Flash Memory Summit

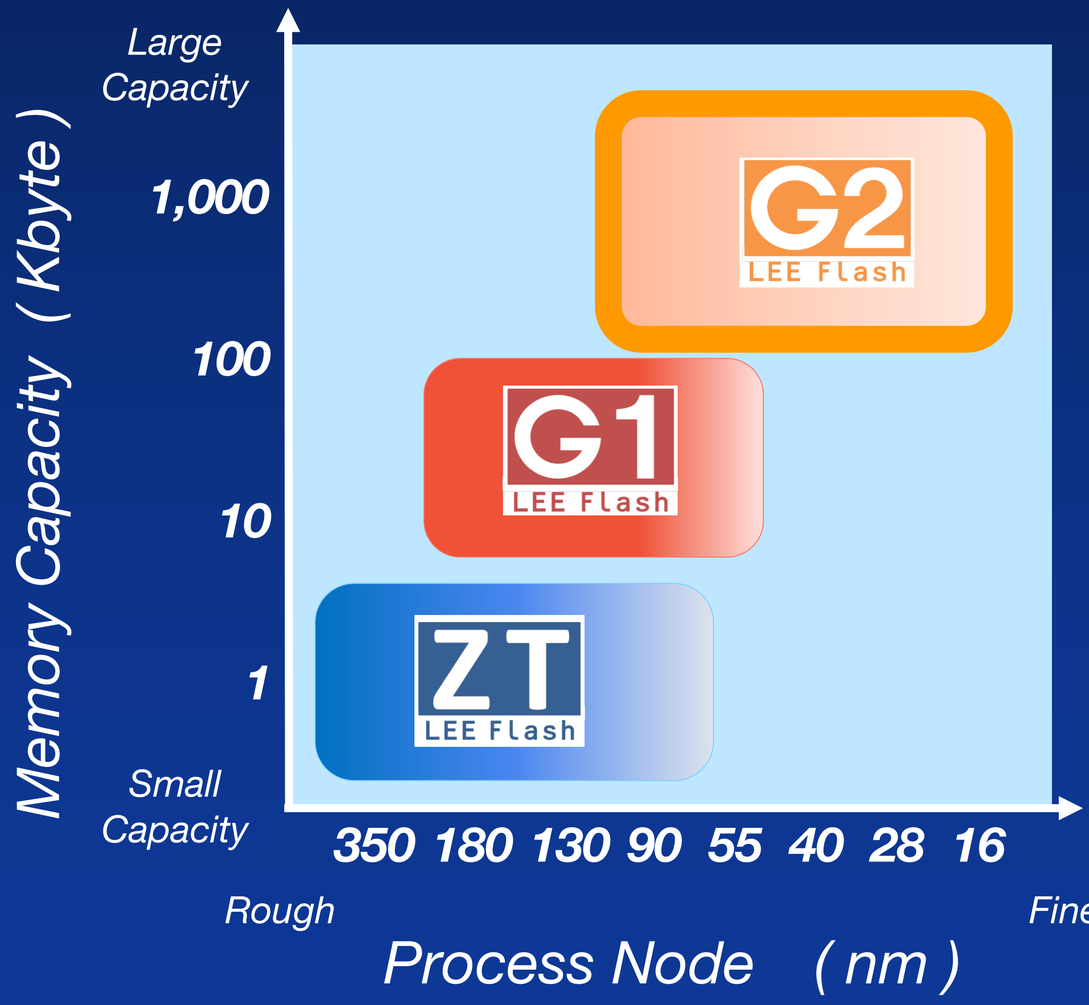
Computing In Memory with Tri-gate SONOS Nonvolatile Multi-level Memory

Y.Taniguchi, S.Yoshida, T.Tamatsu, M.Hishiki, Y.Sasaki,
Y.Kawashima, F.Owada, K.Yanagisawa, K.Okuyama

Floadia Corporation, Tokyo, Japan

IP Portfolio

Floadia offers three kinds of eFlash IP [ZT · G1 · G2]



G2 LEE Flash

- *In development.*
- *Capable of large capacity up to several MB*
- *Suitable for a wide range of devices including automotive LSI, IoT terminal,*

AI (Computing in Memory)

G1 LEE Flash

- *Can be embedded at 1/3 low cost, comparing to competitors, and is being adopted as a medium-range capacity Flash for Smart Phones and Automotive LSI*

ZT LEE Flash

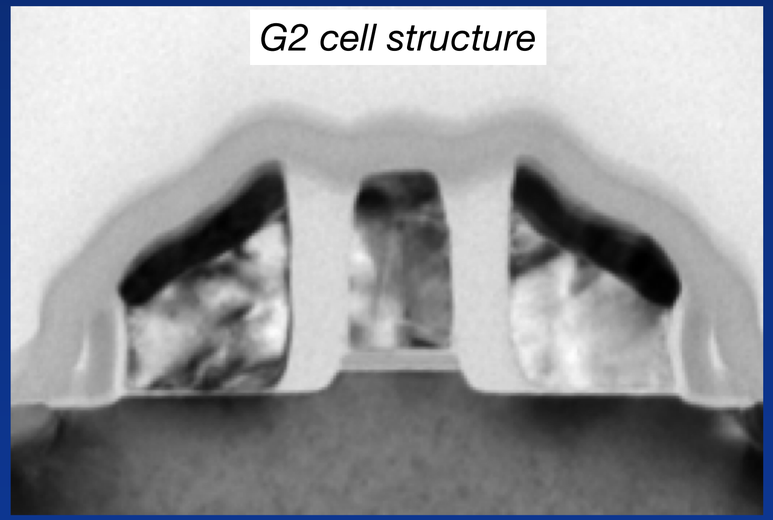
- *Can be embedded without additional mask process.*
- *Has High reliability as the feature, and is widely adopted for Automotive and Industrial applications*

* *All IPs are covered by own patents*

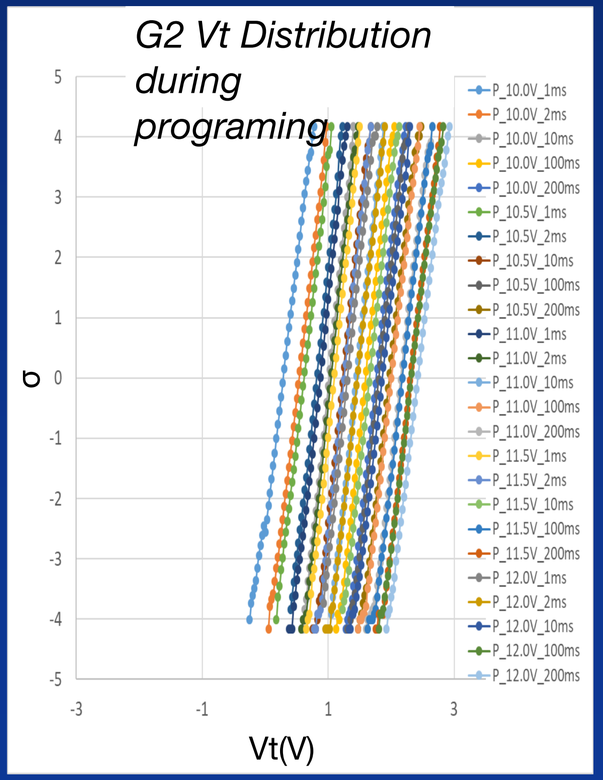
Unique G2 characteristics make CiM possible

- *Tri-Gate Structure* > *Sneak current free architecture*
- *Good Uniformity (no tail bit)* > *Precise program control 64~128 levels*
- *Wide Ion / Ioff ratio (7 digits)* > *1k~10k data in/out memory array*
- *Robust P/E endurance* > *Repeatability of setting weight*

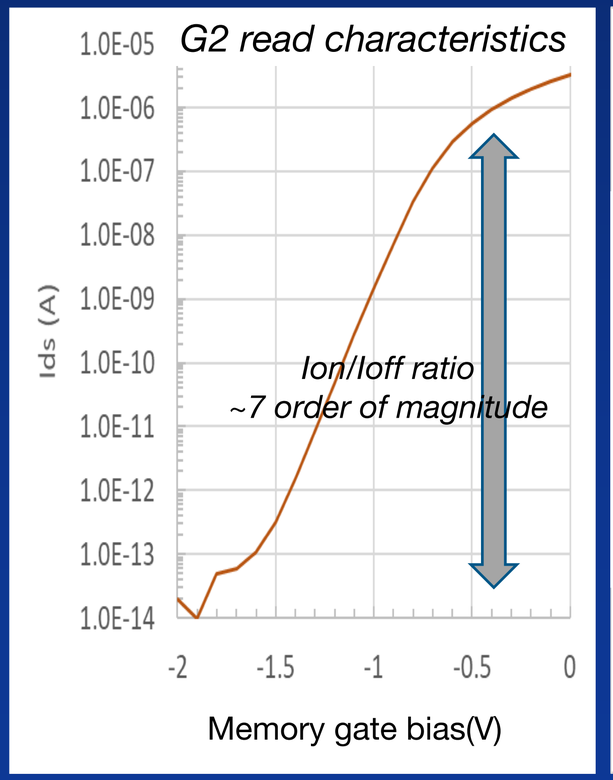
G2 cell structure



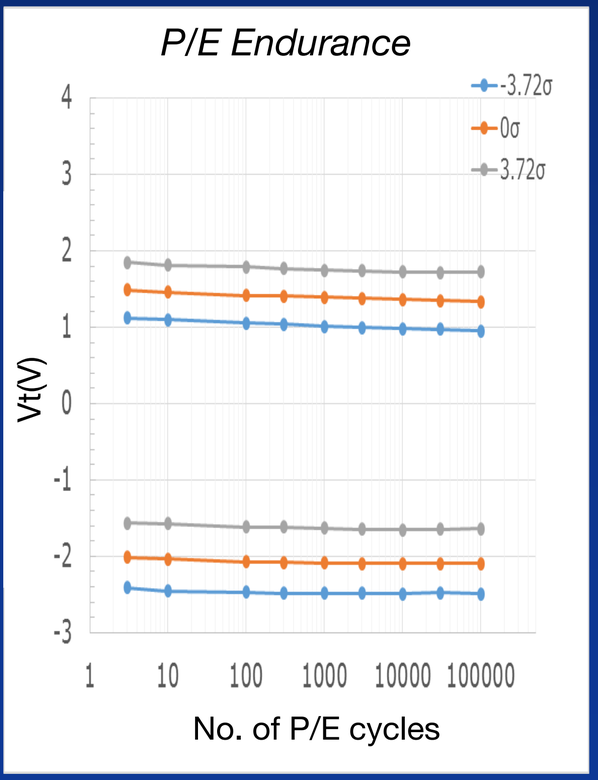
G2 Vt Distribution during programing



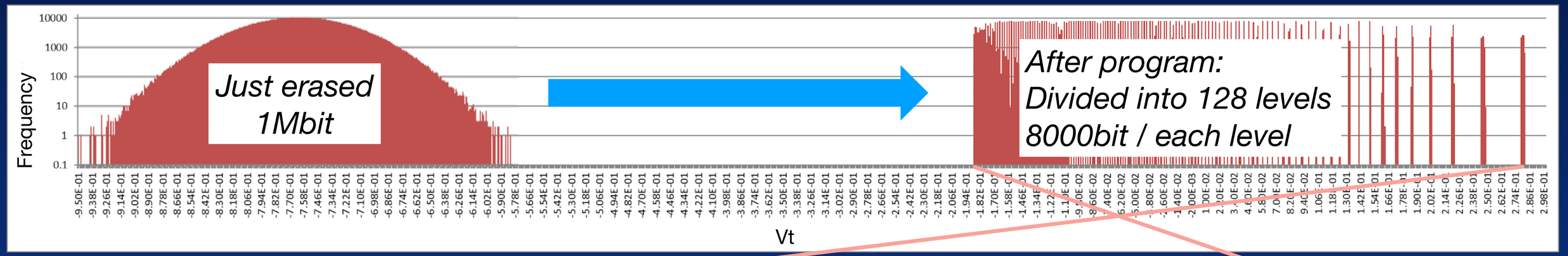
G2 read characteristics



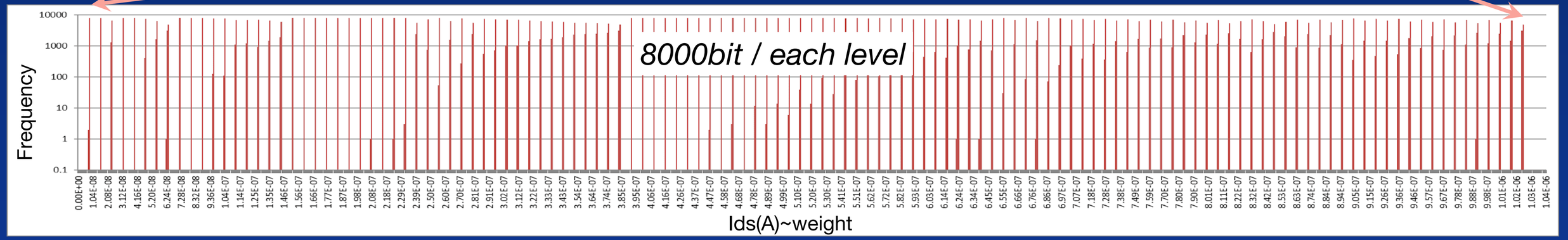
P/E Endurance



Accuracy estimation of setting weight (1M bits simulation)

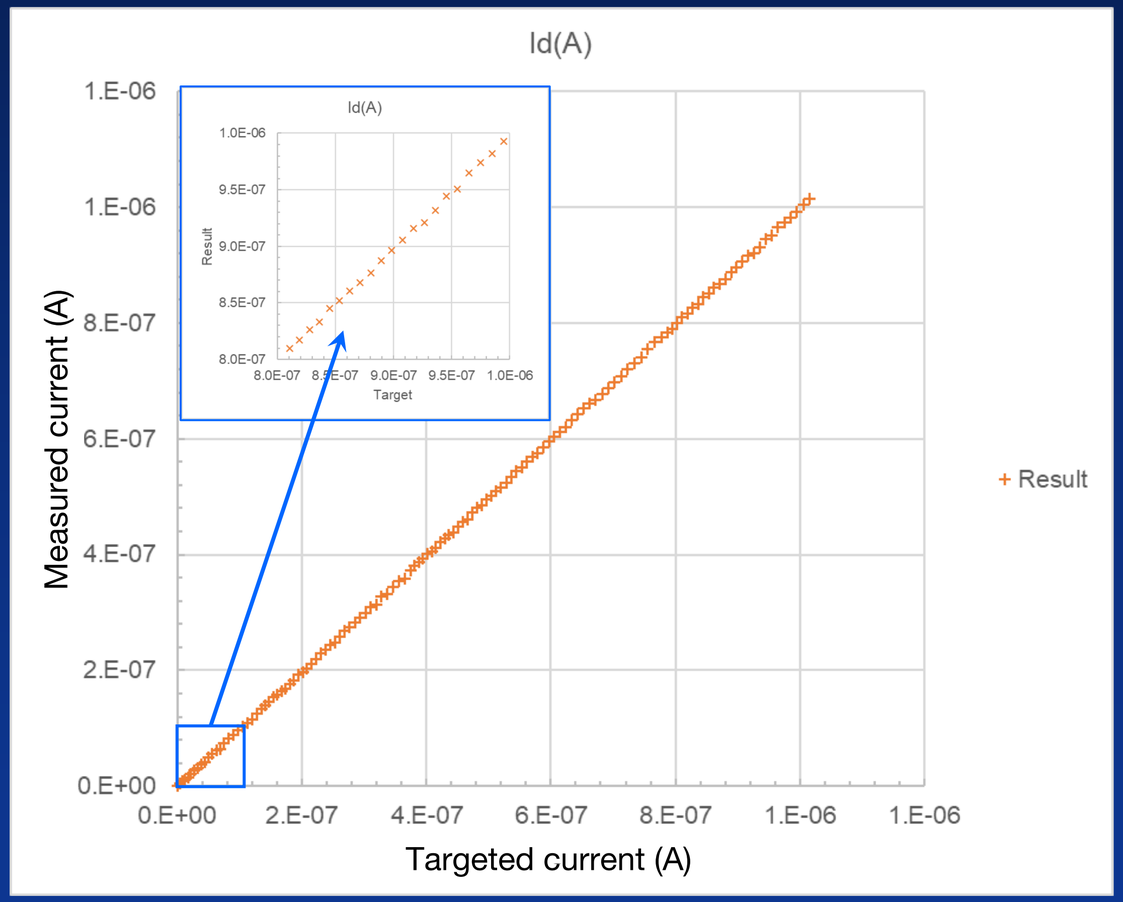


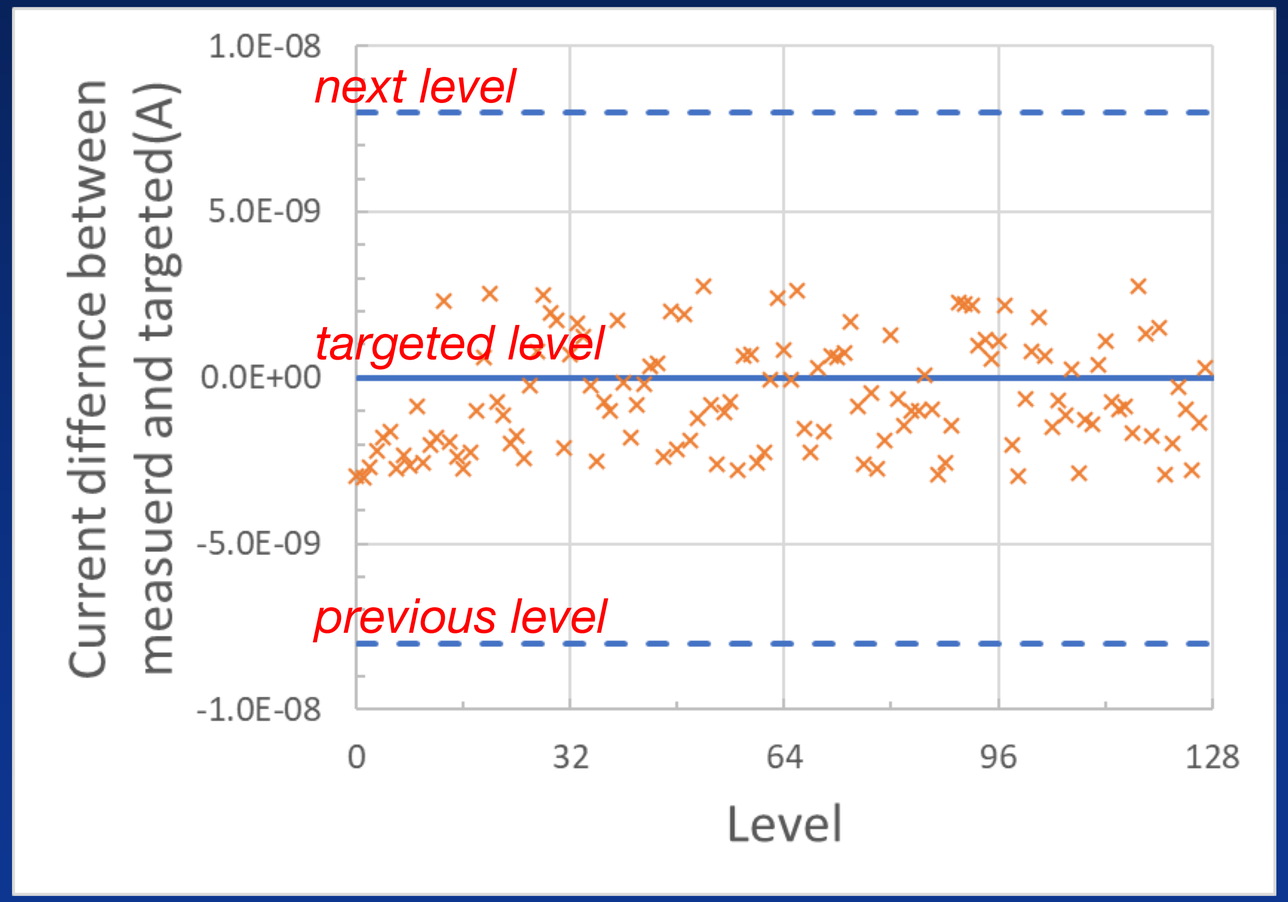
Conversion : Vt > Ids(weight)



- Experiment : 1 Mbit (1024 bits / output × 1000 output)
- Target : Set bits equally at all levels 128 level, 8000 bit / each level
- Vti variation : 40 mV / σ
- Target accuracy and result : Variance within the level is less than 10% to the adjacent level
  → all bits succeeded (0.1 LSB@128levels)

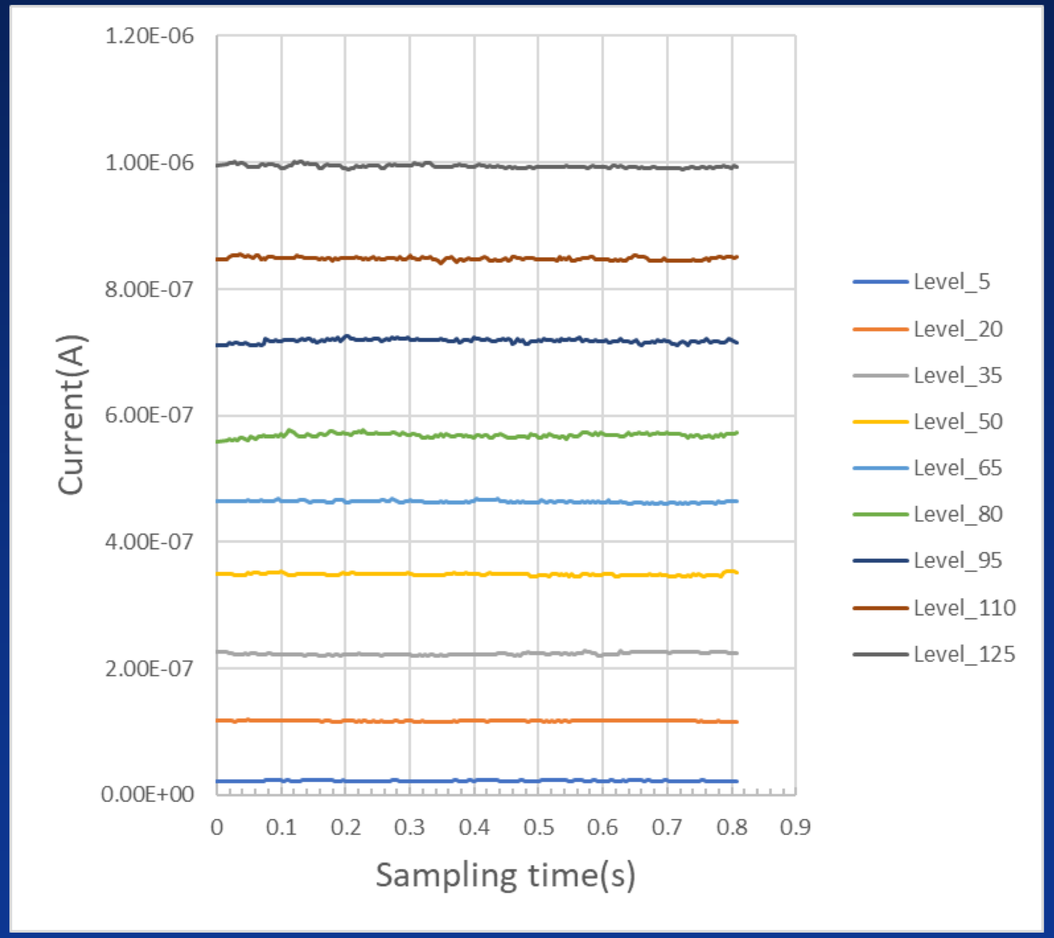
Accuracy evaluation of Weight set (single bit)

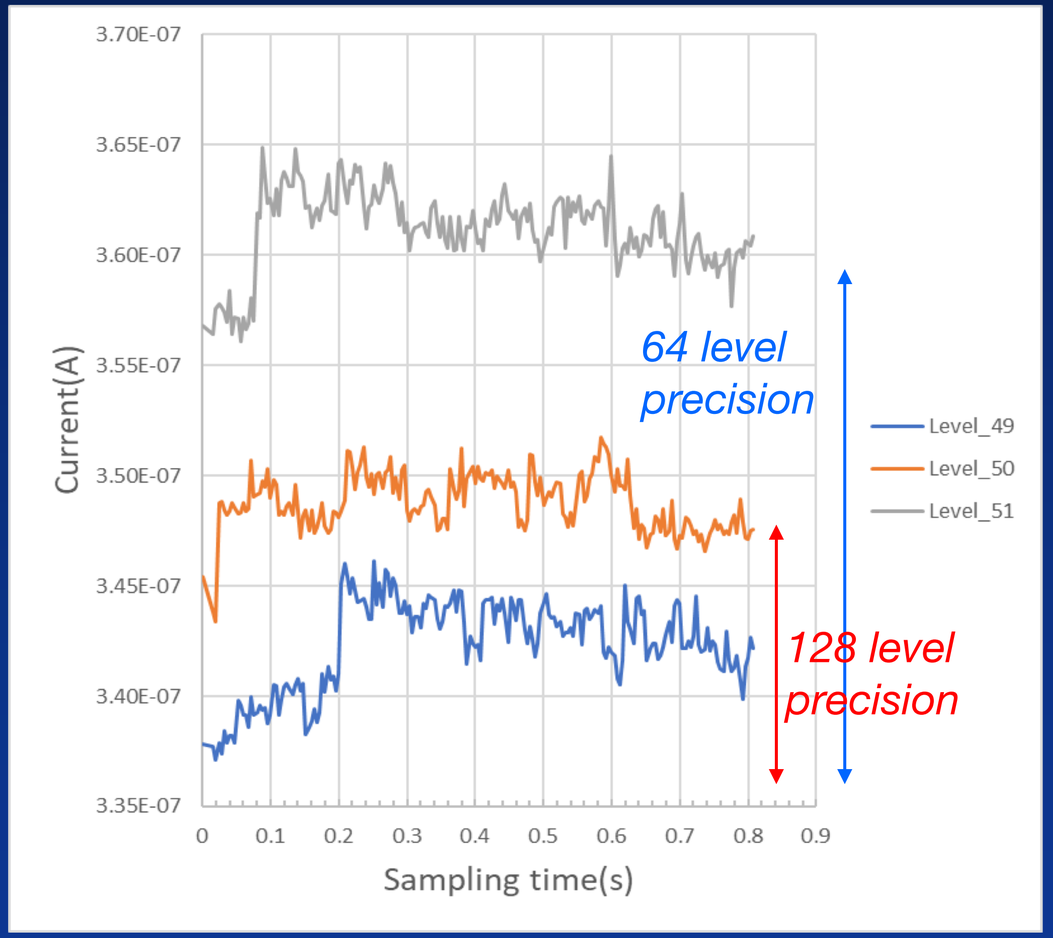




128 levels precision capability has been confirmed.

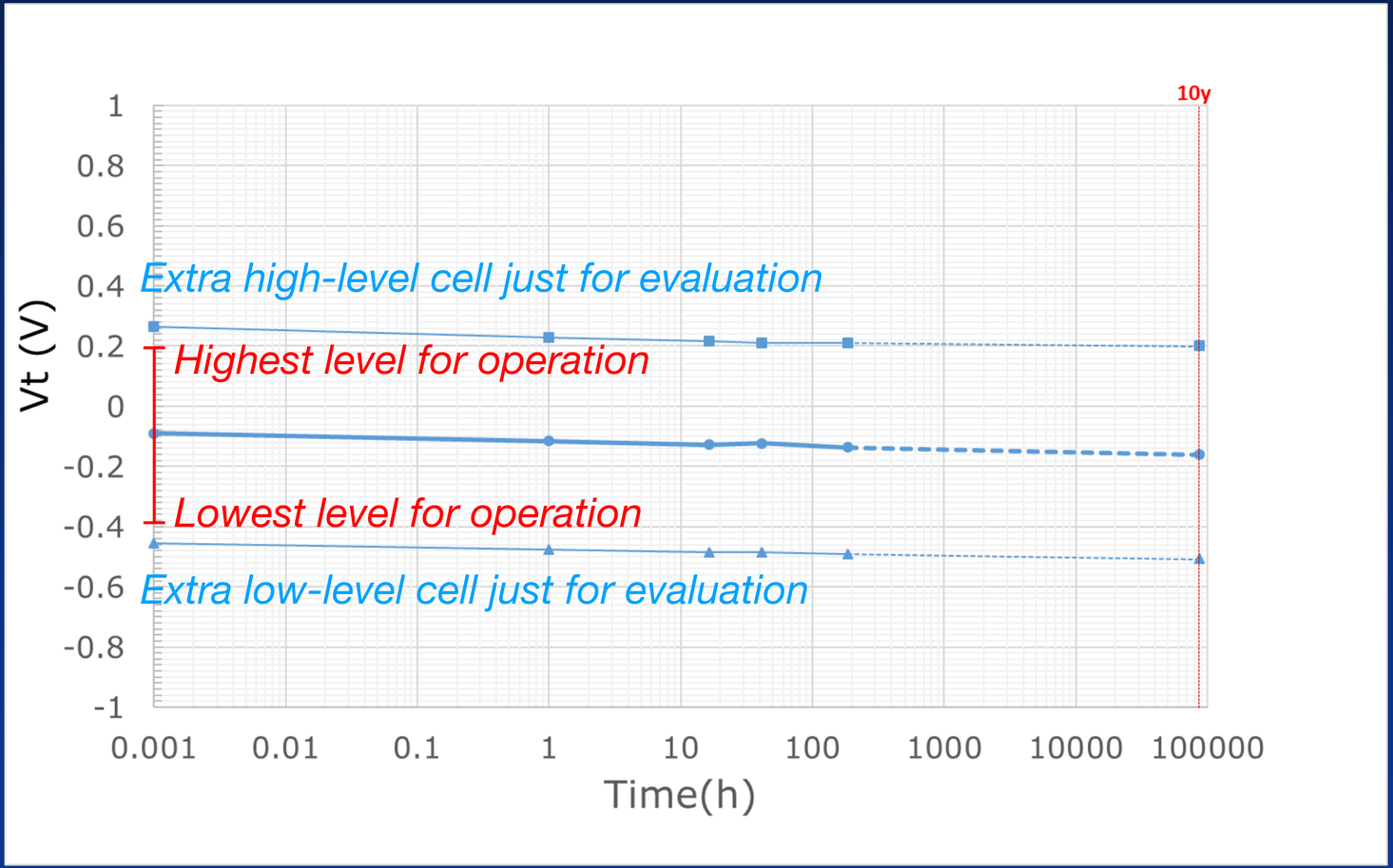
Read operation stability (single bit)

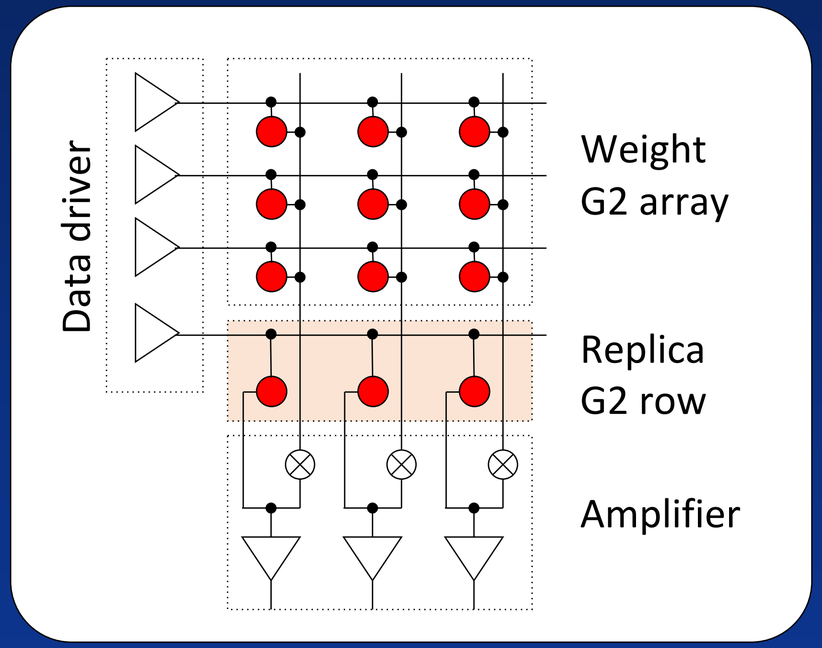




More than 64 levels precision capability has been confirmed.
Improvement to 128 levels is in progress.

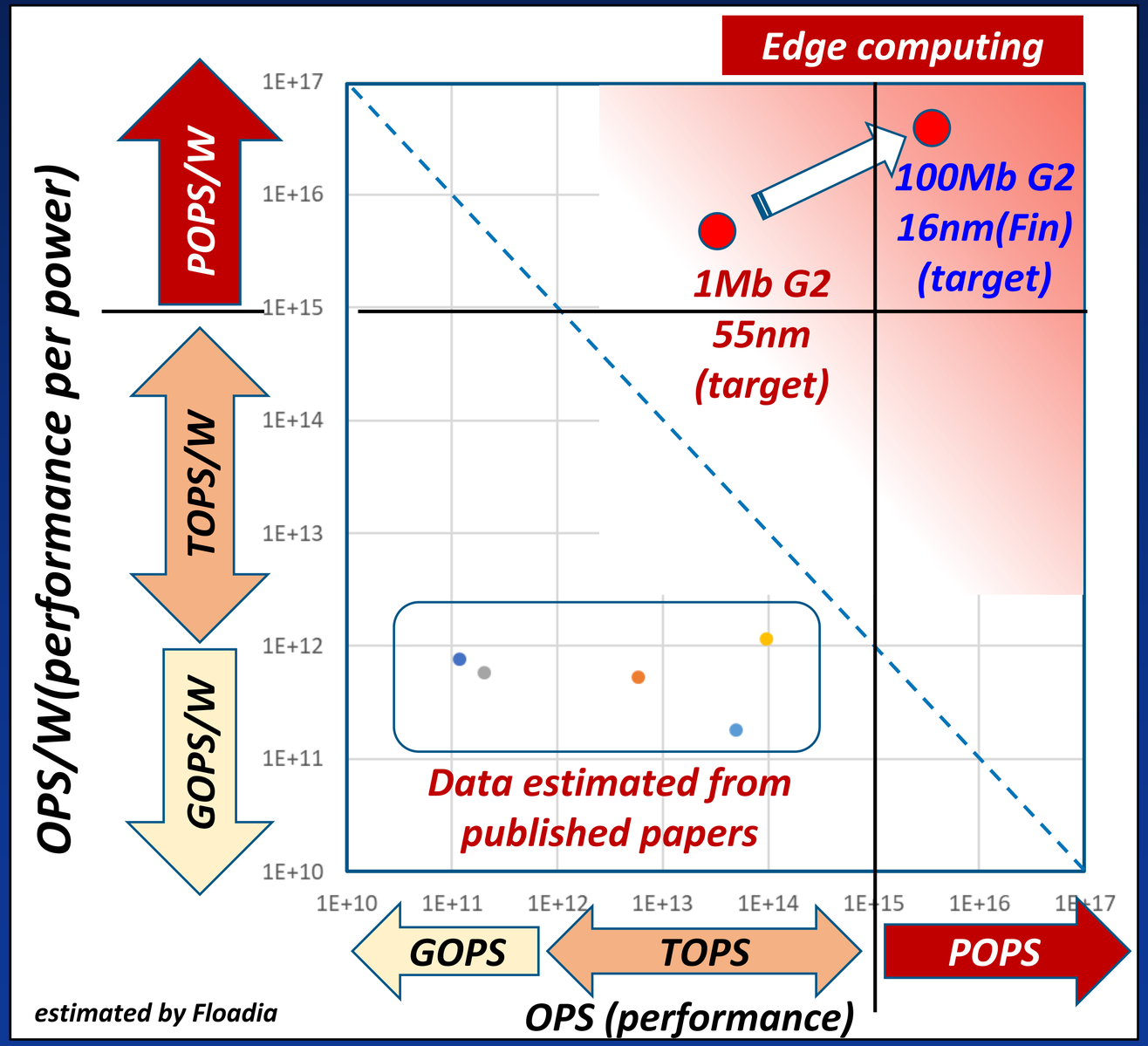
Evaluation of data retention stability





The amounts of Vt shift during storage are uniform regardless of programmed level.

Comparison of Performance per Power



*GOPS: Giga-Operation Per Second
TOPS: Tera-Operation Per second
POPS: Peta-Operation Per Second

Summary & technology comparison

	ReRAM	STT-MRAM	Floadia's G2
Max read current	~1uA	~10uA	<1uA
Ion/Ioff ratio	10~10^3	2	10^7
No. of level	~8	2	64~128
Sneak current	Need to care		Free
Precision control	0.5LSB@4levels : at present	Highest but for only 2 level	0.1LSB@128levels:simulation 0.5LSB@64:Si evaluation (just started)
Resistance change mechanism	Physical structure change	Spin direction (stable)	SiN film capture/release electron (stable)
Status	R&D		• G2 cell has been available • No restriction on migration down to FIN process

Thank you for your attention

For further information, please visit the web site

https://www.floadia.com/

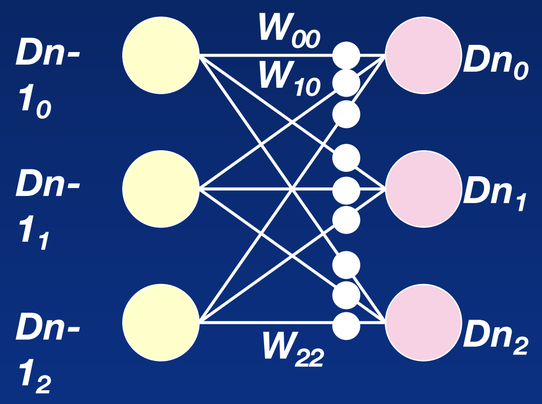
Floadia

A world class expert providing embedded Non-Volatile Technology
applied to various devices such as IoT terminals and Artificial Intelligence.

Headquarters : 1-30-9, Ogawa Higashi-cho, Kodaira-shi, Tokyo 187-0031, Japan
Taiwan Branch : R612/R618, Center of Innovative Incubator, NTHU, Hsinchu, Taiwan 30013, ROC

No Matrix Operation required with Computing in Memory

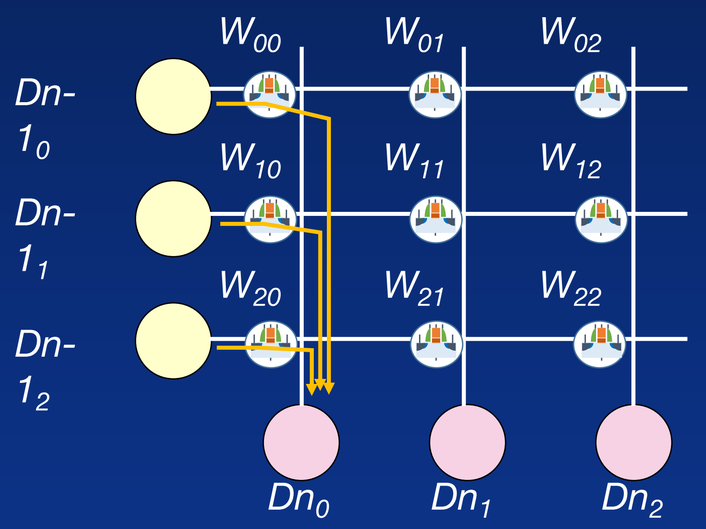
Deep Learning NN Calculation



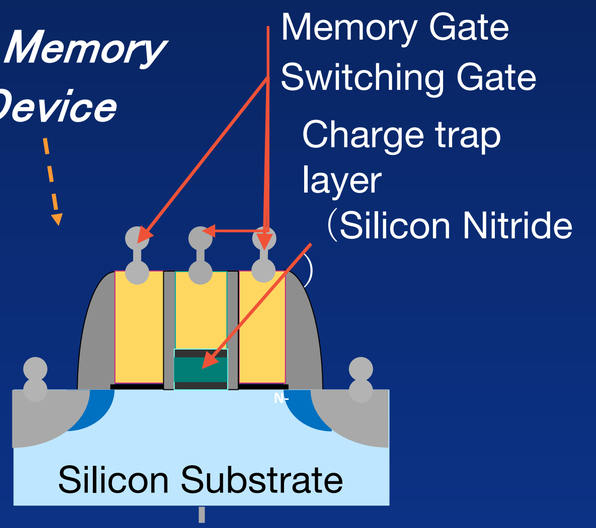
$$Dn\ i+1 = \sum_j (Wi\ j \times Dn\text{-}1\ j)$$

Calculation Method : Product & Sum
Number of Calculation :
【 3 Inputs x 3 Outputs 】
9 times product + 9 times sum
【 1,000 Inputs x 1,000 Outputs 】
1,000,000 times product
+ 1,000,000 times sum

G2 for Computing in Memory



Calculation of the product is done by each input current flow into the G2 memory.



Cross Section of G2 device

Calculation Method : Read-out
Number of Read-out : One time
【 3 Inputs x 3 Outputs 】
Simultaneous Read-out of 3 Outputs lines
【 1,000 Inputs x 1,000 Outputs 】
Simultaneous Read-out of 1,000 Outputs lines