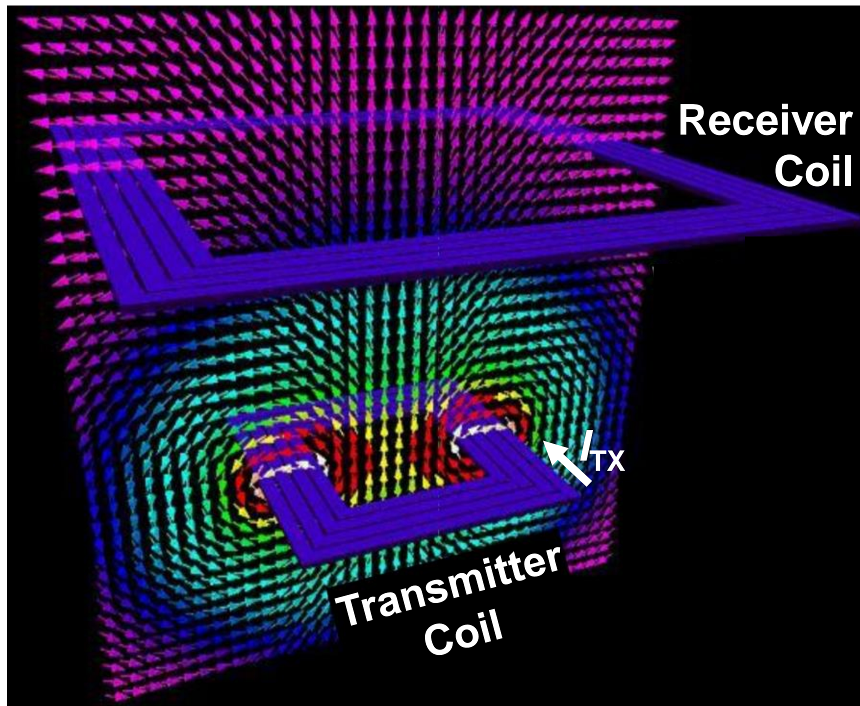
3D-stacked SRAM with TSV and μ-bump [1][2]

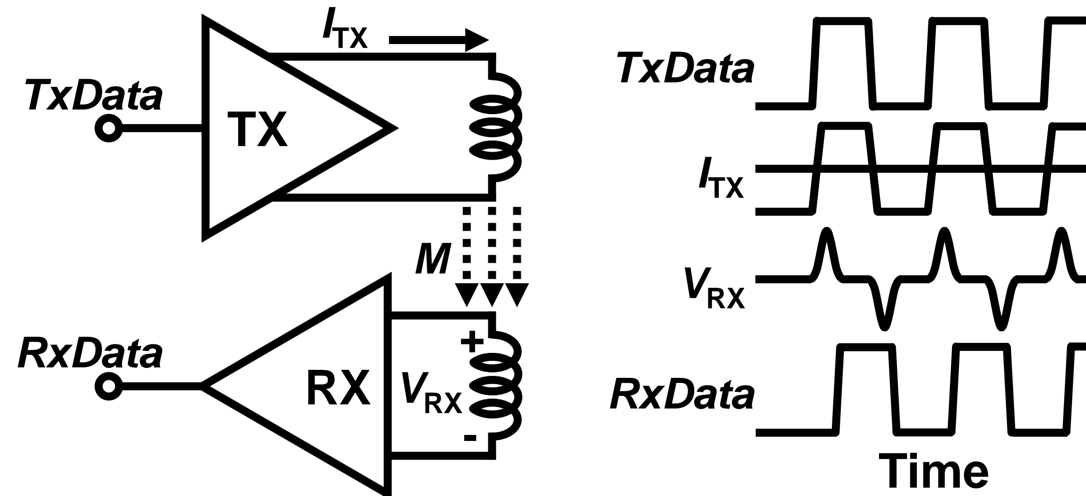
[1] K. Cho, et al., Hot Chips, 2020 [2] S.-K. Seo, et al., ECTC, 2021

Inductive Coupling Technology (TCI)

- To eliminate TSV and μ-bump, ThruChip Interface (TCI) is proposed, which is a wireless version of TSV
- TCI is compatible with standard CMOS process, leading to low cost and high yield

	TSV + μ-bump [1]	TCI [3]
Process	Additional Process	Standard CMOS process
Cost	High	Low
Yield	Low	High





[3] D. Ditzel, et al., Hot Chips, 2014

Conventional 3D-SRAM Using TCI

- A 40-nm 96-MB 3D-stacked SRAM using inductive coupling was proposed
- But limited bandwidth due to large coils is an issue

96-MB 3D-SRAM [5]
14.3 mm
8.5 mm
12MB/die
CH1 CH2 CH7 CH18 CH23 CH24
Inductive Coupling (TCI)
24 Independent Channels
AI Accelerator [4]

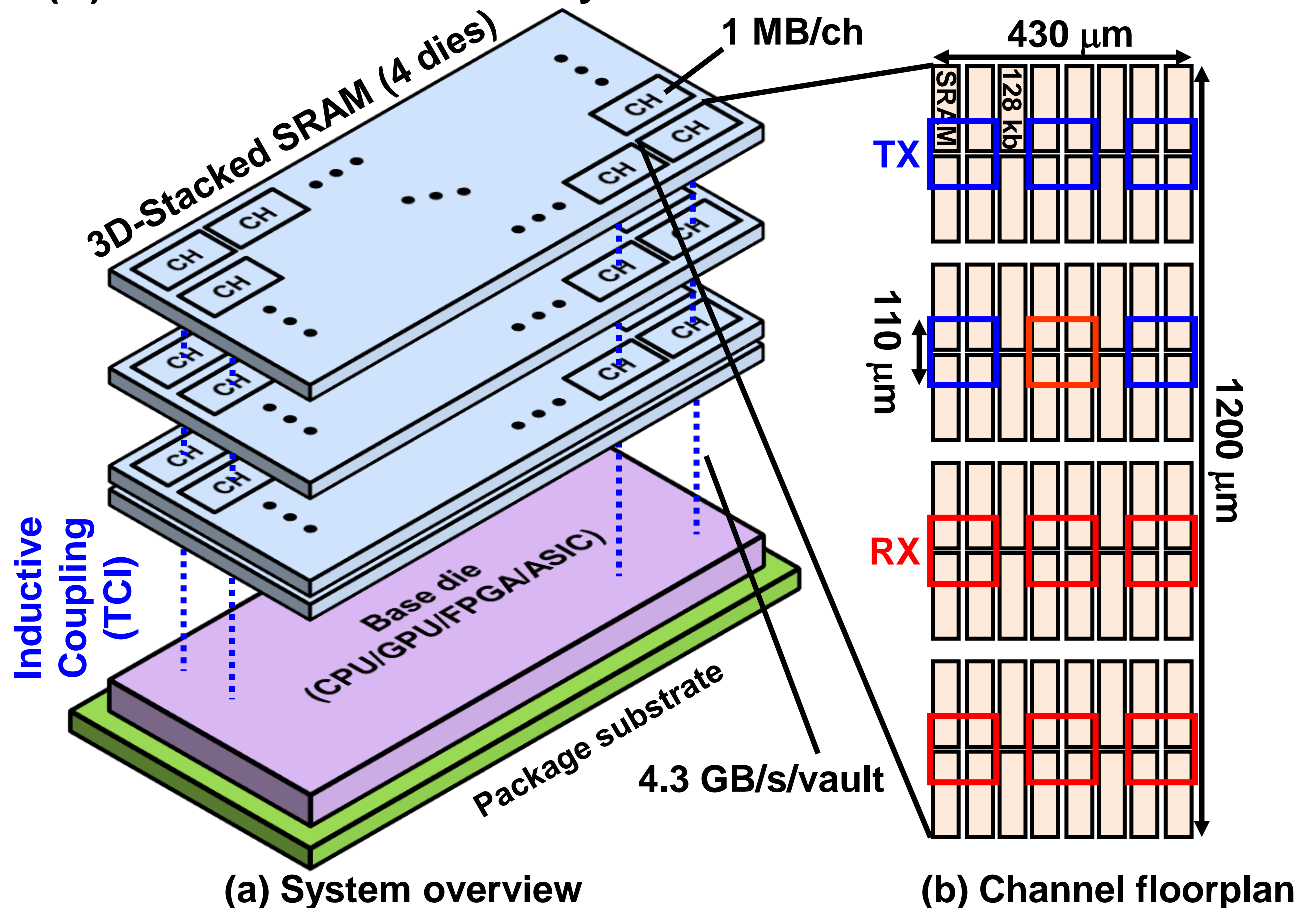
2250 μm
1945 μm
200 μm
SRAM Macro
Total 512-KB SRAM/CH
RX TX
Large coil area

2250 μm
1945 μm
200 μm
TX1 CLK TX2 TX3 TX4 TX5 CS
RX1 RX2 DQS RX3 RX4
Logic

[4] K. Ueyoshi, et al., JSSC, 2019
[5] K. Shiba, et al., TCAS-I, 2020

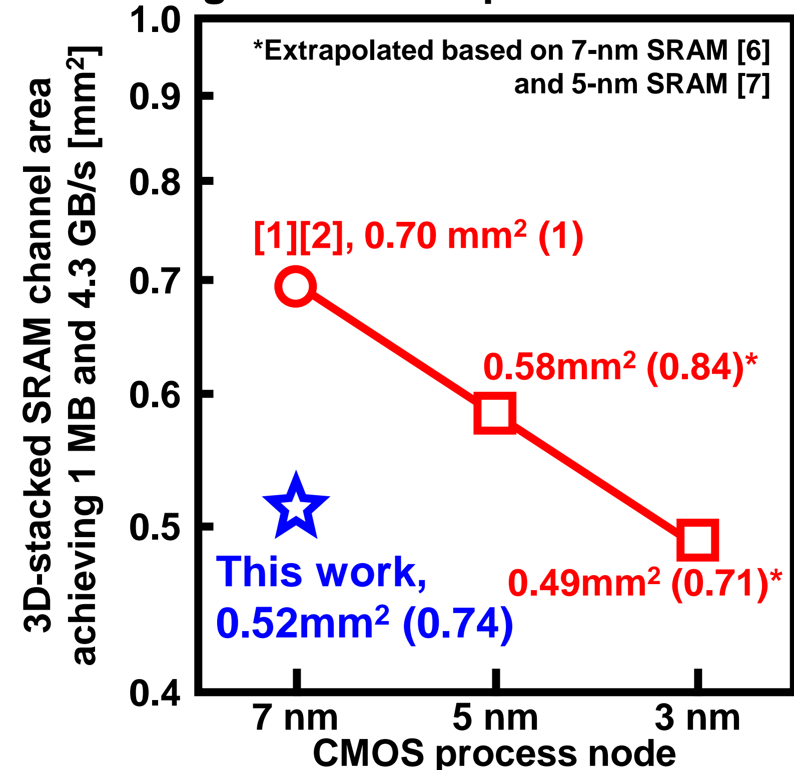
Proposed 3D-SRAM Using TCI

- **This work proposes 3D-stacked SRAM using inductive coupling with minimized area overhead, reducing SRAM die area by 26% vs TSV**
  - (A) Over-SRAM coils: enable high area efficiency while limiting magnetic field attenuation to 30%
  - (B) Manchester-encoded synchronous transceiver: detects small received signal with low power



(a) System overview

(b) Channel floorplan

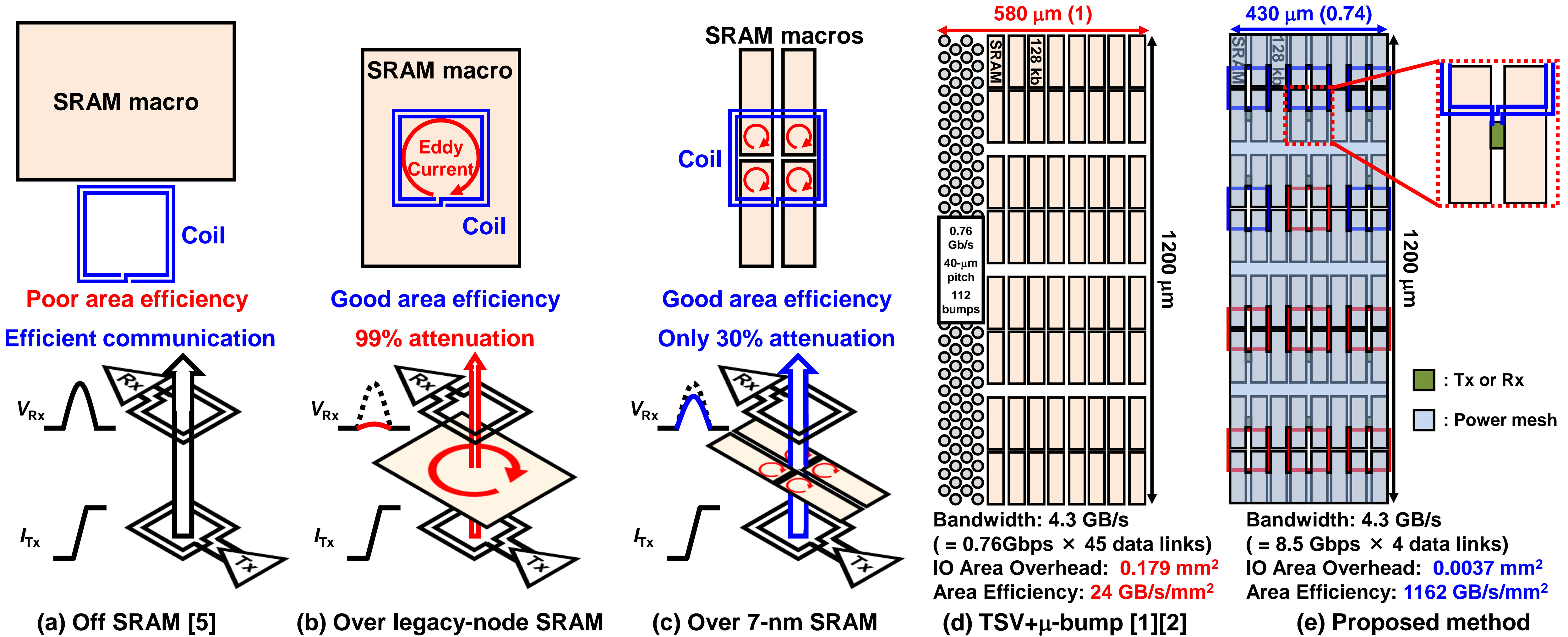


(c) Scaling trend of SOTA SRAM

[6] J. Chang, et al., ISSCC, 2017
[7] J. Chang, et al.,ISSCC 2020.

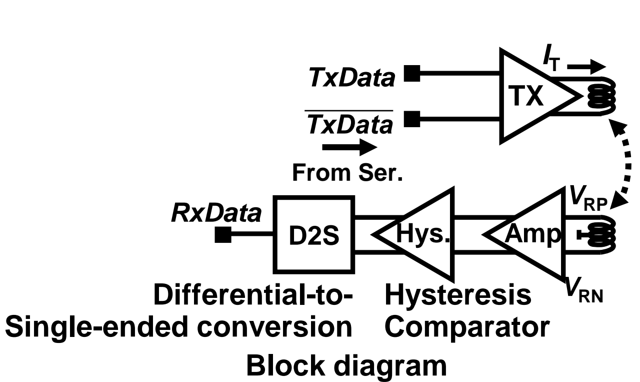
(A) Over-SRAM Coil

- **Proposed physical layout method of coils over off-the-shelf SRAM macros suppresses magnetic field attenuation due to eddy currents on SRAM macros**

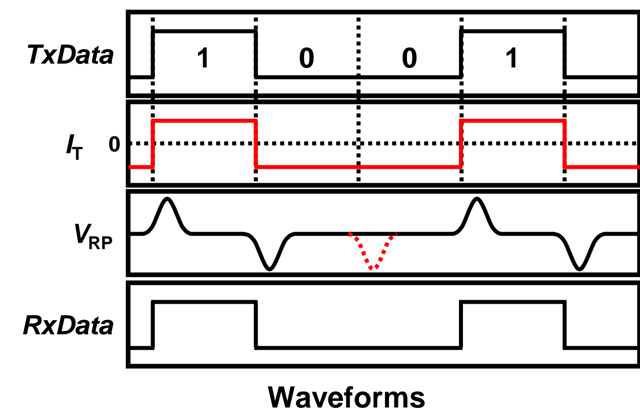


(B) Manchester-encoded Synchronous TRx

- (a) Clocked comparator and (b) Manchester encoding achieve detection of small pulse signal with low transmission power
  - (a) detects low-swing pulse by utilizing clock-triggered positive feedback, leading to low transmission power
  - (b) generates pulse signal in every cycle for clock-triggered data reception

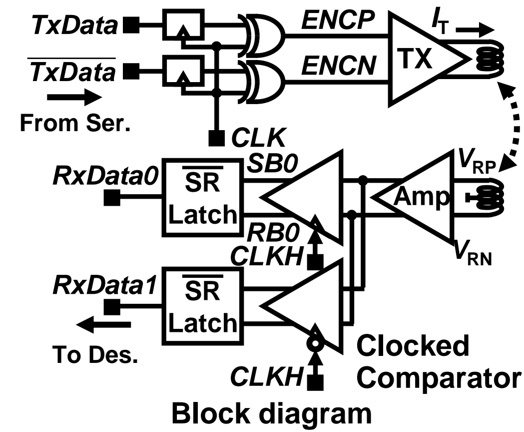


Block diagram

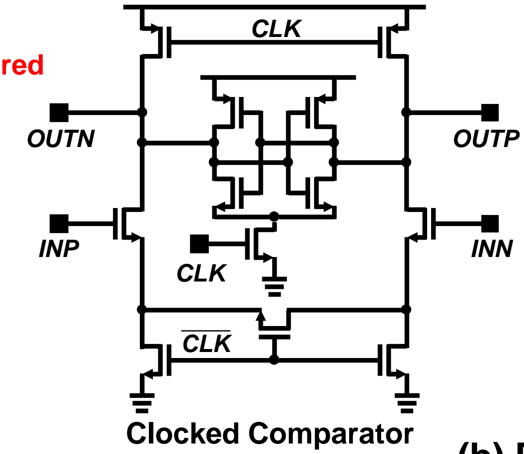


Waveforms

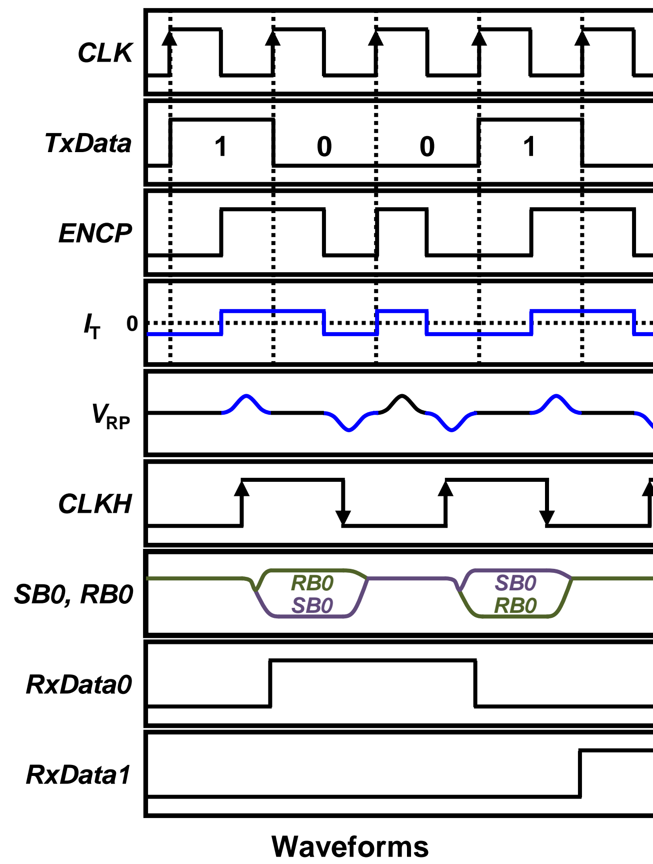
(a) Conventional TCI



Block diagram



Clocked Comparator

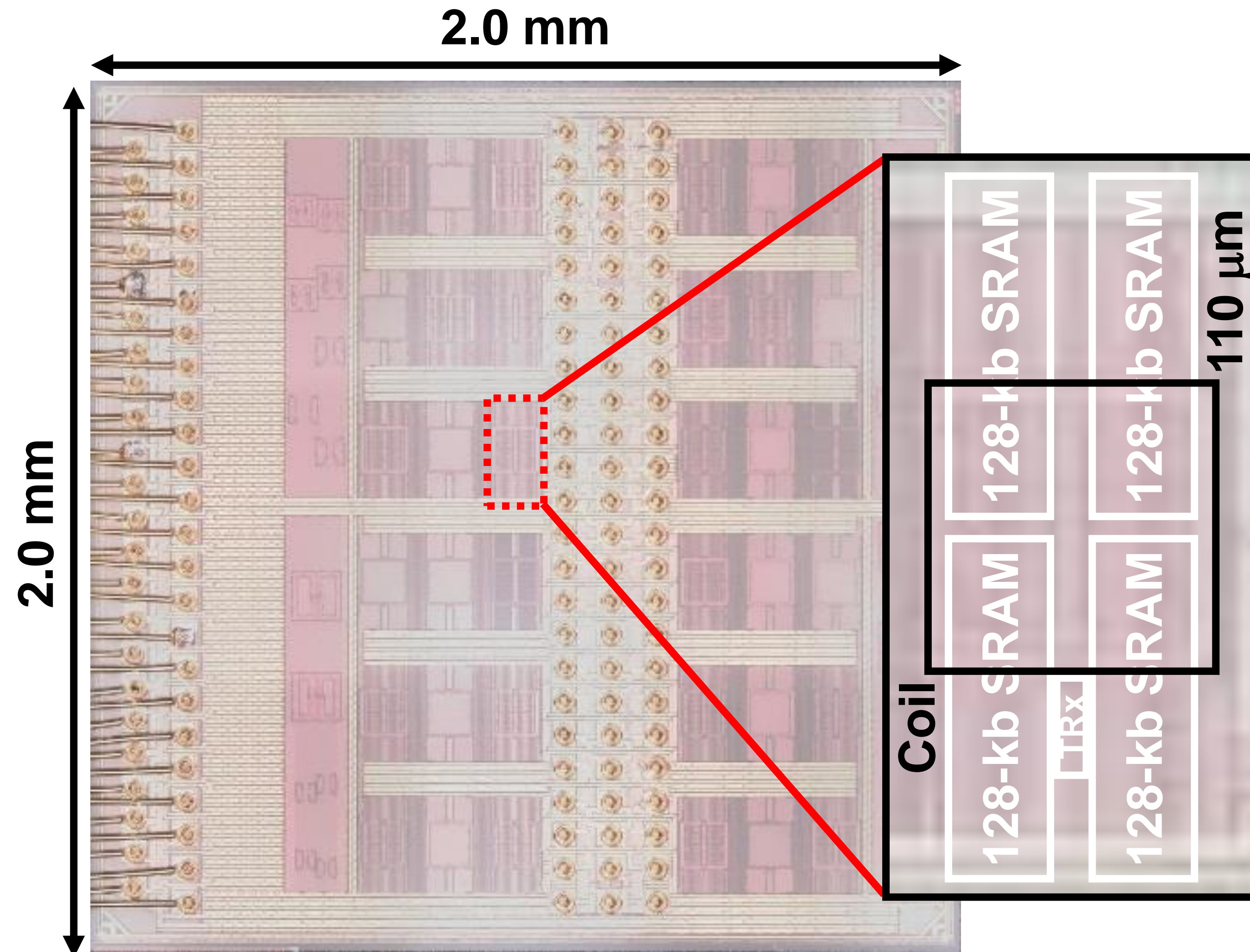


Waveforms

(b) Proposed TCI

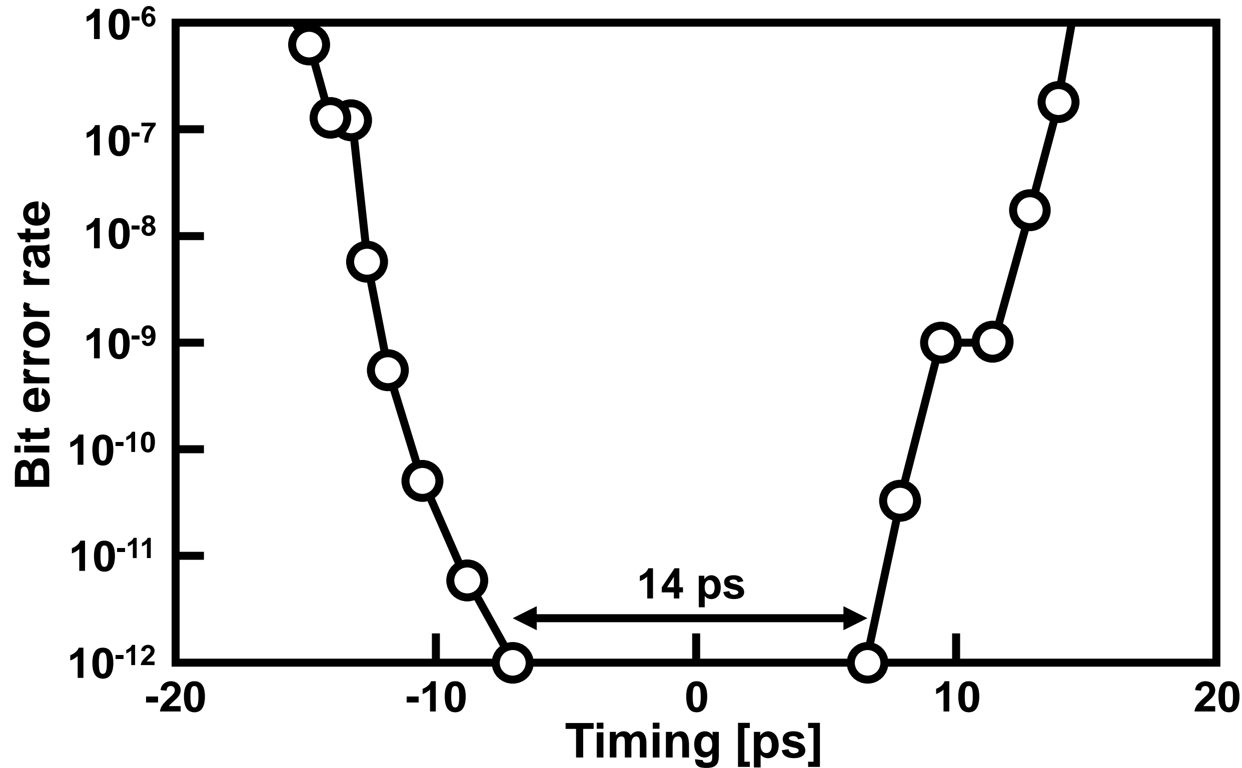
Test Chip

- Test chip was fabricated in a 7-nm FinFET process

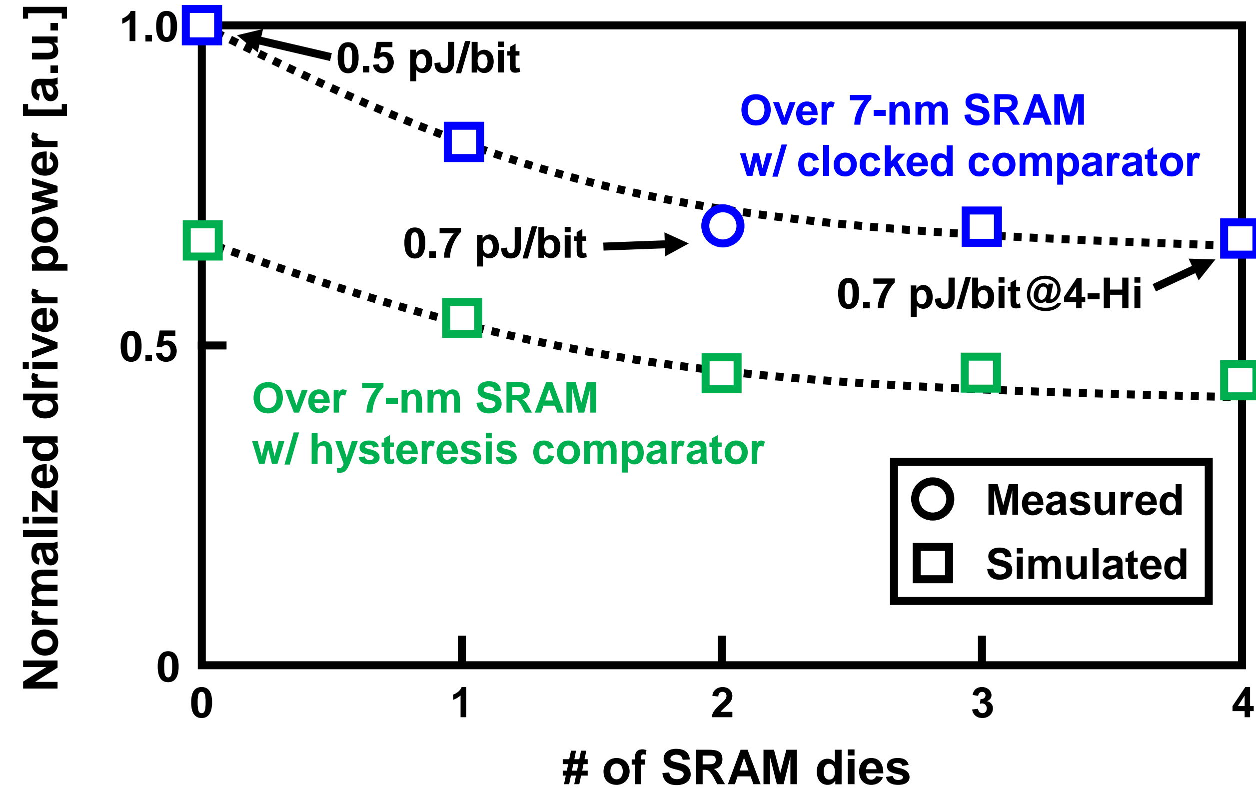


Measurement Results

- Inter-chip wireless communication at 0.7 pJ/bit, 8.5 Gbps/link measured for a 2-hi 3D-SRAM
- A 4-hi 3D-SRAM estimated to achieve 1.2 TB/s/mm^2, a two-orders-of-magnitude improvement over TSV-based 3D-SRAM [1]



(a) Measured bathtub curve



(b) Effect of sandwiched SRAMs on TCI

Performance Comparisons

RaaS 東京大学 THE UNIVERSITY OF TOKYO d.lab Systems Design Lab

	MICRO'17 [8]	ISSCC'20 [9]	Hot Chips'20 [1]	Hot Chips'20 [1] (Extrapolated to 4 Hi)	This work
Technology	20-nm DRAM	1y-nm DRAM	7-nm FinFET	7-nm FinFET	7-nm FinFET
Memory type	HBM2 DRAM	HBM2E DRAM	SRAM	SRAM	SRAM
Data bus	Bi-directional	Bi-directional	Uni-directional	Uni-directional	Uni-directional
Stack #	8	12	1	4	4
Bandwidth	256 GB/s	640 GB/s	24.3 GB/s	24.3 GB/s	4.3 GB/s
μ-bump pitch	48 / 55 μm	48 / 55 μm	40 μm	40 μm	-
IO area overhead (*1)	2.8 mm^2	2.8 mm^2	0.92 mm^2	0.92 mm^2	0.0037 mm^2
Bandwidth per IO area overhead	92 GB/s/mm^2	231 GB/s/mm^2	26 GB/s/mm^2	26 GB/s/mm^2	1162 GB/s/mm^2
Data-rate	2.0 Gb/s	5.0 Gb/s	0.76 Gb/s	0.76 Gb/s	8.5 Gb/s
I/O energy consumption	~ 2 pJ/bit	N/A (~2.5 pJ/bit(*2))	0.1 pJ/bit	0.4 pJ/bit (*3)	0.7 pJ/bit
Interface type	TSV + μ-bump	TSV + μ-bump	TSV + μ-bump	TSV + μ-bump	TCI
Chip size	12mm × 8mm	11mm × 10mm	9.0mm × 9.0mm	-	2.0mm × 2.0mm

*1: IO area only for signal excluding power
*2: Estimated from ratio of the squared voltage and stack # of HBM2 (1.2 V, 8 Hi, [8]) and HBM2E (1.1 V, 12 Hi, [9])
*3: Capacitance load of 4×# of Rx's, μ-bumps and TSVs driven by Tx compared with 1 Hi

[8] M. O'Connor, MICRO, 2017 [9] C.-S. Oh, et al., ISSCC, 2020

Conclusion

- **A 3D-stacked SRAM using inductive coupling is proposed with two new methods to minimize area overhead.**
  - **(A) Over-SRAM coils: enable high area efficiency while limiting magnetic field attenuation to 30%.**
  - **(B) Manchester-encoded synchronous transceiver: detects small received signal with low power.**
- **Test chip was fabricated in a 7-nm FinFET process.**
  - **Inter-chip wireless communication at 0.7 pJ/bit, 8.5 Gbps/link was measured for a 2-hi 3D-SRAM**
  - **A 4-hi 3D-SRAM achieves 1.2 TB/s/mm^2, a two-orders-of-magnitude improvement over conventional TSV-based 3D-SRAM.**

References

[1] K. Cho, et al., "SAINT-S: 3D SRAM Stacking Solution based on 7nm TSV technology," *IEEE Hot Chips*, Aug. 2020.

[2] S. -K. Seo, et al., "CoW Package Solution for Improving Thermal Characteristic of TSV-SiP for AI-Inference," *IEEE ECTC*, June 2021.

[3] D. Ditzel, et al., "Low-cost 3D chip stacking with ThruChip wireless connections," *IEEE Hot Chips*, Aug. 2014.

[4] K. Ueyoshi, et al., "QUEST: Multi-purpose log-quantized DNN inference engine stacked on 96-MB 3-D SRAM using inductive coupling technology in 40-nm CMOS," *IEEE JSSC*, vol. 54, no. 1, pp. 186-196, Jan. 2019.

[5] K. Shiba, et al., "A 96-MB 3D-Stacked SRAM Using Inductive Coupling with 0.4-V Transmitter, Termination Scheme and 12:1 SerDes in 40-nm CMOS," *IEEE TCAS-I*, vol. 68, no. 2, pp. 692-703, Feb. 2021.

[6] J. Chang et al., "A 7nm 256Mb SRAM in high-k metal-gate FinFET technology with write-assist circuitry for low-VMIN applications," *IEEE ISSCC*, Feb. 2017.

[7] J. Chang et al., "A 5nm 135Mb SRAM in EUV and High-Mobility-Channel FinFET Technology with Metal Coupling and Charge-Sharing Write-Assist Circuitry Schemes for High-Density and Low-VMIN Applications," *IEEE ISSCC*, Feb. 2020.

[8] M. O'Connor, et al., "Fine-Grained DRAM: Energy-Efficient DRAM for Extreme Bandwidth Systems," *MICRO-50*, Oct. 2017.

[9] C. -S. Oh et al., "A 1.1V 16GB 640GB/s HBM2E DRAM with a Data-Bus Window-Extension Technique and a Synergetic On-Die ECC Scheme," *IEEE ISSCC*, Feb. 2020.

Acknowledgement

The authors would like to thank UltraMemory Inc. and Jedat Inc. for their technical support in design, implementation, and evaluation. This work was supported by JST, ACT-X Grant Number JPMJAX210A and JSPS KAKENHI Grant Number 21J11729.