



A 65nm Logic-Compatible Embedded AND Flash Memory for In-Memory Computation of Artificial Neural Networks

Junjie Mu, Bongjin Kim

*School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore*

Agenda

- **Motivation and Background**
- **Proposed AND eFlash In-Memory Computation Macro**
 - Proposed AND eFlash cell
 - Operation of multiplication and accumulation
 - Reconfigurable Bit-precision Weights Operation
 - Multi-cycle Program-and-verify
- **Simulation Results**
- **Conclusion**

Agenda

- **Motivation and Background**
- **Proposed AND eFlash In-Memory Computation Macro**
 - Proposed AND eFlash cell
 - Operation of multiplication and accumulation
 - Reconfigurable Bit-precision Weights Operation
 - Multi-cycle Program-and-verify
- **Simulation Results**
- **Conclusion**

Motivation



fingerprint recognition



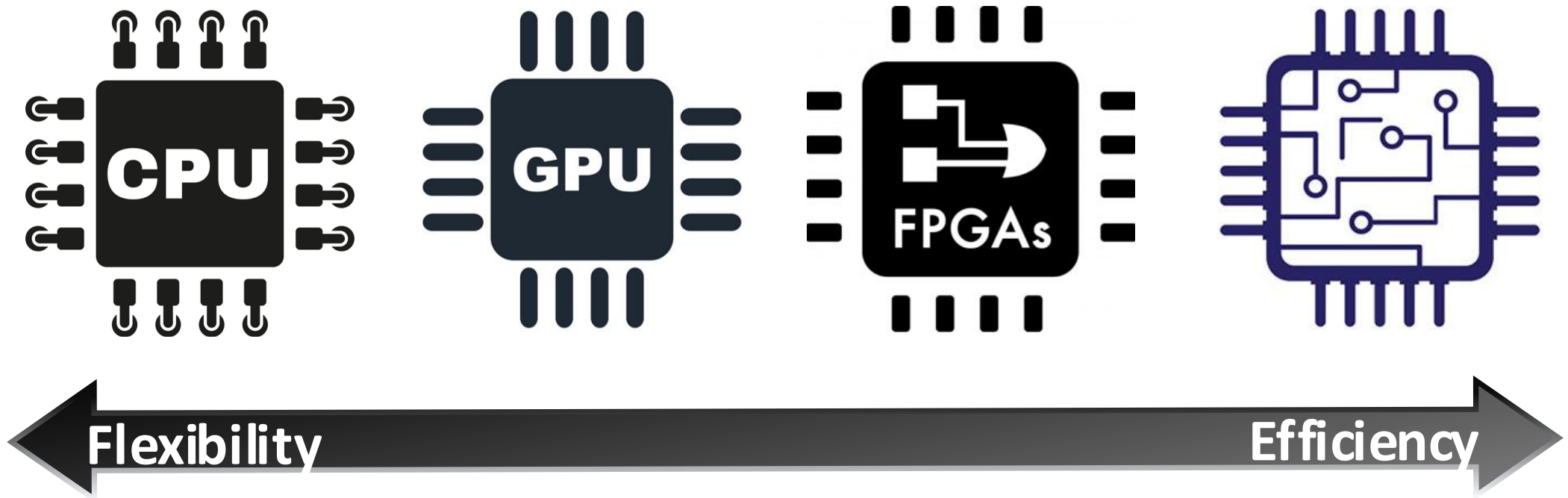
face recognition



speech recognition

- **ANNs has demonstrated excellent capability in pattern recognition.**
- **Low-power & low-latency AI chips are needed.**

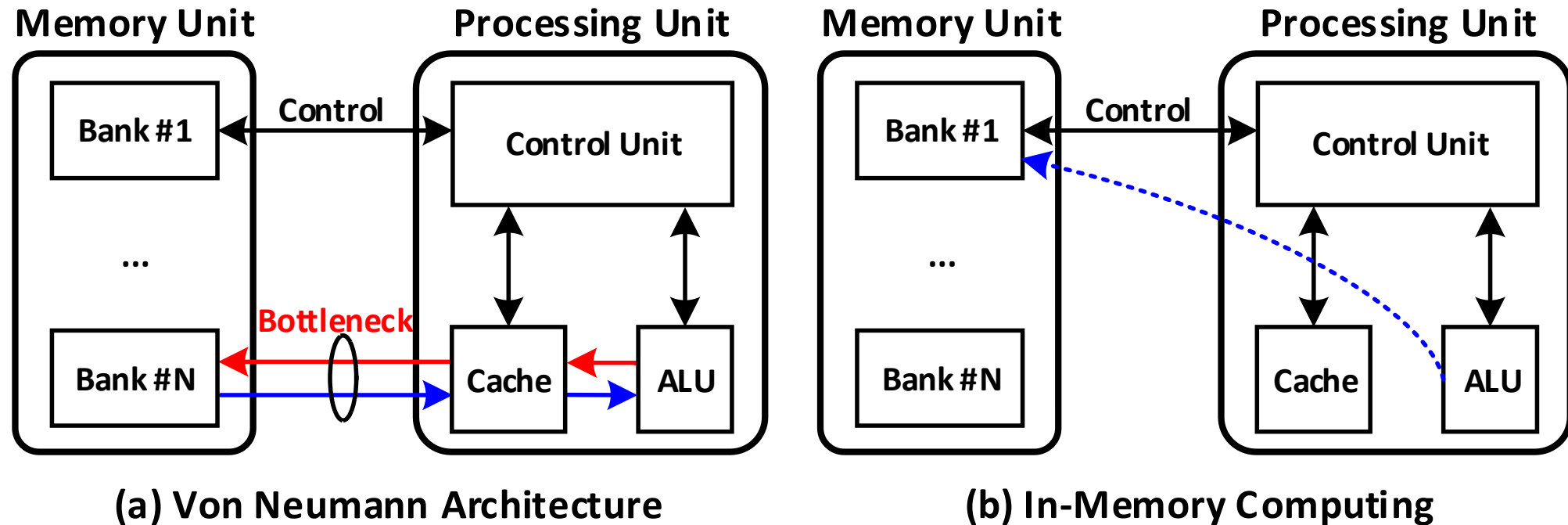
Motivation



- General Purpose
- Serialized Workloads
- Power Hungry

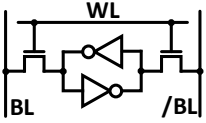
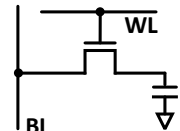
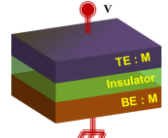
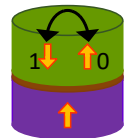

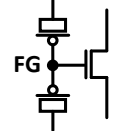
- Special Purpose
- Dedicated Workloads
- Low Hungry

In-Memory Computing

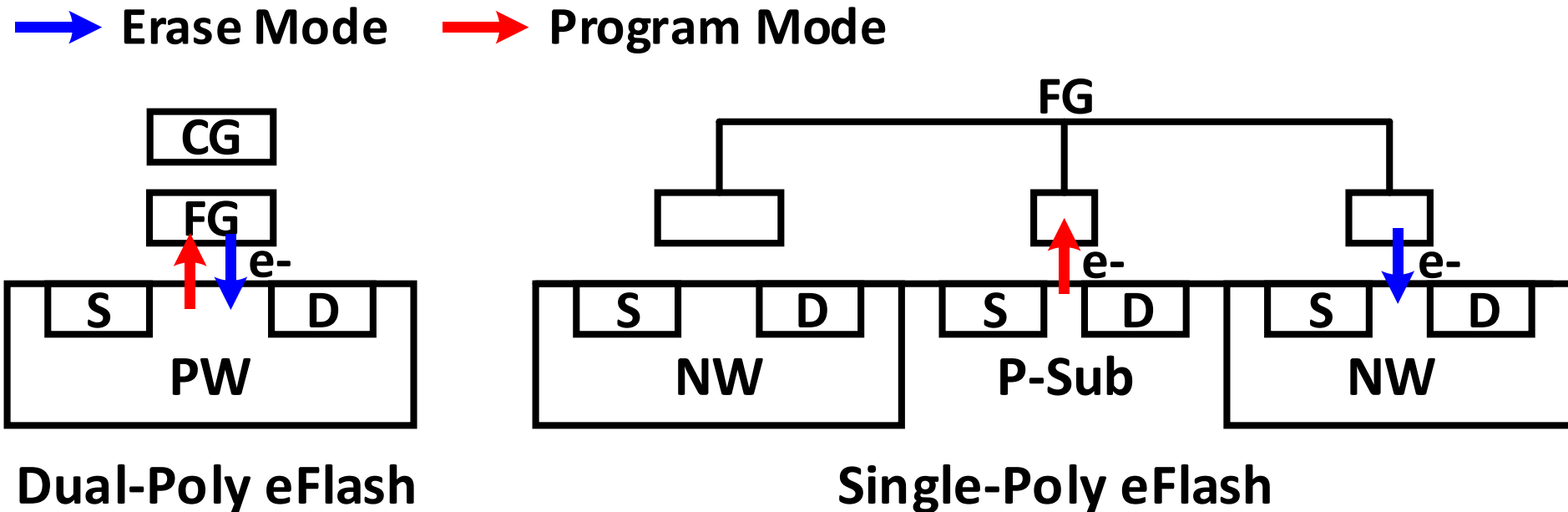


- **Von-Neumann bottleneck:** Throughput & energy cost by memory access
- **In-Memory Computing** brings processing into memory

Memory Options

Device	SRAM	eDRAM	RRAM	MRAM	PRAM	eFlash
Cell Configuration						
Nonvolatile	No	No	Yes	Yes	Yes	Yes
Logic Compatible	Yes	No	No	No	No	Yes
Storage	Latch	Capacitor	Resistance	Resistance	Resistance	Floating Gate
Cell Leakage	Middle	High	Low	Low	Low	Low
Cell Size(F ²)	150	40	60	40	40	~550
Multi-level storage	NO	NO	YES	NO	YES	YES

Dual-poly vs Single-poly eFlash



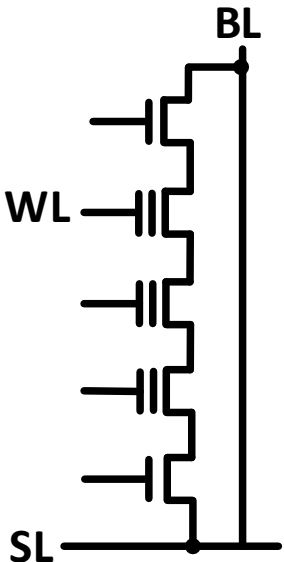
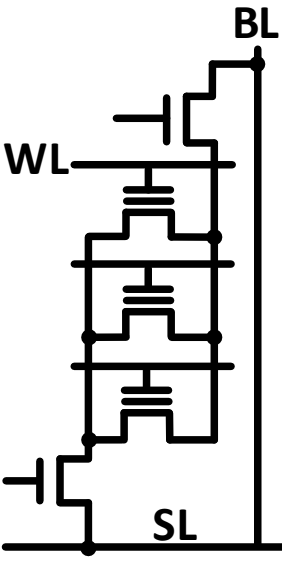
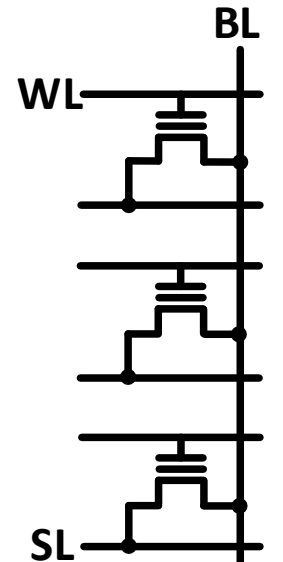
S. Song, JSSC'13[2]

- Dual-Poly eFlash: additional process is required for floating gate.
- Single-Poly eFlash: Back-to-back connected gate, logic-compatible.

Agenda

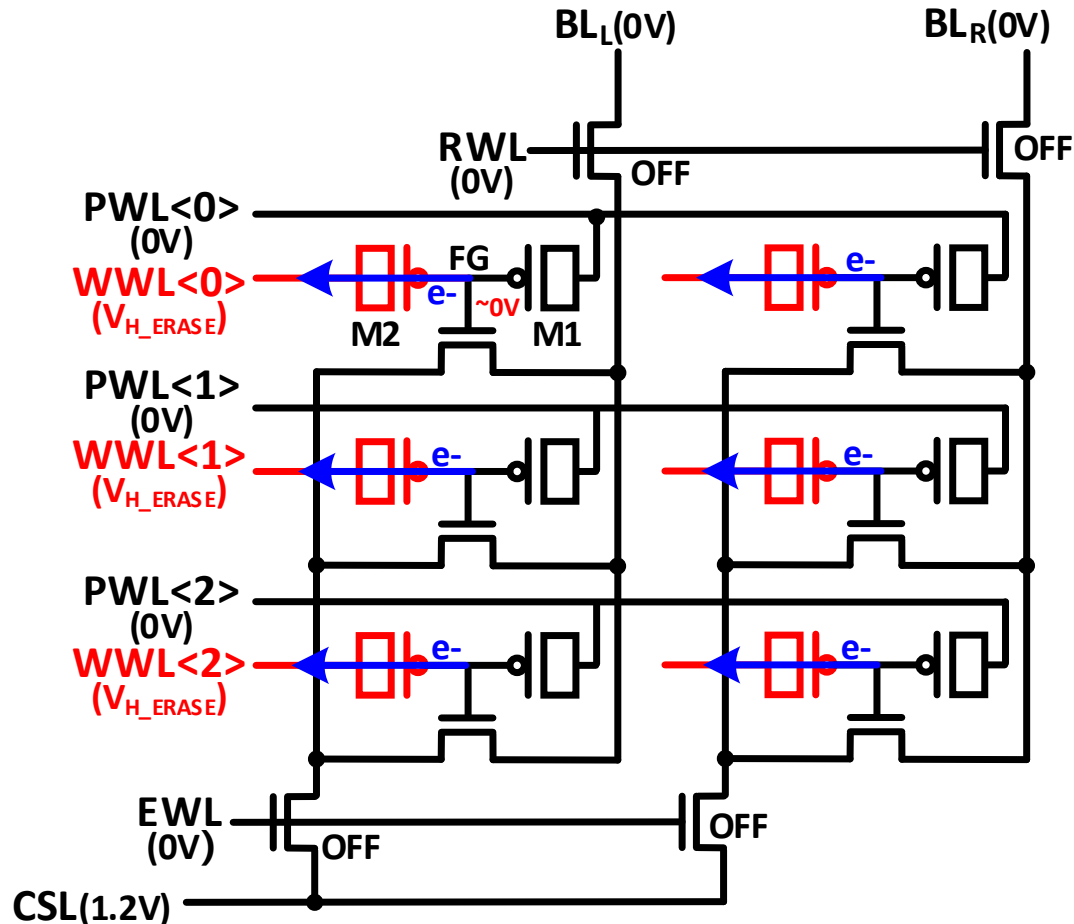
- Motivation and Background
- **Proposed AND eFlash In-Memory Computation Macro**
 - Proposed AND eFlash cell
 - Operation of multiplication and accumulation
 - Reconfigurable Bit-precision Weights Operation
 - Multi-cycle Program-and-verify
- Simulation Results
- Conclusion

eFlash Structure Comparison

Structure	NAND	AND	NOR
Cell Array			
Size	$4F^2$	$8F^2$	$10F^2$
Access	Sequential	Random	Random

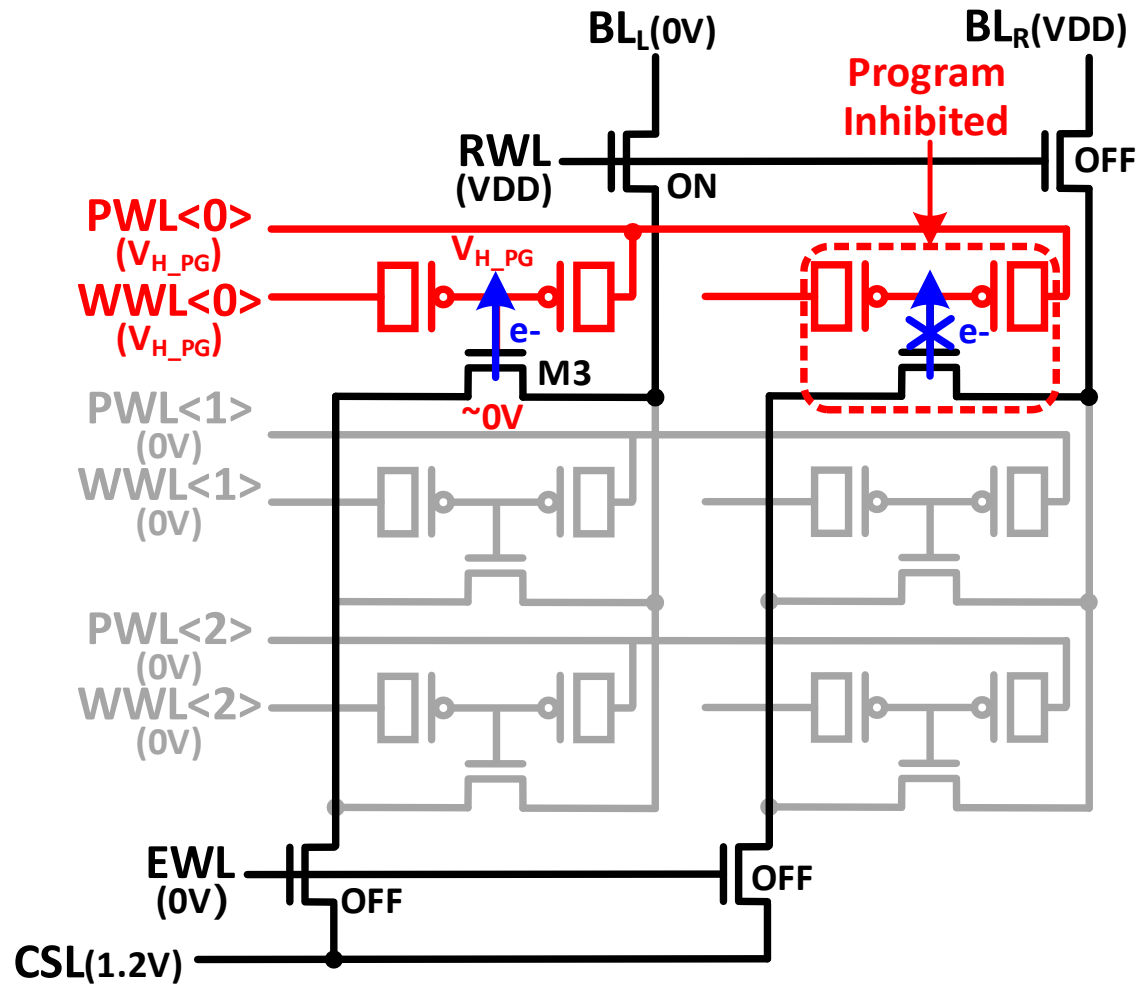
- Higher density
- Faster access speed

And eFlash Operation—Erase



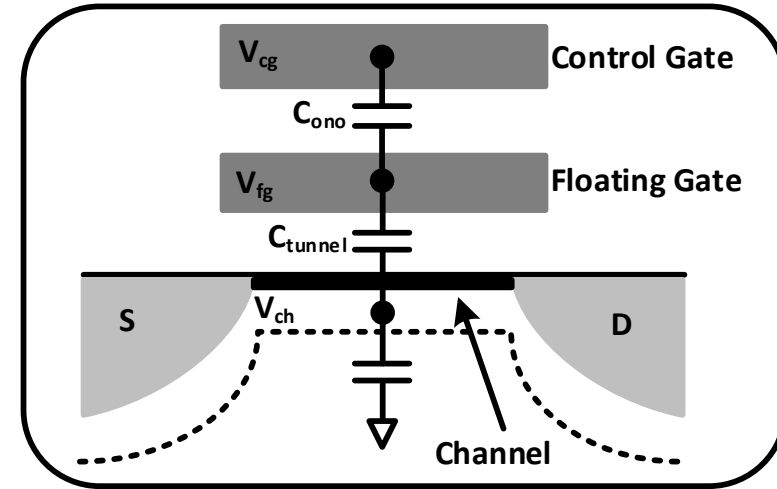
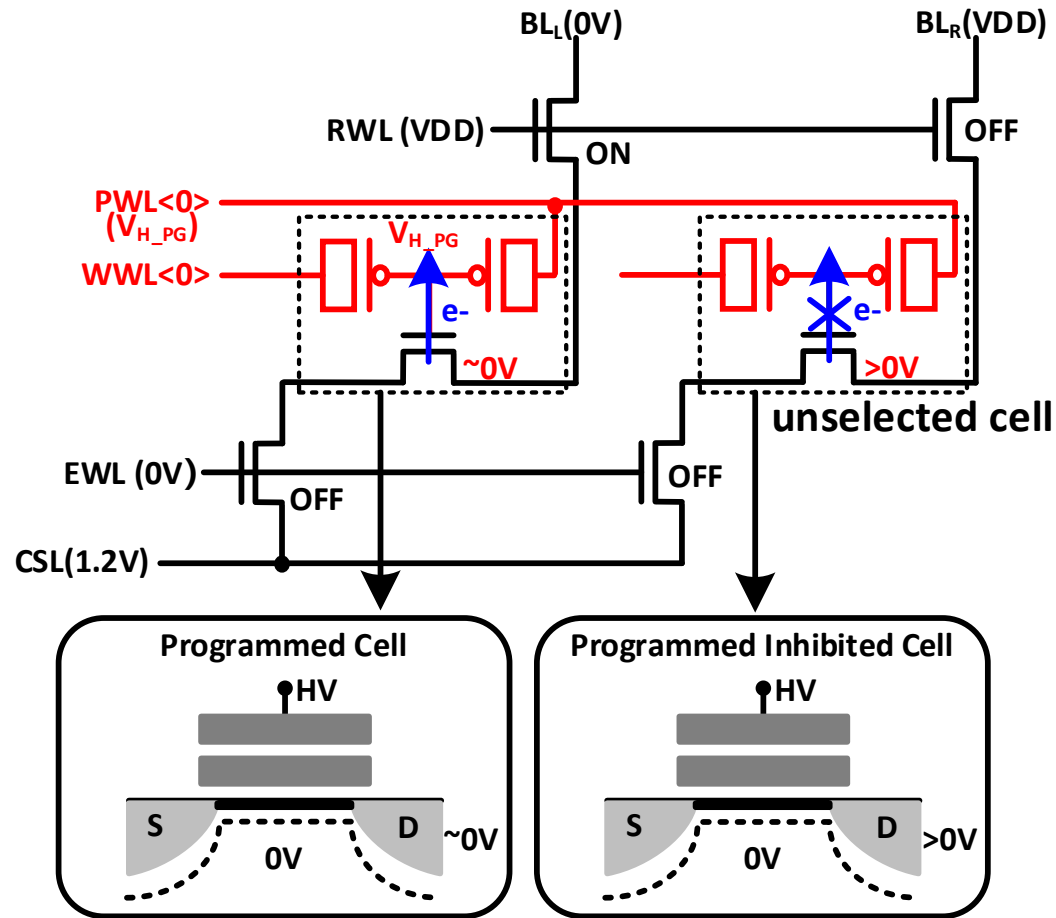
- Erase: the process of write '1'.
- V_{H_ERASE} is applied to all WWLs, erasing all cells in parallel.
- Fowler–Nordheim tunnelling utilized for erasing.

And eFlash Operation—Program



- **Program: the process of write '0'.**
- **V_{H_PG} is applied to selected PWL/WWL, cell-by-cell program.**
- **Fowler–Nordheim tunnelling utilized for programming.**

And eFlash Operation—Program Inhibit



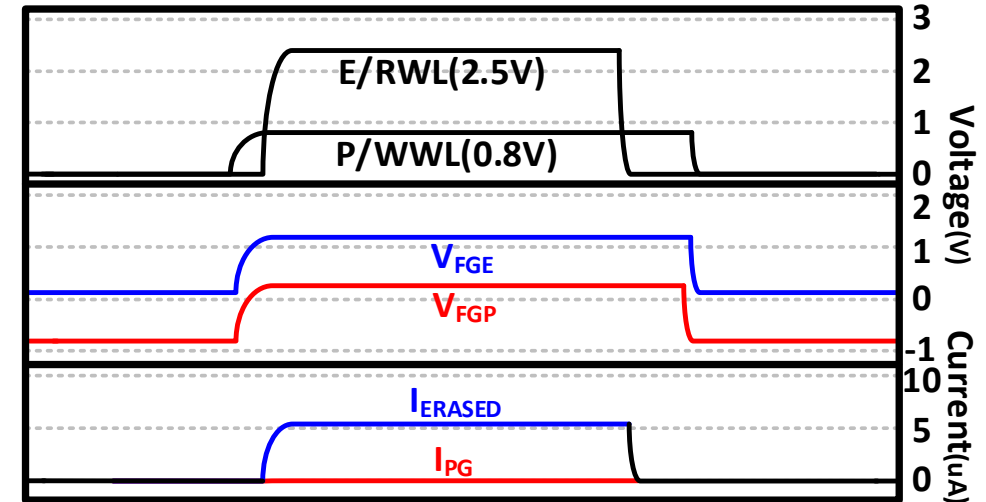
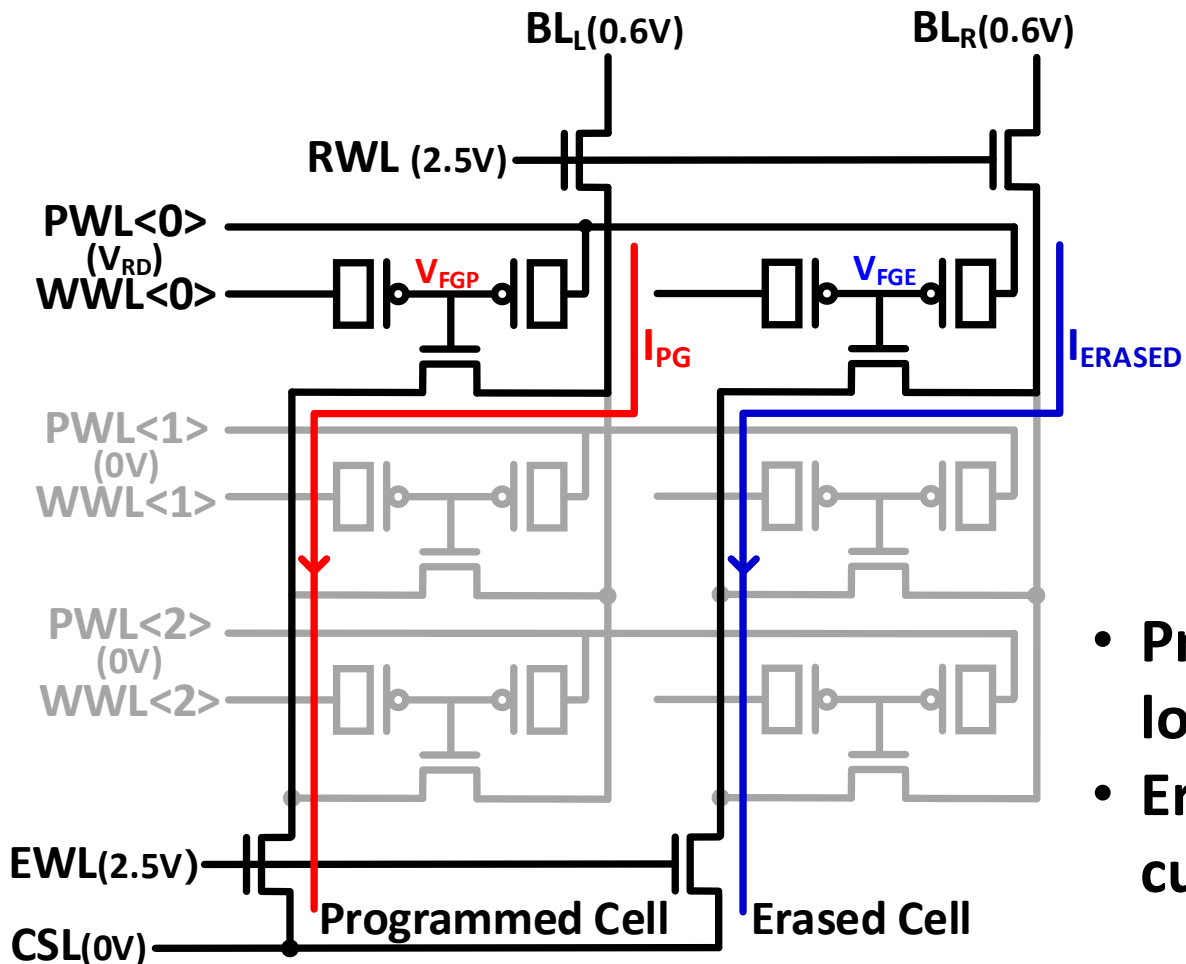
$$V_{ch} = \frac{C_{ins}}{C_{ins} + C_{channel}} V_{cg}$$

$$C_{ins} = C_{ono} || C_{tunnel}$$

Joe E. Brewer[3]

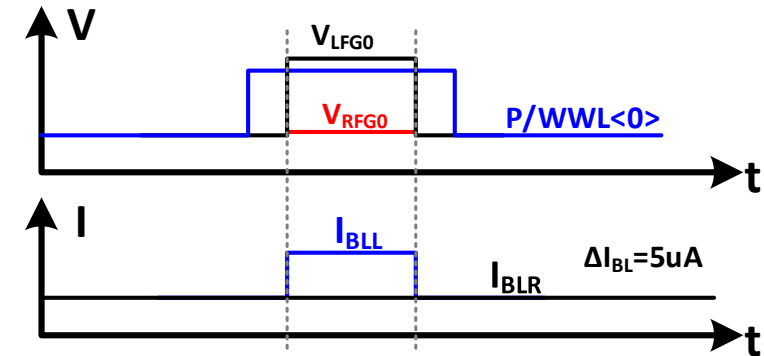
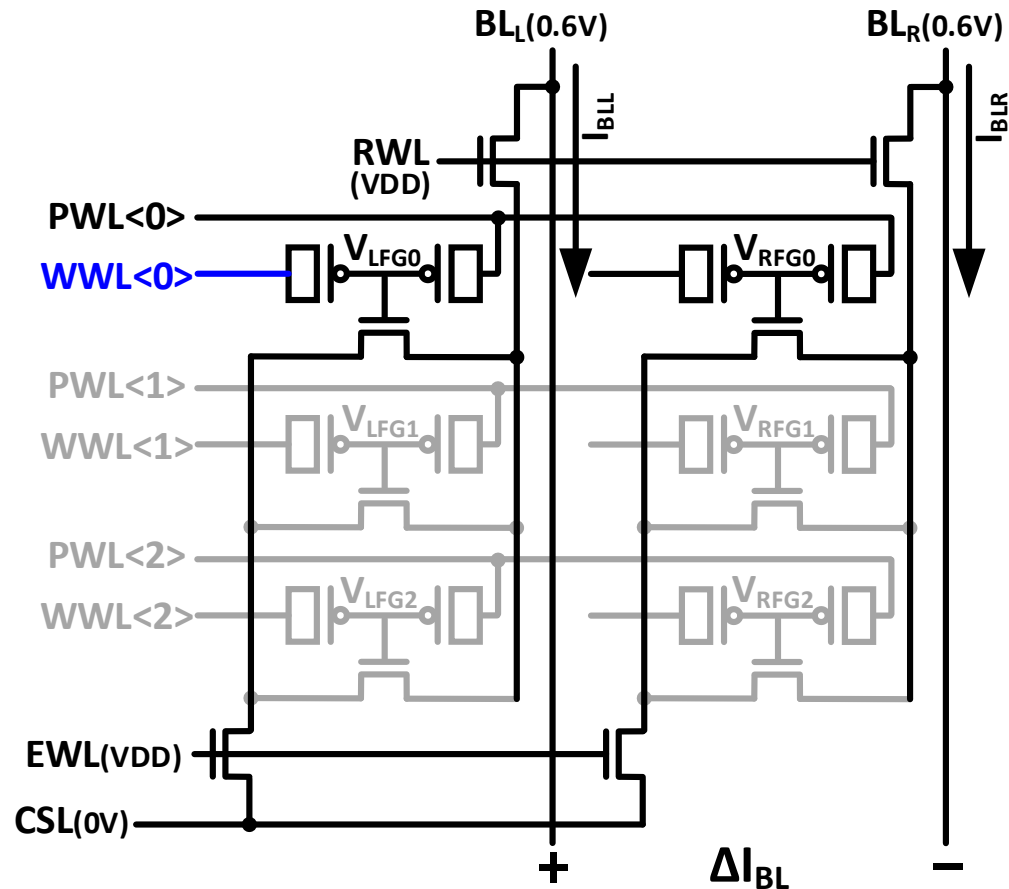
- Program inhibit of unselected cells via self-boosting.

And eFlash Operation—Read



- Programmed Cell: Low floating gate voltage, low BL current(~0).
- Erased Cell: High floating gate voltage, high BL current.

Multiplication

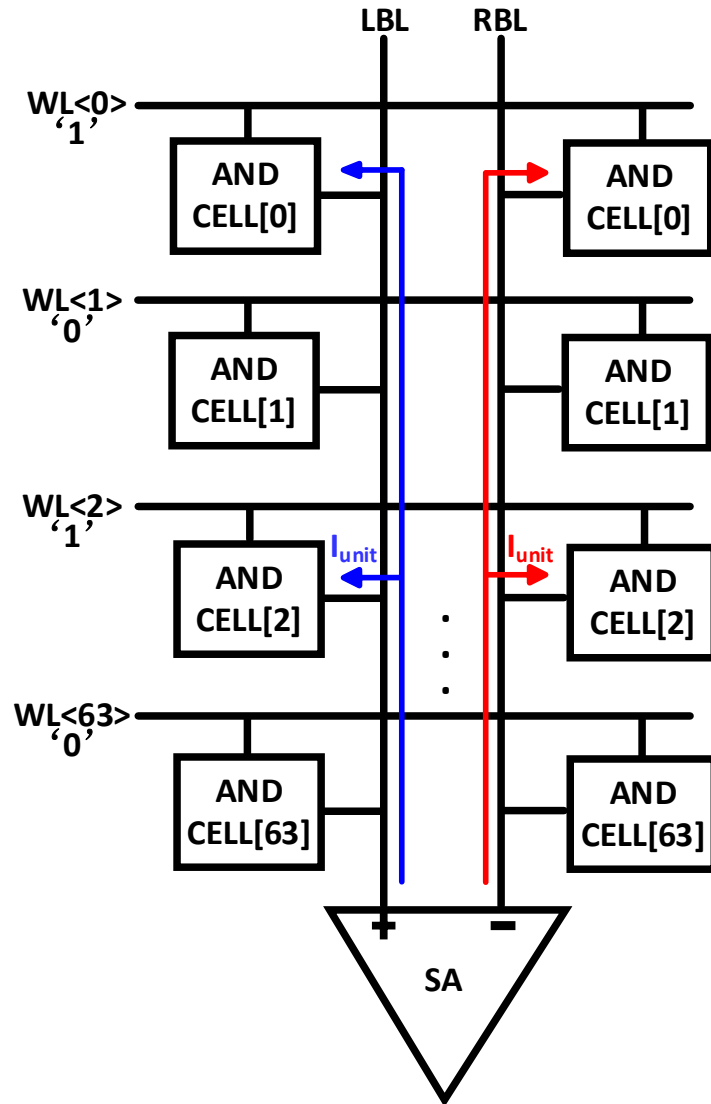


Input Weight	0 (*WL=0V)	1 (*WL=0.8V)
-1 ($V_{LFG0}=L, V_{RFG0}=H$)	0 μA	-5 μA
0 ($V_{LFG0}=L, V_{RFG0}=L$)	0 μA	0 μA
1 ($V_{LFG0}=H, V_{RFG0}=L$)	0 μA	5 μA

*WL=PWL=WWL for Read OP

- 2-level weight storage & Read 1 cell/cycle.
- ΔI_{BL} represents the product between input and weight.

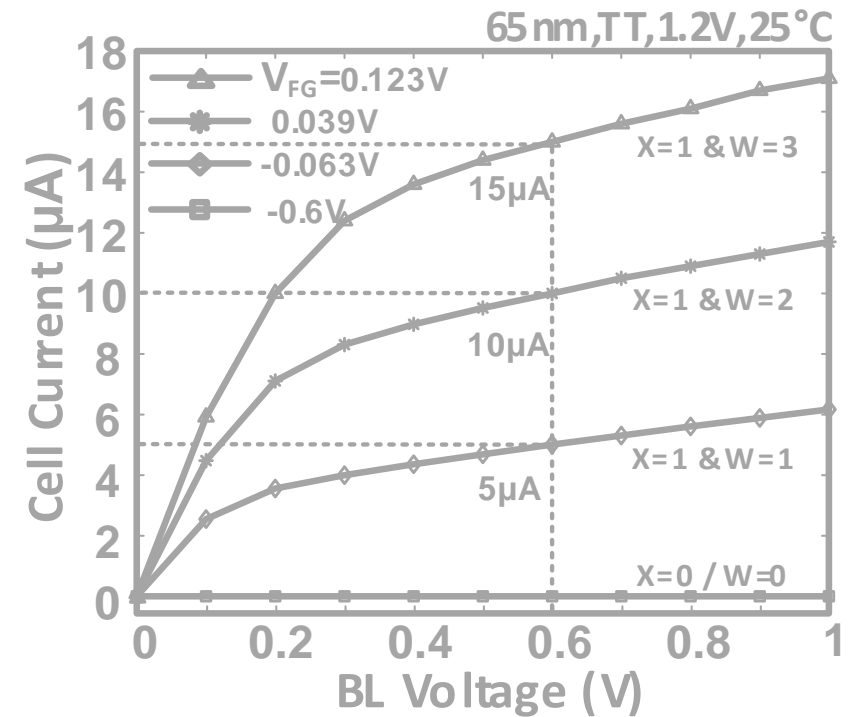
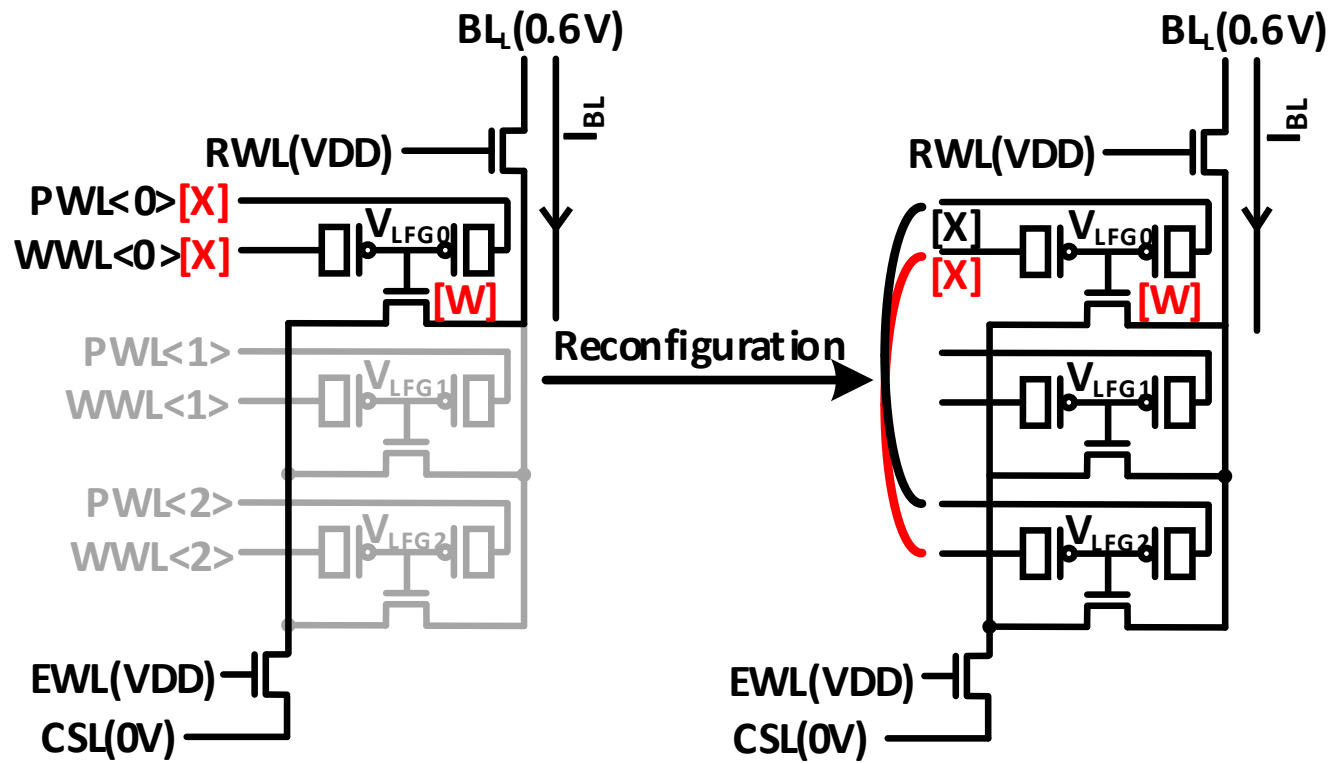
Accumulation



$$I_{ACC} = \sum_{i=1}^n W_i X_i$$

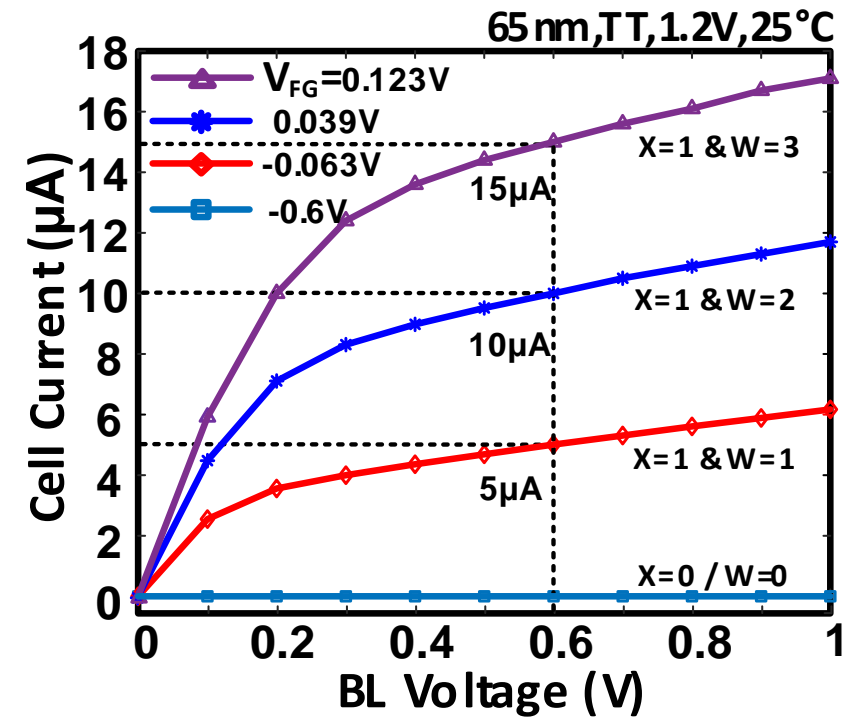
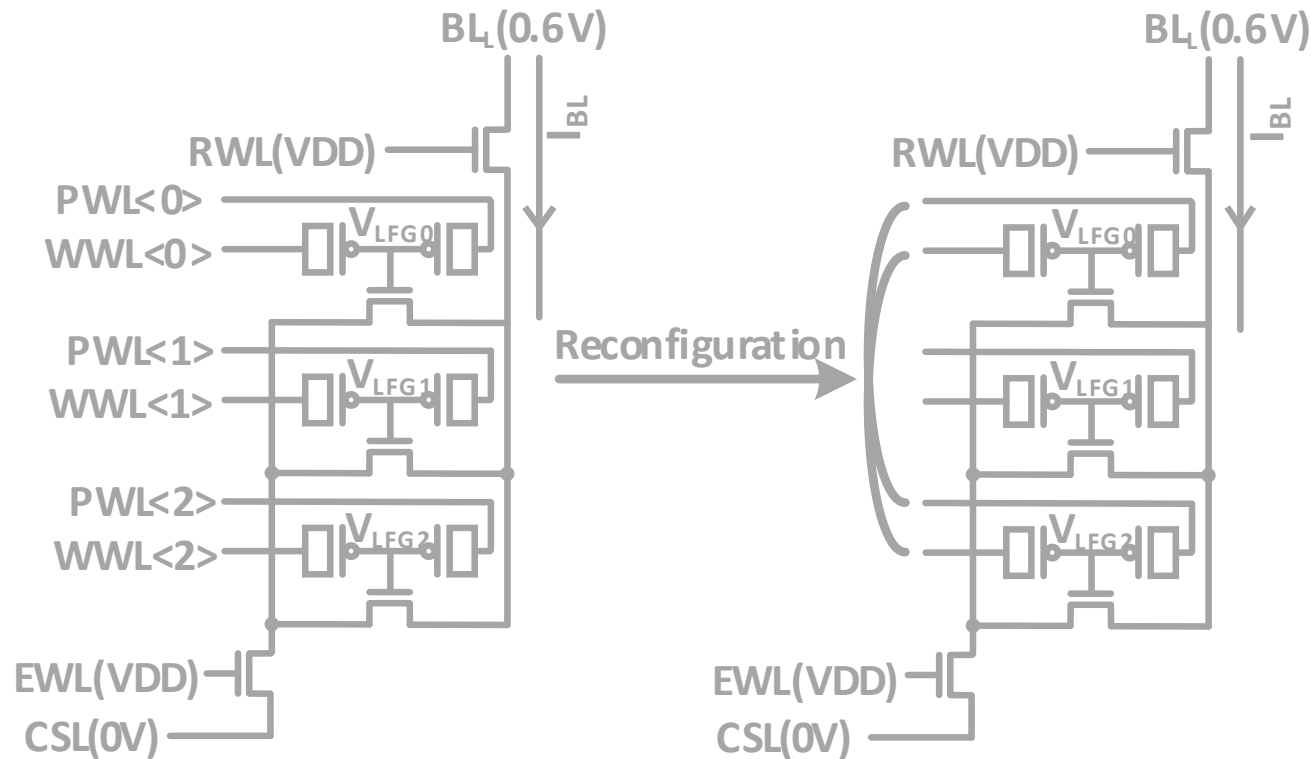
- 64 AND cells are connected in one column.
- The total current on BL is the accumulation result.
- Adjacent two columns BL voltage are compared through SA as the binary output activation.

Multi-Level Weight



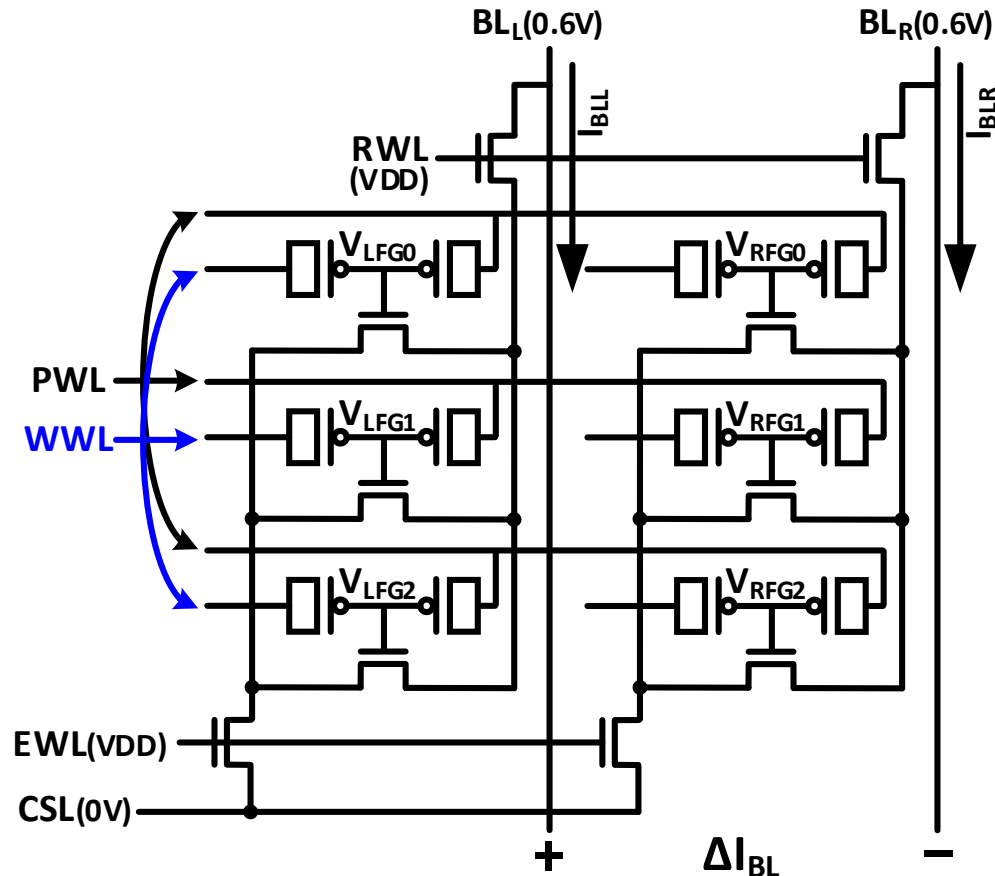
- 3x 3T eFlash cells are grouped as a computation unit, one-cycle operation.

Multi-Level Weight



- Different cell currents can be programmed to implement multi-level weight storage.

Multi-Level Weight Multiplication



Input Weight	0 (*WL=0V)	1 (*WL=0.8V)
-6	0 uA	-30 uA
-5	0 uA	-25 uA
-4	0 uA	-20 uA
-3	0 uA	-15 uA
-2	0 uA	-10 uA
-1	0 uA	-5 uA
0	0 uA	0 uA
1	0 uA	5 uA
2	0 uA	10 uA
3	0 uA	15 uA
4	0 uA	20 uA
5	0 uA	25 uA
6	0 uA	30 uA

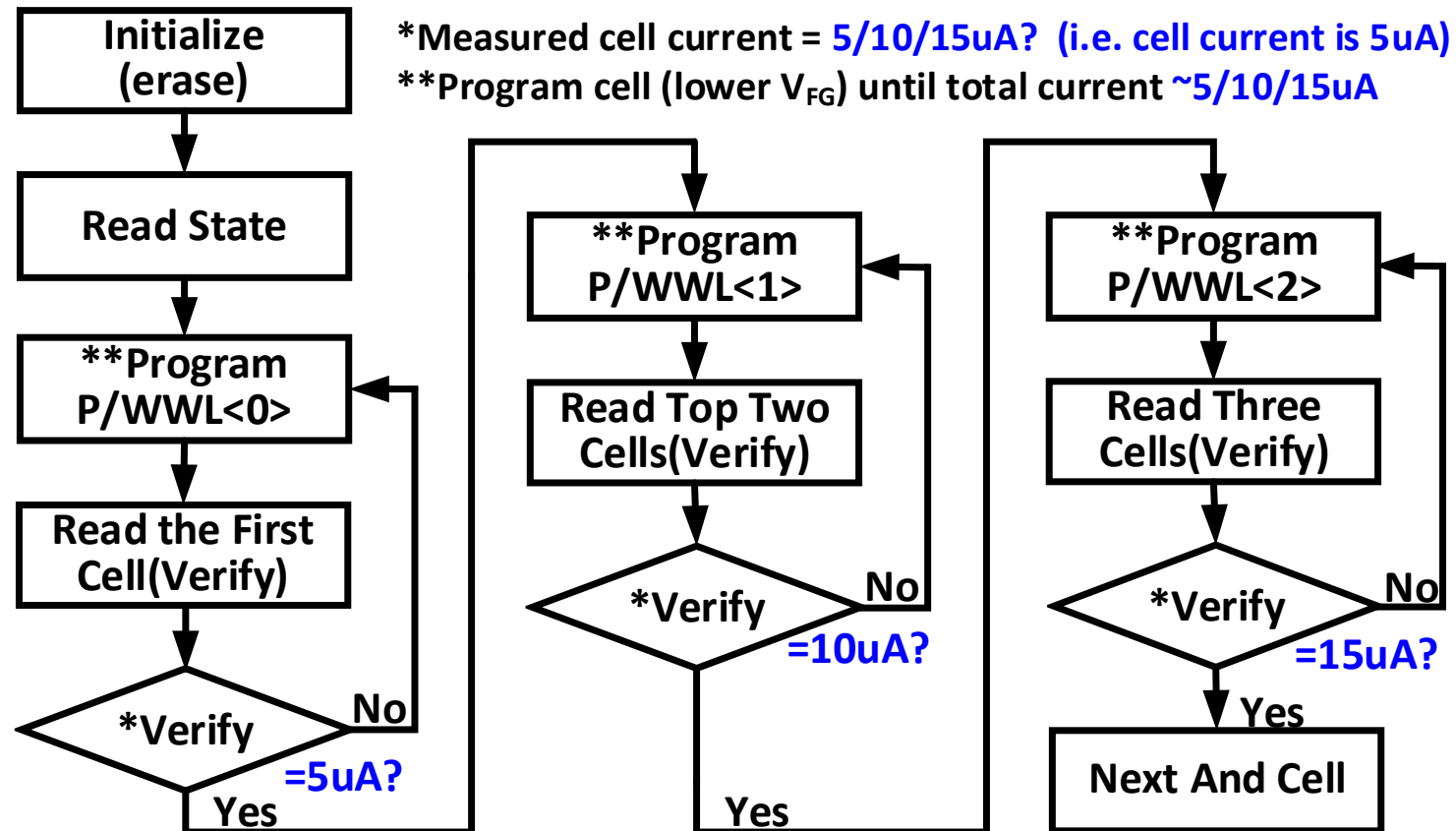
*WL=PWL=WWL for Read OP

- 3x 3T eFlash cells are grouped as one computation unit.
- Each eFlash cell stores three-level weight, 13-level(3.7bit) weight and binary input operation.

Reconfigurable Bit-Precision Weights

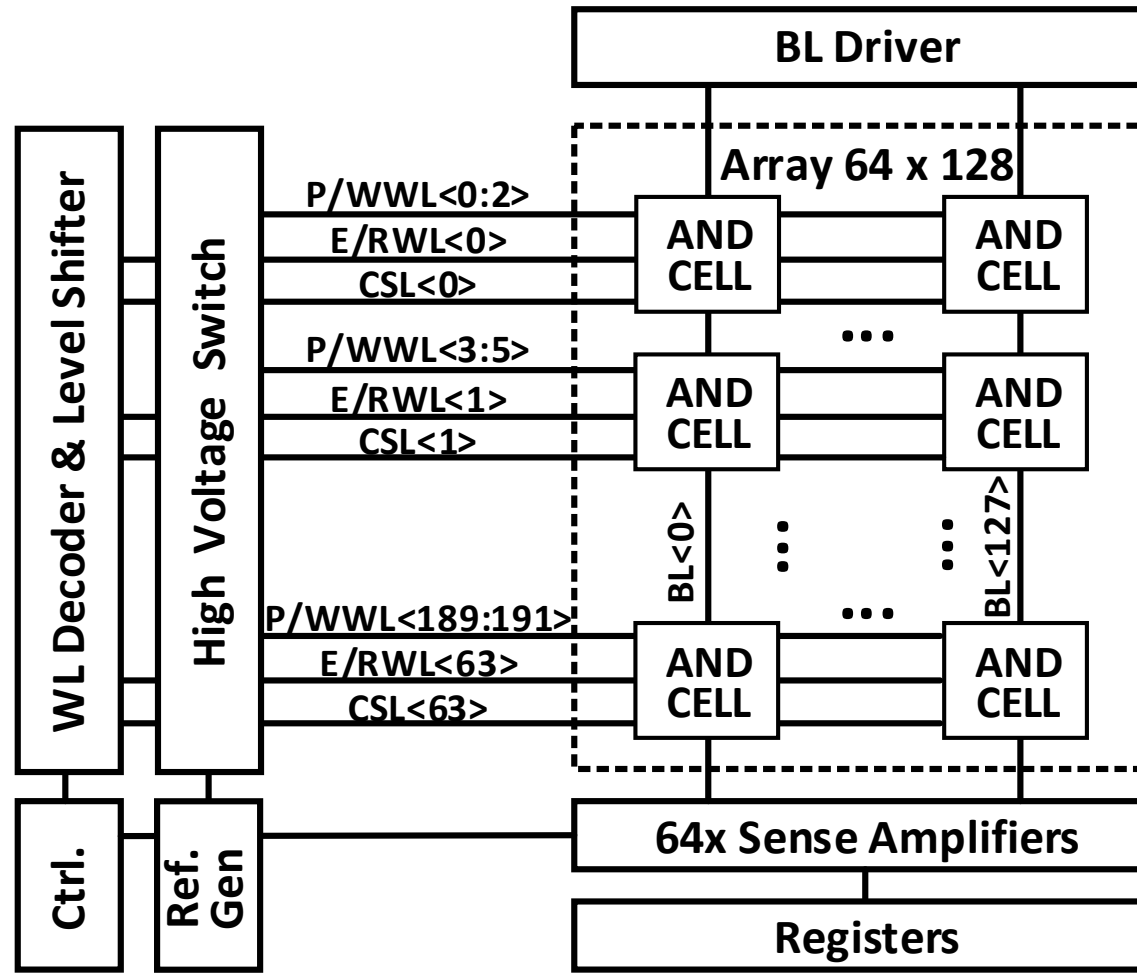
Weight Bit-Precision		Binary Data Storage	Three-Level Data Storage
Three 3T eFlash (11T)	Three-Cycle Operation	1.58b	2.32b
	One-Cycle Operation	2.81b	3.7b
Five 3T eFlash (17T)	Five-Cycle Operation	1.58b	2.32b
	One-Cycle Operation	3.46b	4.39b
Seven 3T eFlash (23T)	Seven-Cycle Operation	1.58b	2.32b
	One-Cycle Operation	3.91b	4.86b

Proposed Cell-by-cell Calibration



- A three-step program-and-verify operation to improve the linearity.
- Iterate program and verify until current is 5/10/15uA.

Array Architecture

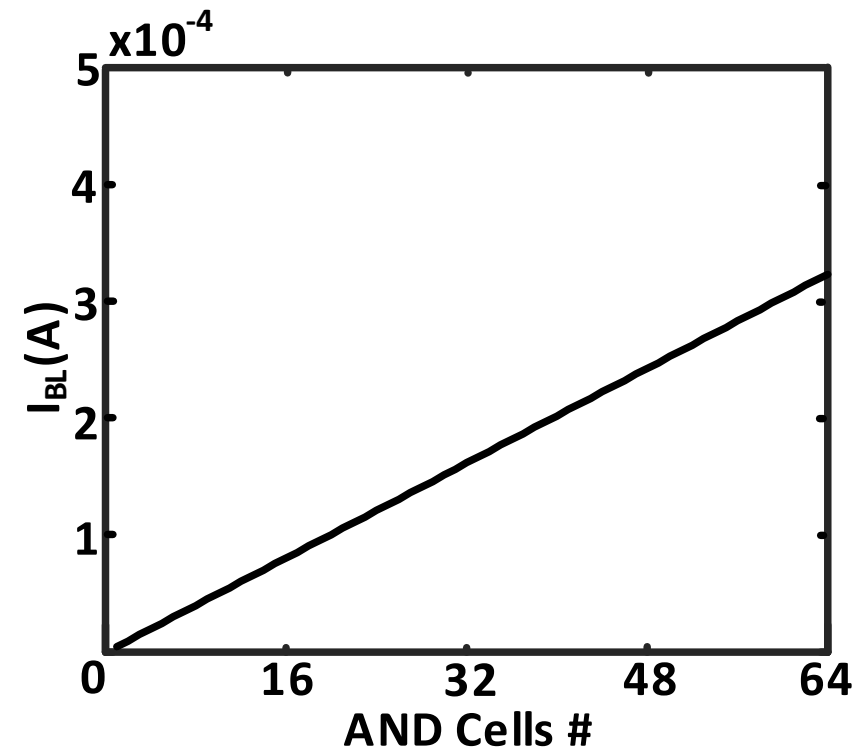
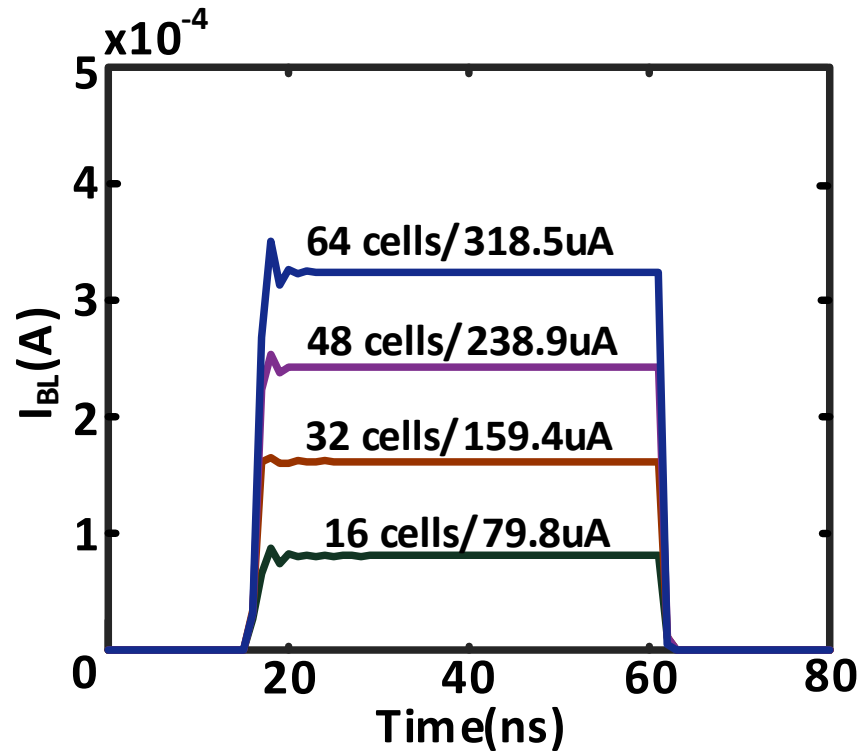


- Compute 128x parallel dot-product using binary inputs and multi-level weights.

Agenda

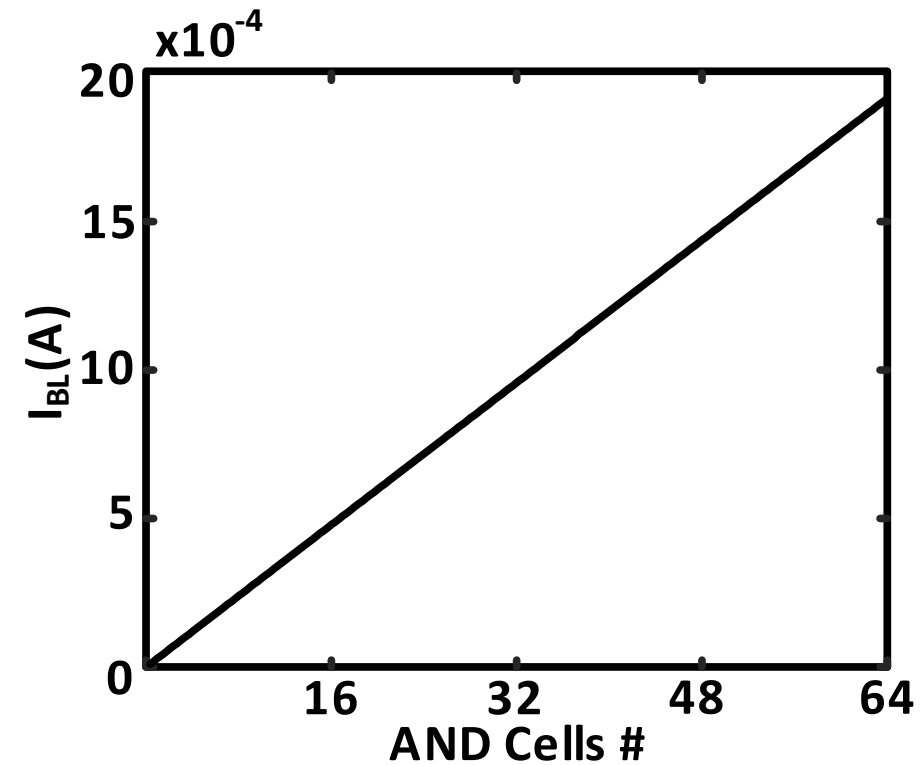
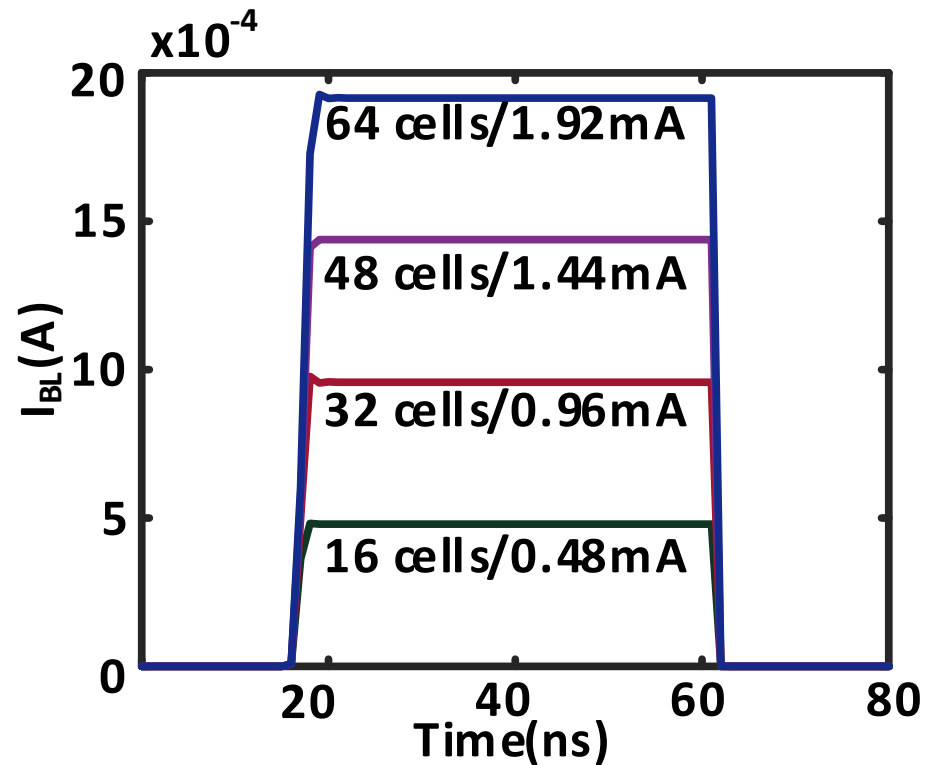
- Motivation and Background
- Proposed AND eFlash In-Memory Computation Macro
 - Proposed AND eFlash cell
 - Operation of multiplication and accumulation
 - Reconfigurable Bit-precision Weights Operation
 - Multi-cycle Program-and-verify
- **Simulation Results**
- Conclusion

Linearity Simulation Result



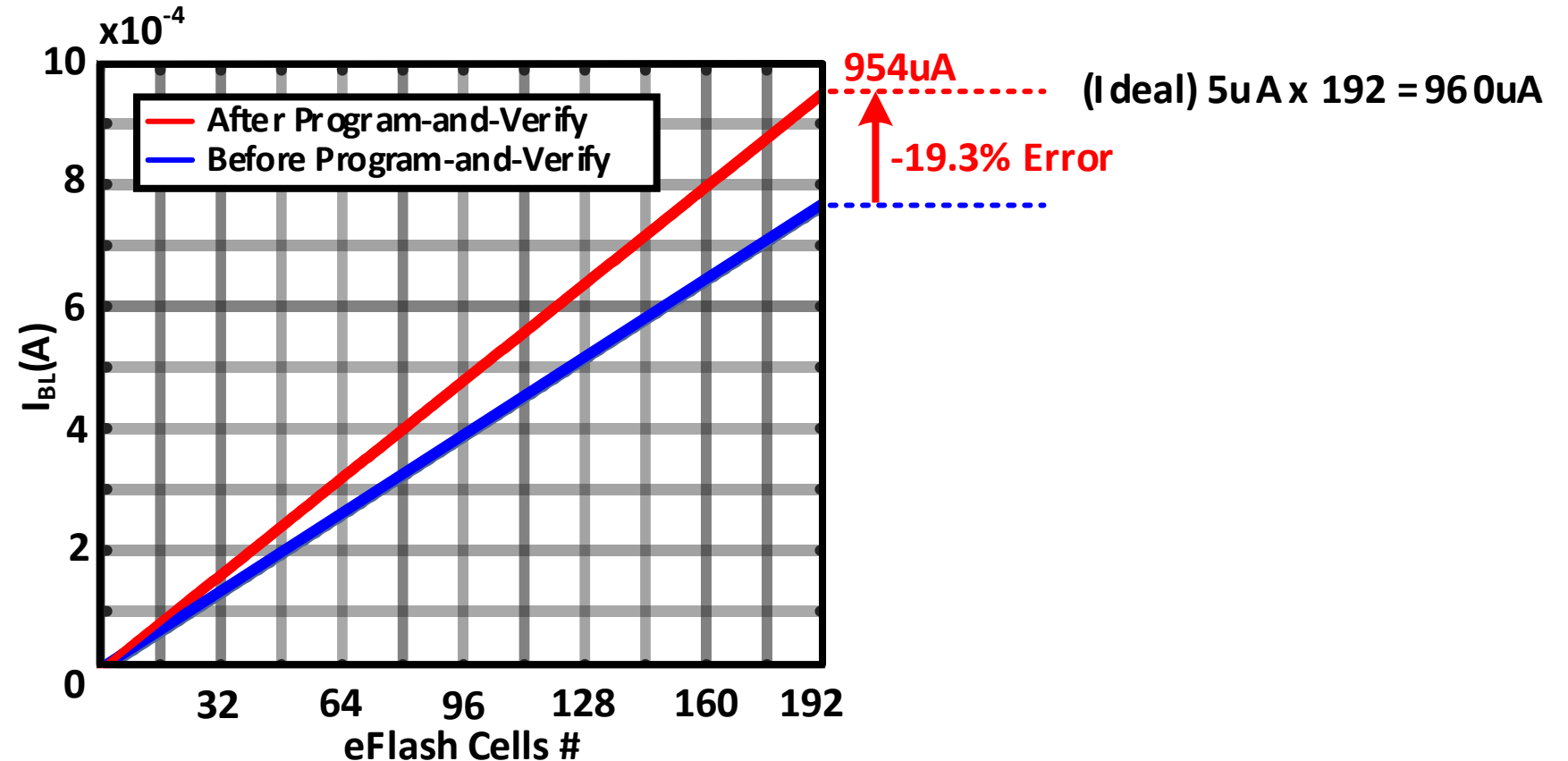
- Single column consists of 64 AND cells, two-level weight storage and three-cycle operation.
- Uniform Interval / good linearity.

Linearity Simulation Result



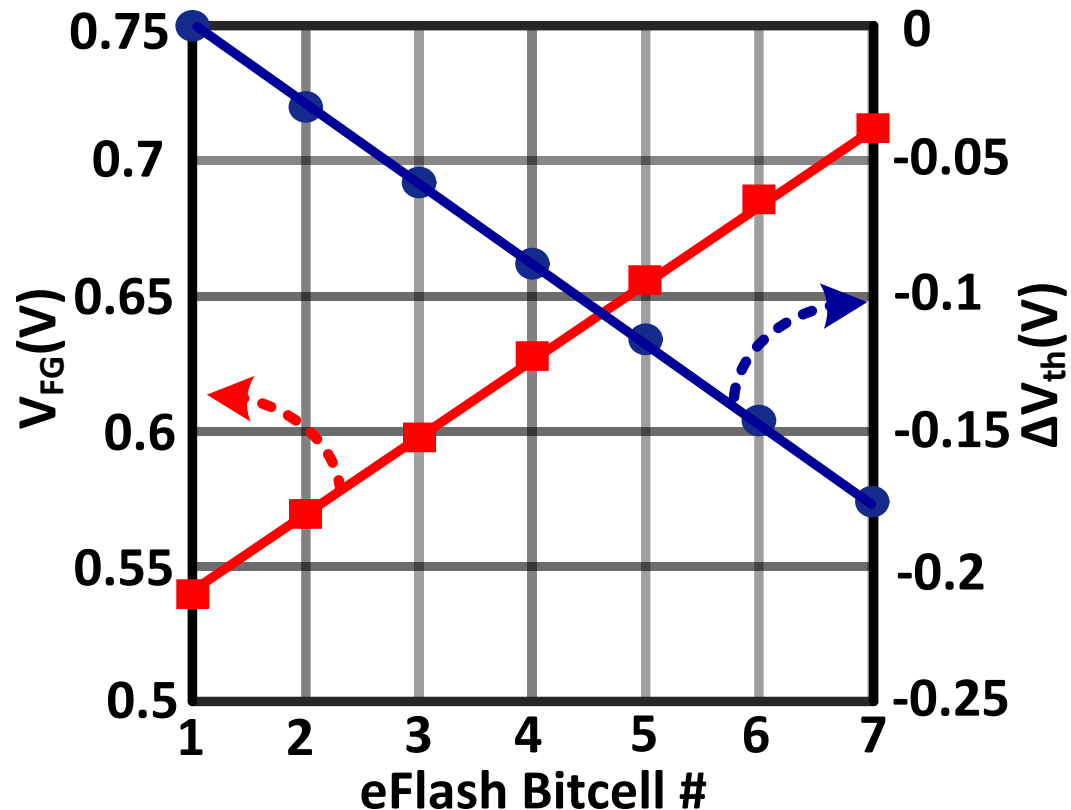
- Single column consists of 64 AND cells, three-level weight and one-cycle operation.
- Uniform Interval / good linearity.

Proposed Verify Simulation



- The BL current increases from 0 to 954 μ A (i.e. -0.6% compared to the ideal 960 μ A).
- The current deviation is reduced by 19.3% after the program-and-verify operation.

Endurance and Retention



- The more the number of eFlash bitcells, the less wearing incurred by program operation.
- Better endurance and retention performance are realized.

Performance Table /Comparison

Technology	65nm LP CMOS
Memory	AND eFlash
Unit And Cell Area	1.29 x 20.7 μm^2
Array Size	(64X3) x 128
Supply Voltage	1.2 V
#Level for Input	Binary(0/1)
#Level for Weight	3,5,7 and 13
Column Sensing	Differential

	IEDM'17 [13]	IEDM'17 [16]	JSSC'19 [2]	This Work
Cell	NOR eFlash	ReRAM	SRAM	AND eFlash
Technology	180nm	150nm	28nm	65nm
Logic Compatible	NO	NO	YES	YES
MAC Array Size	101,780	16Mb	256x64	(64x3)x128
Accumulator Type	current-mode	voltage-mode	voltage-mode	current-mode
Precision (weight/input)	1.6b/1b	2b/1b	1b/6b	1.58-3.7b/1b

Agenda

- Motivation and Background
- Proposed AND eFlash In-Memory Computation Macro
 - Proposed AND eFlash cell
 - Operation of multiplication and accumulation
 - Reconfigurable Bit-precision Weights Operation
 - Multi-cycle Program-and-verify
- Simulation Results
- **Conclusion**

Conclusion

- **AND-Type eFlash memory based in-memory computing macro for processing dot-products is proposed.**
- **Reconfigurable bit-precision weights is realized by using one- or multi-cycle operation and multi-level data storage in one eFlash bitcell.**
- **Multi-cycle program-and-verify process to improve the linearity of the accumulated bitcell current in a multi-bit weight computing.**

Thank you!

References

- 1) M. Kim et al., “A 68 parallel row access neuromorphic core with 22K multi -level synapses based on logic-compatible embedded flash memory technology,” IEEE IEDM, Dec. 2018.
- 2) S. Song et al., “A logic-compatible embedded flash memory for zero-standby power system-on-chips featuring a multi-story high voltage switch and a selective refresh scheme,” JSSC, May 2013.
- 3) Joe E. Brewer, Manzur Gill, “Non-volatile memory technologies with emphasis on flash,” John Wiley & Sons, Inc., Hoboken, New Jersey, 2008. pp. 235–236 .