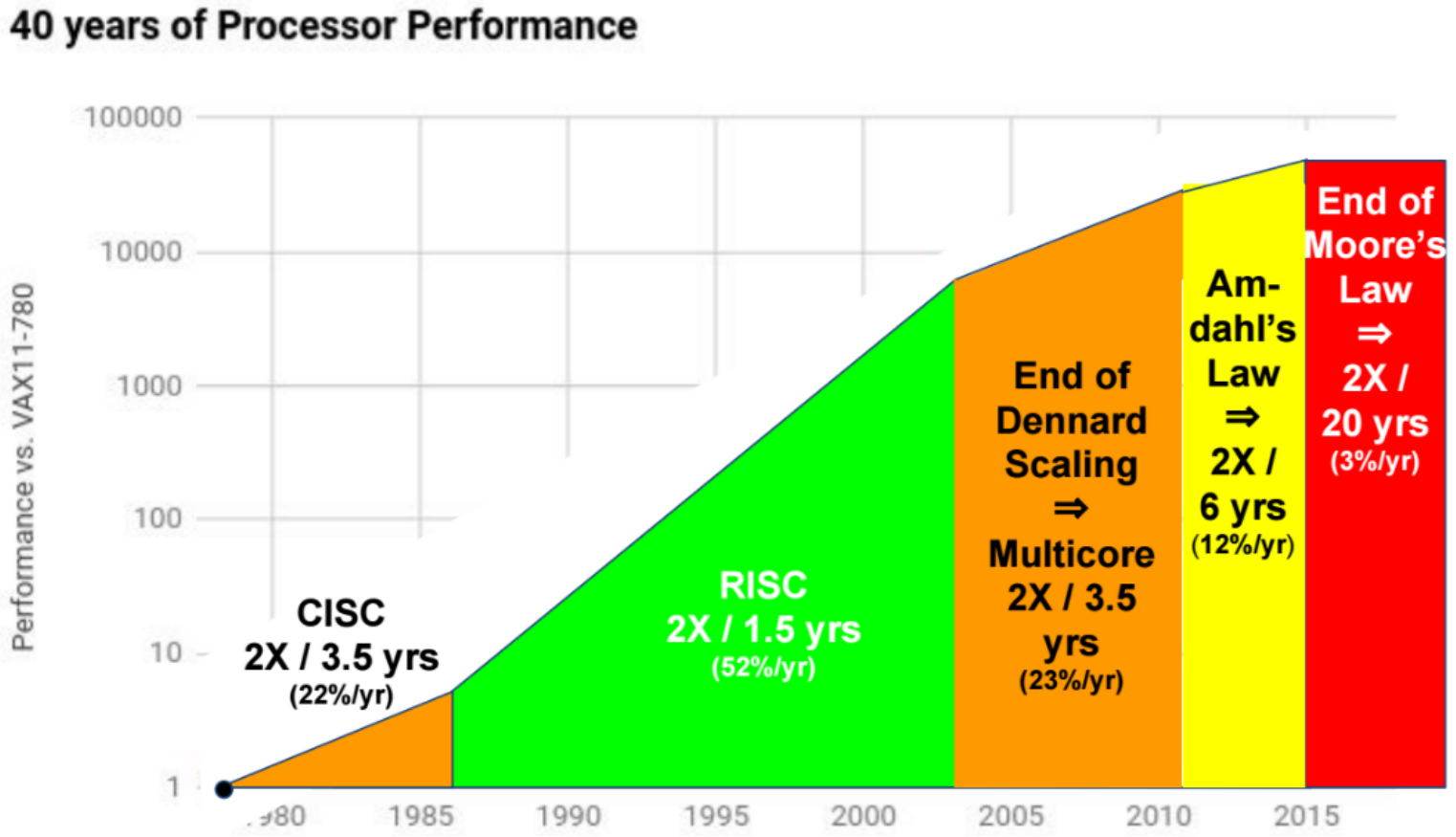
Reliability enhancement of phase change memory for neuromorphic applications

SangBum Kim, Ph.D.
Research Staff Member
Novel memory and cognitive applications
IBM T.J. Watson Research Center

SangBum.Kim@us.ibm.com

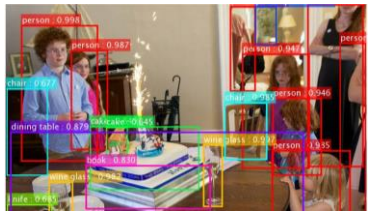
End of Moore's Law

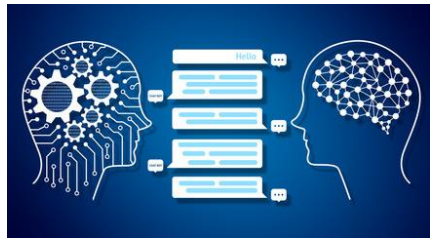


D. Patterson, NAE Regional Meeting, 2017.

- Moore's Law is ending : No more processor performance growth.
- "*More-than Moore*" solution : Build a new processor which does *a few tasks extremely well*. → Which task should we choose to accelerate?

Recent Breakthroughs in Artificial Intelligence

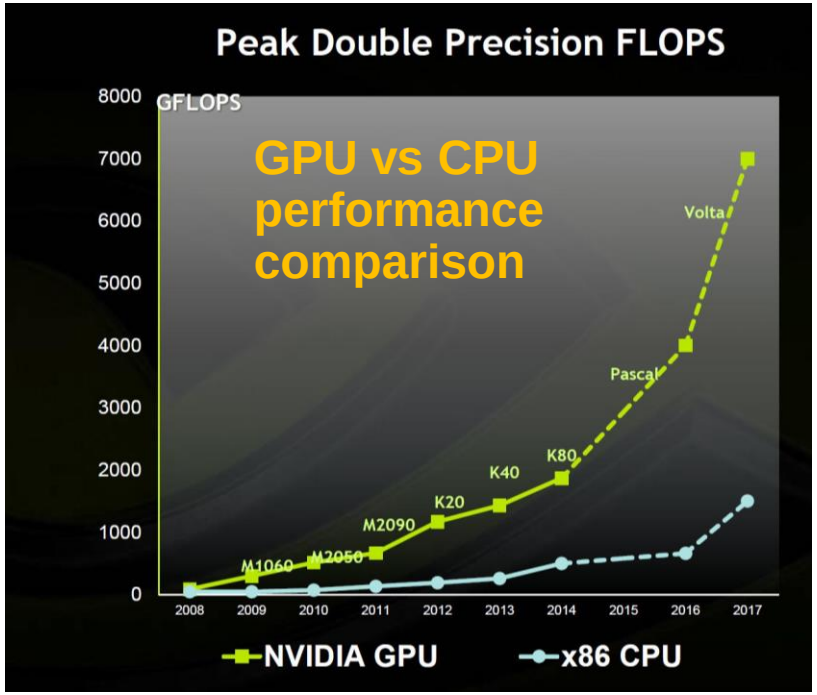


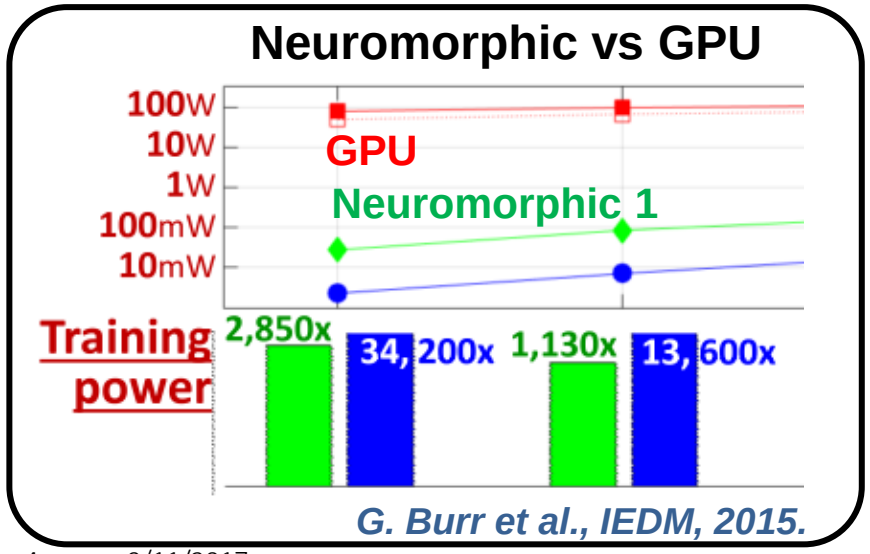


- Artificial Intelligence became so important and companies are developing *a dedicated processors for artificial intelligence*.
  - Google: TPU, Intel: Nervana Engine, Qualcomm, Apple

CADE METZ BUSINESS 10.28.16 7:00 AM
HOW AI IS SHAKING UP THE CHIP MARKET
Wired Magazine, 2016

Processors dedicated for artificial intelligence





performance/watt for artificial intelligence (neural network) computation

CPU
Digital
Now-5 years

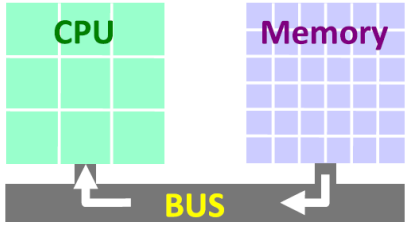
10-100x

GPU, ASIC, FPGA
Digital

10-1000x

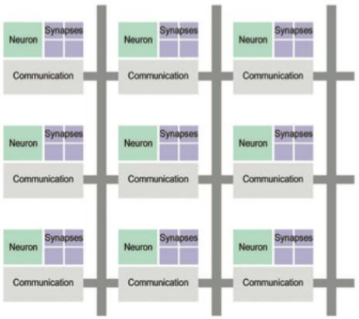
Neuromorphic
Analog
5-10 years

Von-Neumann



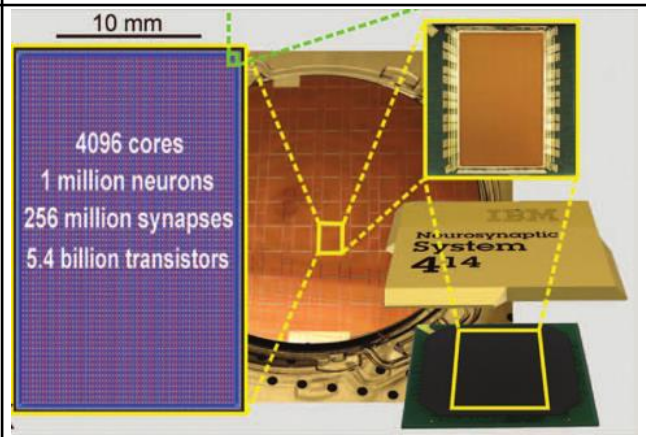
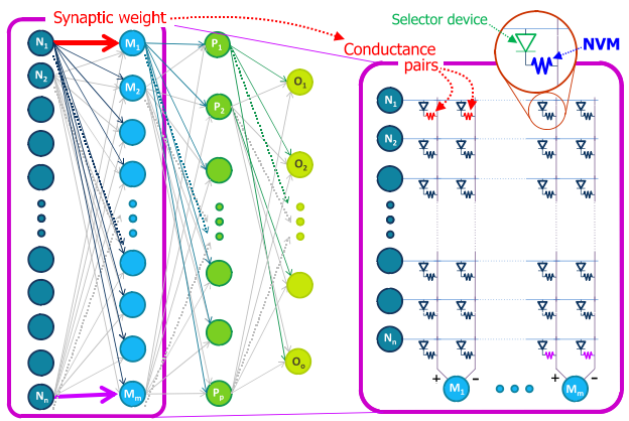
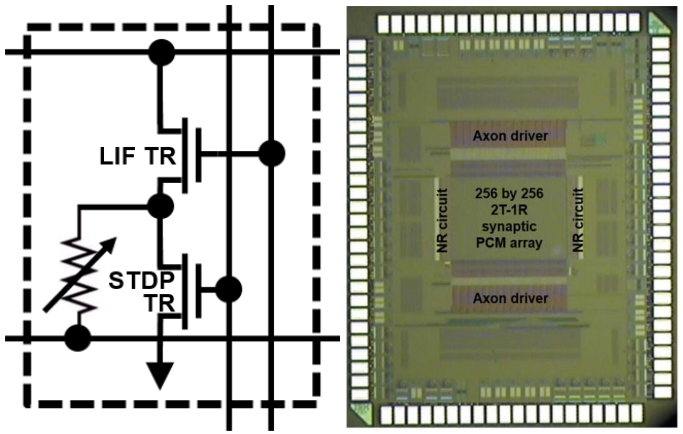
Von-Neumann memory bottleneck

Non Von-Neumann

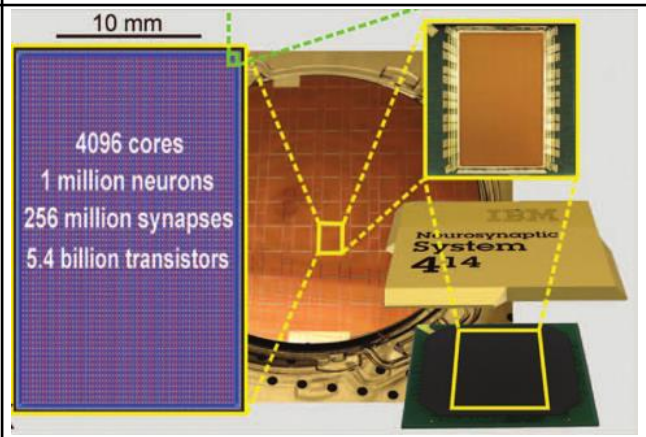
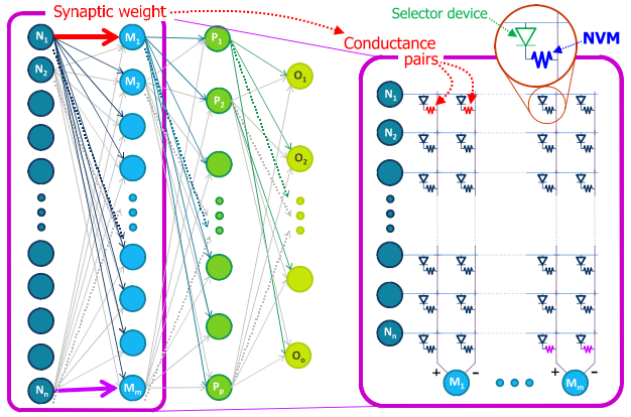
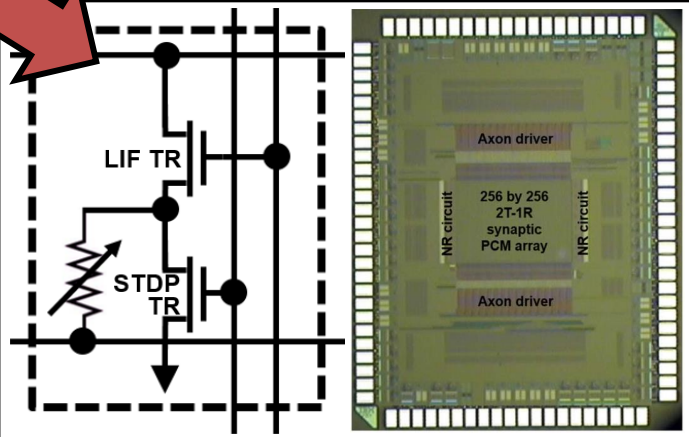


converged logic+memory

Neuromorphic processor development in IBM

	Artificial neural network (ANN)	Spiking neural network (SNN)
Conventional digital memory device		 P. Merolla et al., Science, 2014. ❑ TrueNorth ○ SRAM ○ Off-chip learning
Analog non-volatile memory (NVM) device	 G. Burr et al., IEDM, 2014. ❑ NVM ML accelerator ○ PCM ○ On-chip learning	 S. Kim et al., IEDM, 2015. ❑ NVM low-power ML ○ PCM ○ On-chip learning

Neuromorphic processor development in IBM

	Artificial neural network (ANN)	Spiking neural network (SNN)
Conventional digital memory device	**Today's main topic: NVM + SNN**	 *P. Merolla et al., Science, 2014.* ❑ TrueNorth ○ SRAM ○ Off-chip learning
Analog non-volatile memory (NVM) device	 *G. Burr et al., IEDM, 2014.* ❑ NVM ML accelerator ○ PCM ○ On-chip learning	 *S. Kim et al., IEDM, 2015.* ❑ NVM low-power ML ○ PCM ○ On-chip learning

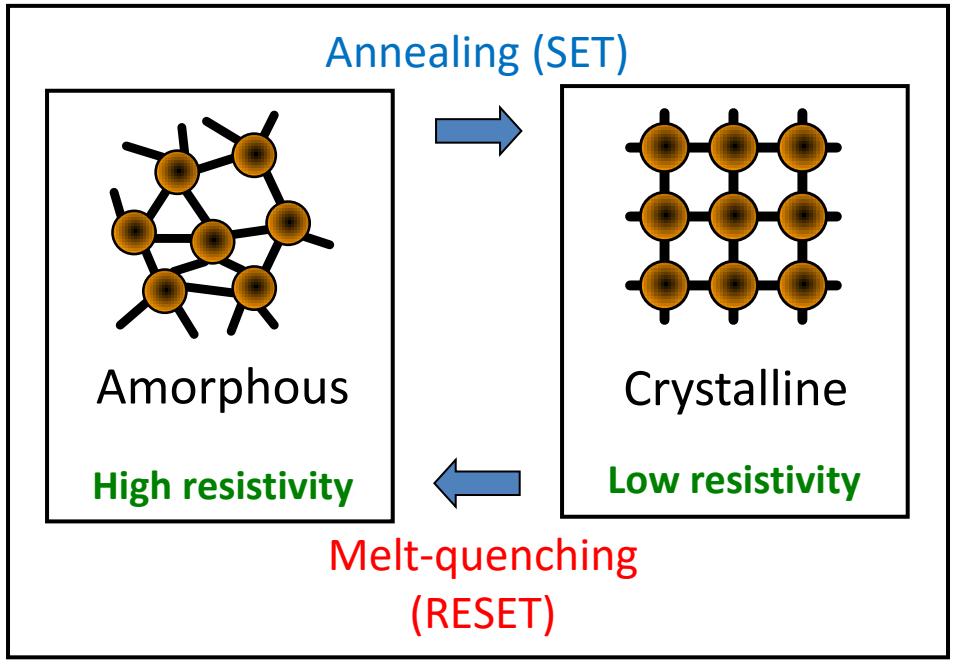
Mismatch between algorithms and devices

- All the *scalable learning algorithms* demonstrated so far requires *near-perfect devices* to match the state-of-the-art clarification accuracy and power & area efficiency.

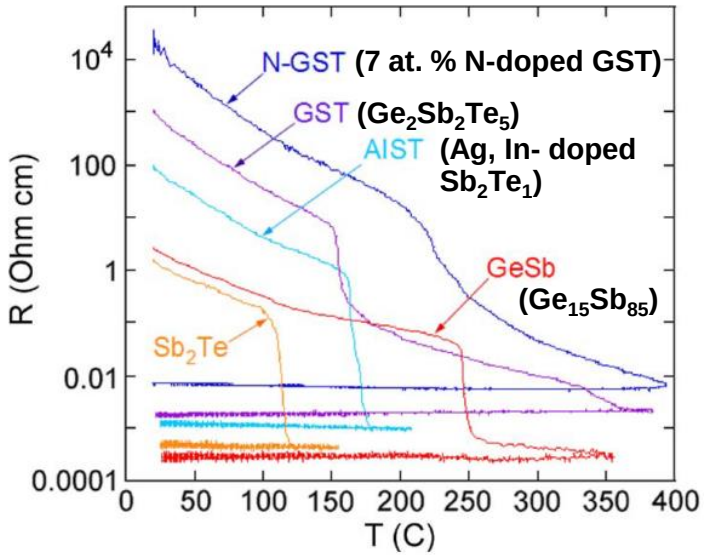
  - How to improve synaptic device characteristics?

  - Can we modify learning algorithms such that they can work with imperfect synaptic devices?

Phase change memory (PCM)

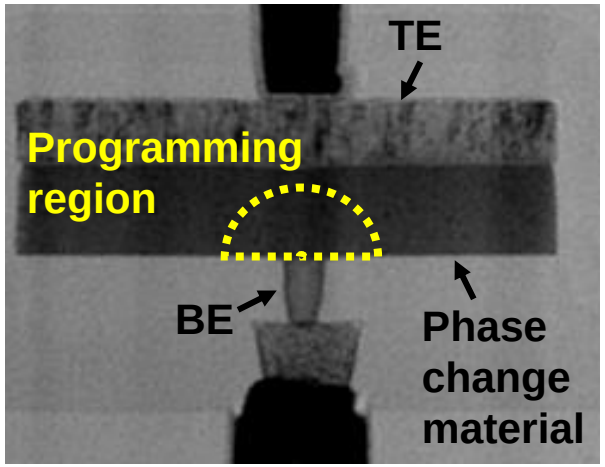


Various phase change material



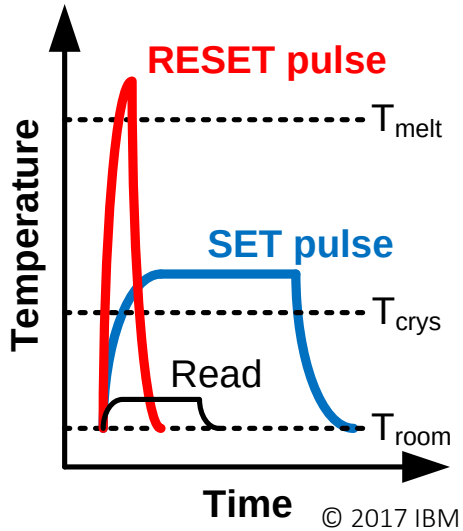
P. Wong et al., IEEE Proc., 2010

Typical cell structure

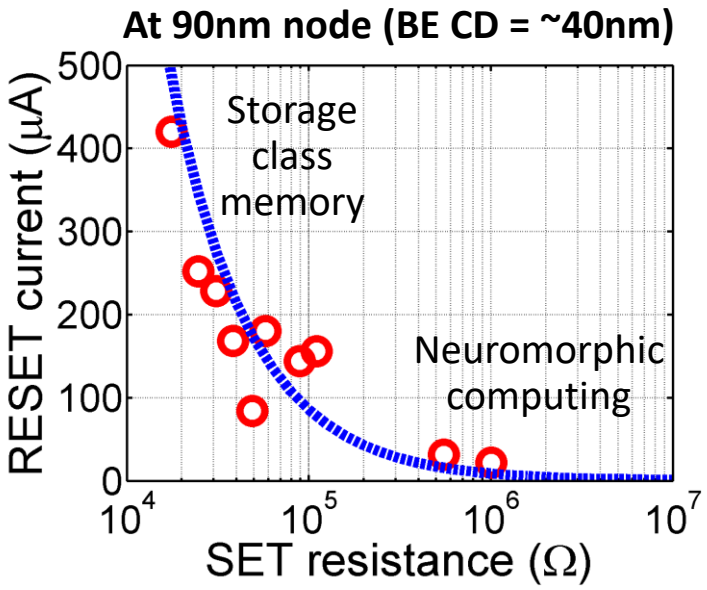


- Memory states : Amorphous and poly-crystalline state
- Program mechanism : Joule heating
- Read mechanism : Resistance

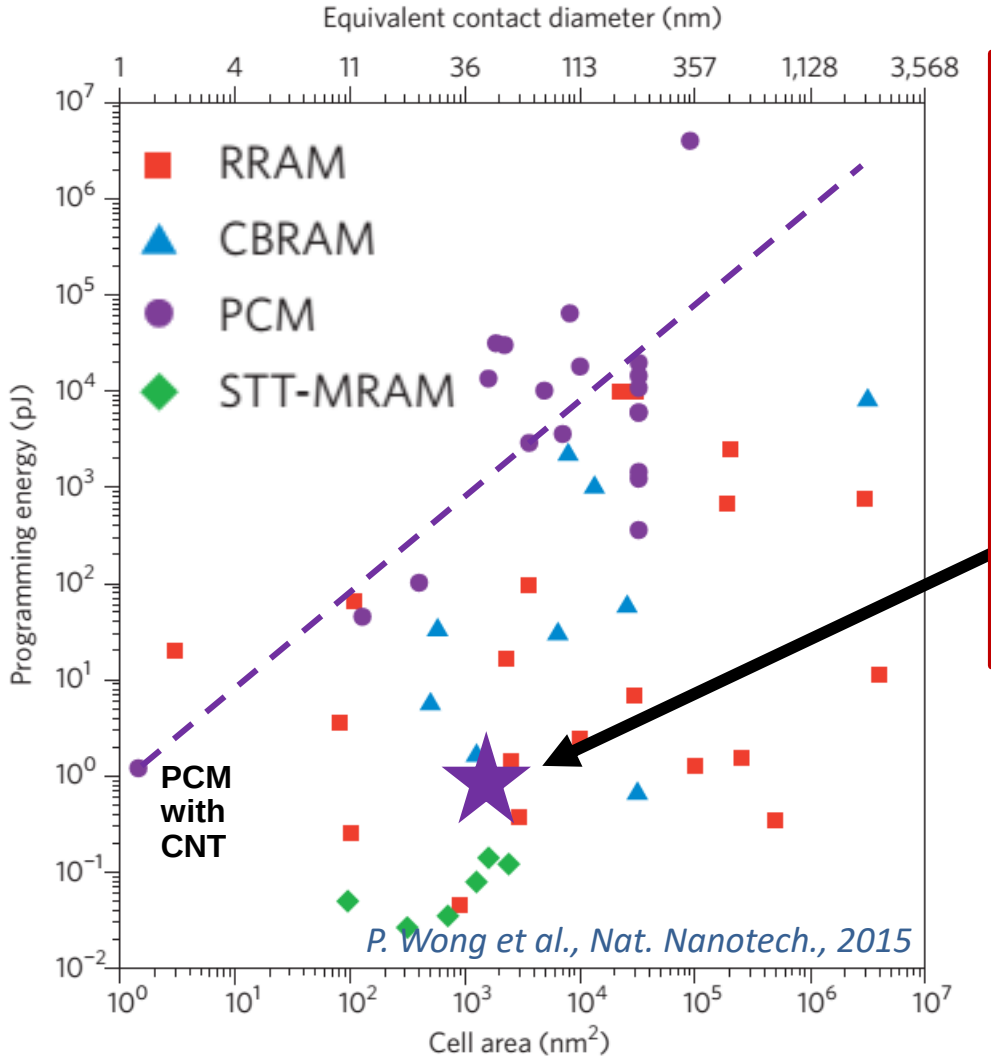
Program and Read

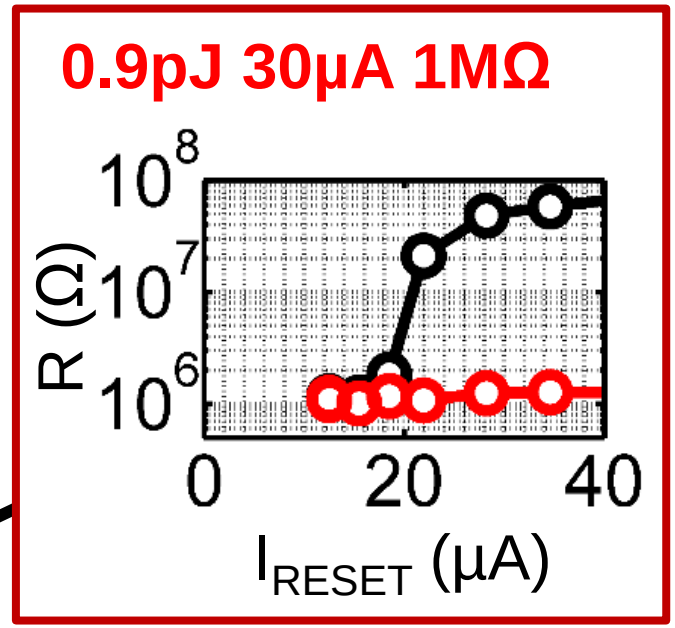


Power efficient PCM cell for neuromorphic applications



Neuromorphic computing
- Large SET R is OK.
- Programming energy can be further scaled down.





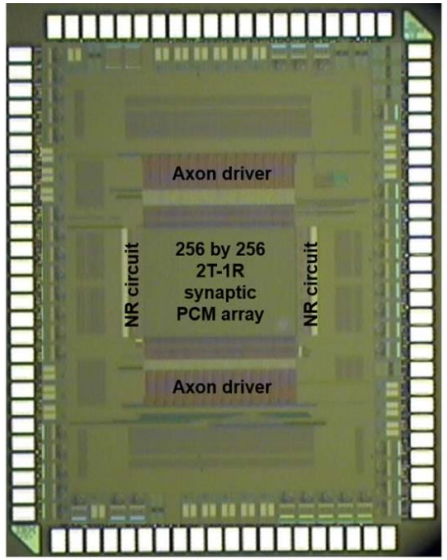
S. Kim et al., IEDM, 2015

We significantly lowered the RESET programming current by heavily doping GST to increase the resistivity.

❑ *PCM power consumption* can be further optimized for neuromorphic applications.

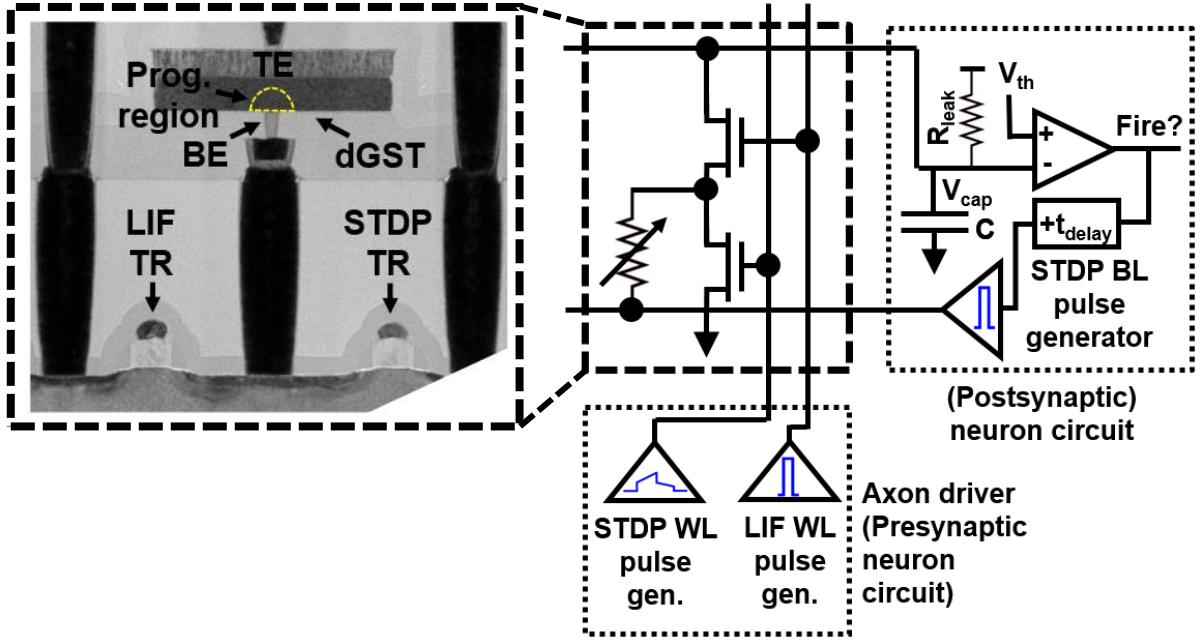
PCM neuromorphic core

PCM neuromorphic core *S. Kim et al., IEDM 2015*



Synapse	NVM PCM
	Analog
Neuron / chip	256
Synapse / chip	64k
Technology node	90nm
Unit cell	2T-1R
Neural network type	Spiking
Learning	On-chip
	Continuous
	STDP mechanism

2T-1R synaptic unit cell

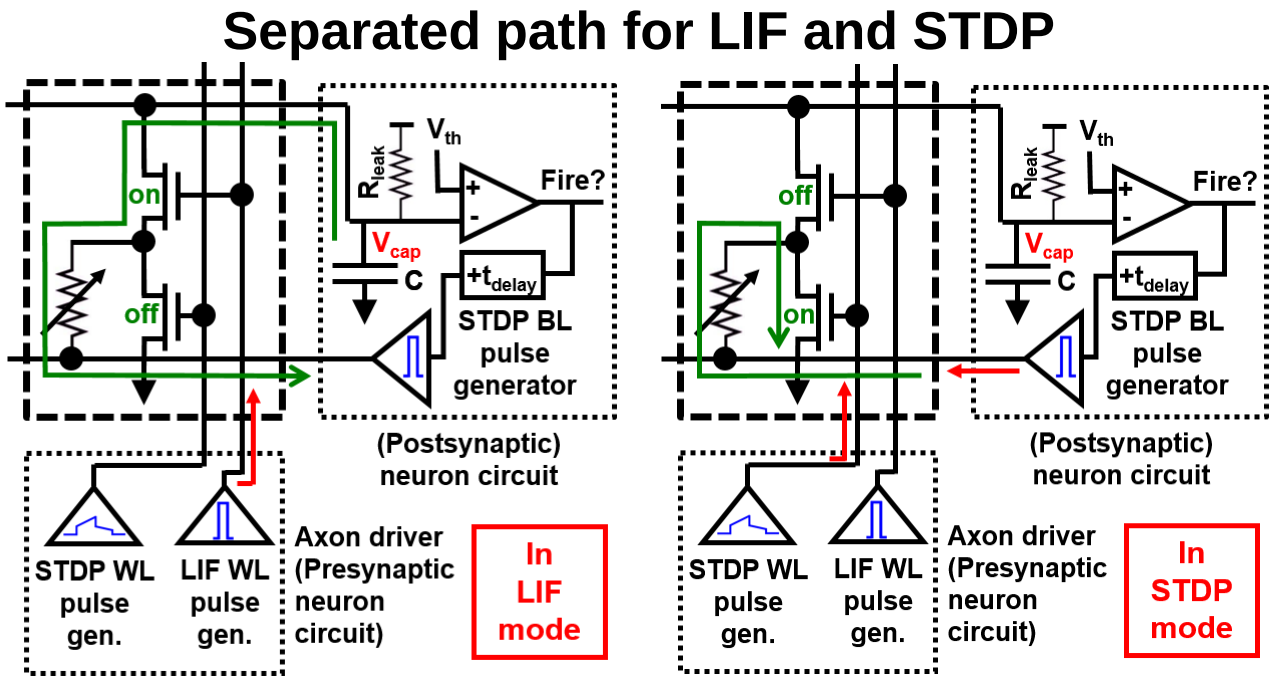


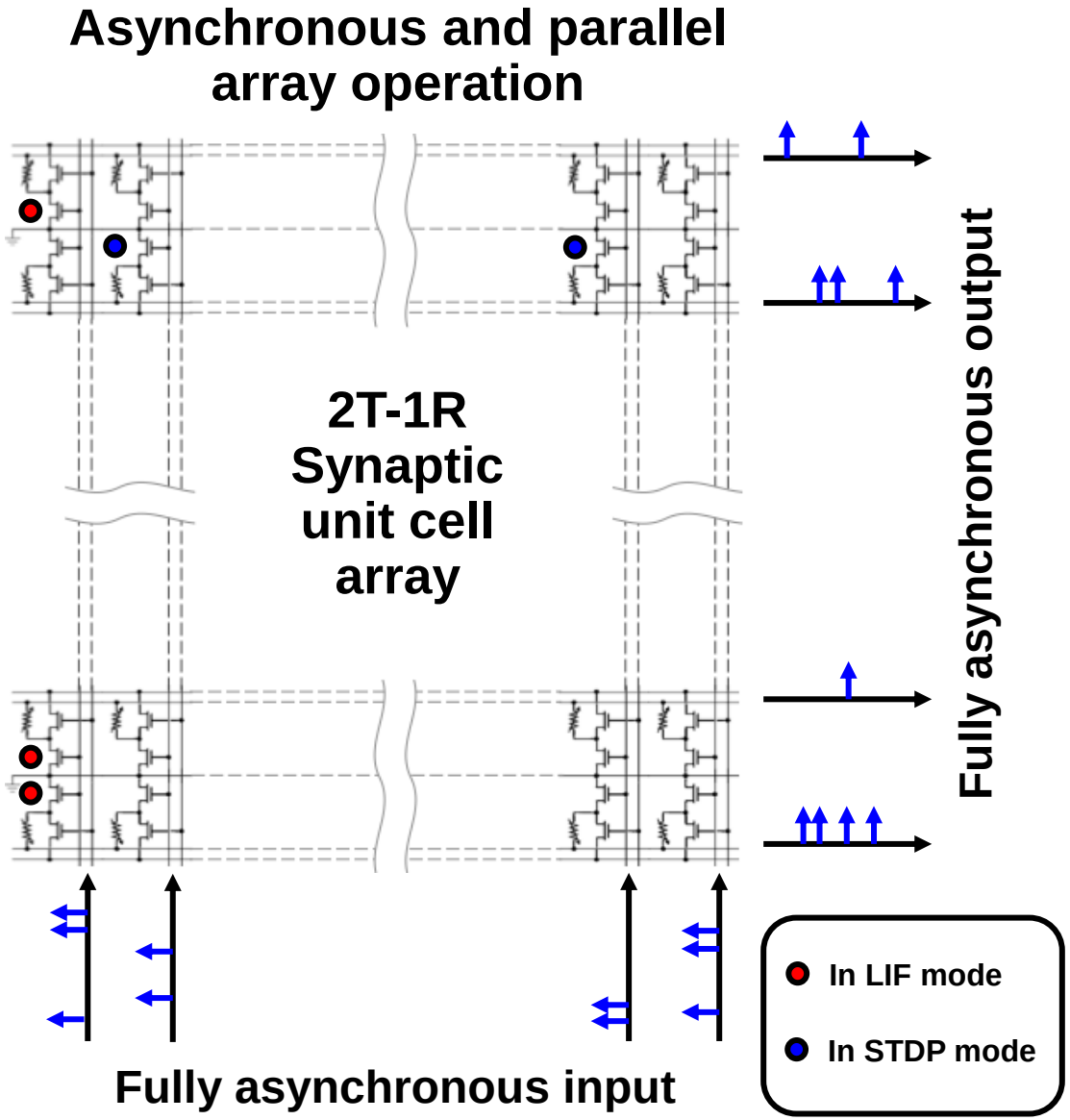
- *World-first* demonstration of sizable spiking neural network using NVM
  - PCM 2T-1R synaptic operation
  - *On-chip spiking and learning neuron circuit*
  - Fully asynchronous and parallel operation

- Invented *2T-1R synaptic unit cell* to enable
  - *massively parallel* operation
  - *asynchronous* LIF and STDP operation

- *Area-efficient*
  - [2T-1R / 5-10 bits] vs [SRAM: 6T / 1 bit]

Fully asynchronous and parallel operation

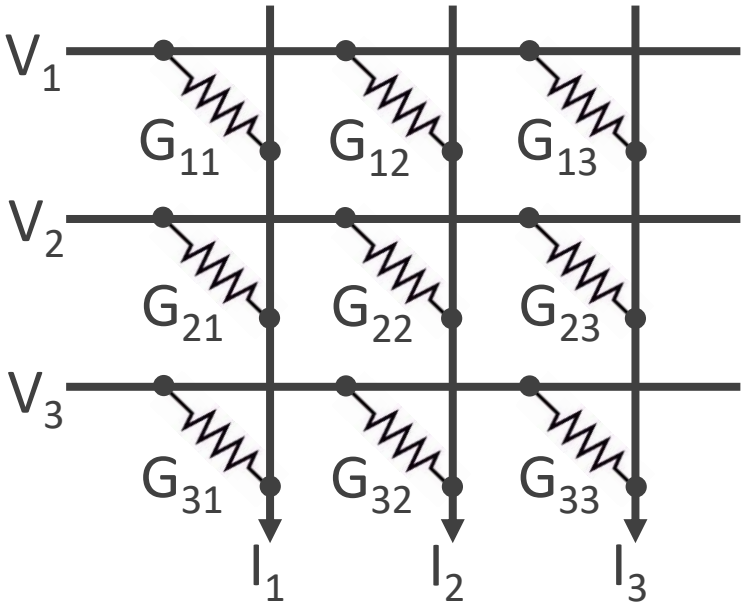




- Fully asynchronous and parallel operation enabled by
  - *Low programming power* of PCM cell
  - *Separated path* for LIF and STDP using *2T-1R* unit cell

NVM's main role in neuromorphic processor

Resistor network



Matrix-vector multiplication using Kirchhoff's law

$$\begin{bmatrix} I_1 \\ I_2 \\ I_3 \end{bmatrix} = \begin{bmatrix} G_{11}V_1 + G_{21}V_2 + G_{31}V_3 \\ G_{12}V_1 + G_{22}V_2 + G_{32}V_3 \\ G_{13}V_1 + G_{23}V_2 + G_{33}V_3 \end{bmatrix}$$

Multiply-and-accumulate

$$\int I_k dt = \int (G_{1k}V_1(t) + G_{2k}V_k(t) + G_{3k}V_k(t))dt$$

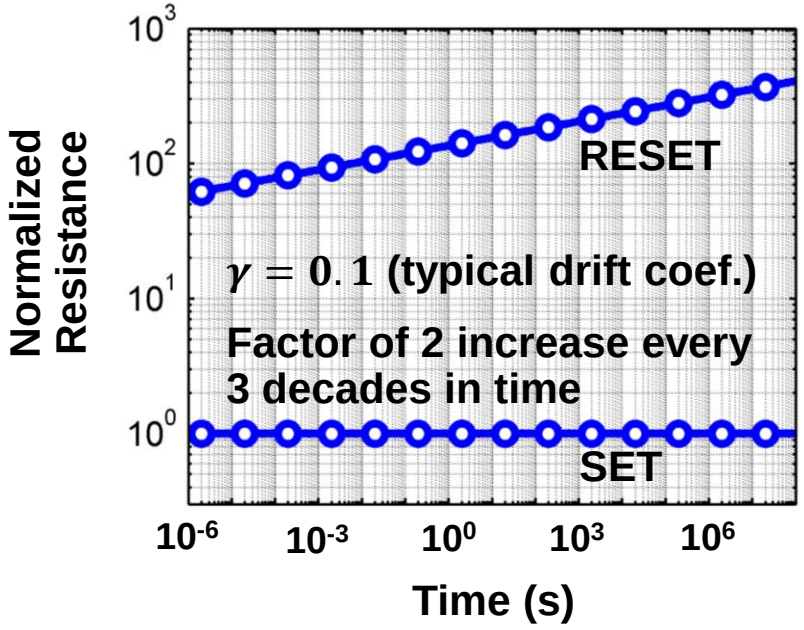
- Heavily used computation in neural network
  - Artificial neural network (Deep learning) : *matrix-vector multiplication*.
  - Spiking neural network : *multiply-and-accumulate*.
- *Resistor network* can perform matrix-vector multiplication and multiply-and-accumulate computation *much faster* in a *massively parallel fashion*.
  - Conductances should be reliably stored in the resistive memory.

PCM reliability issues for neuromorphic applications

- **Resistance drift**
  - PCM cell conductance representing synaptic weight changes over time.

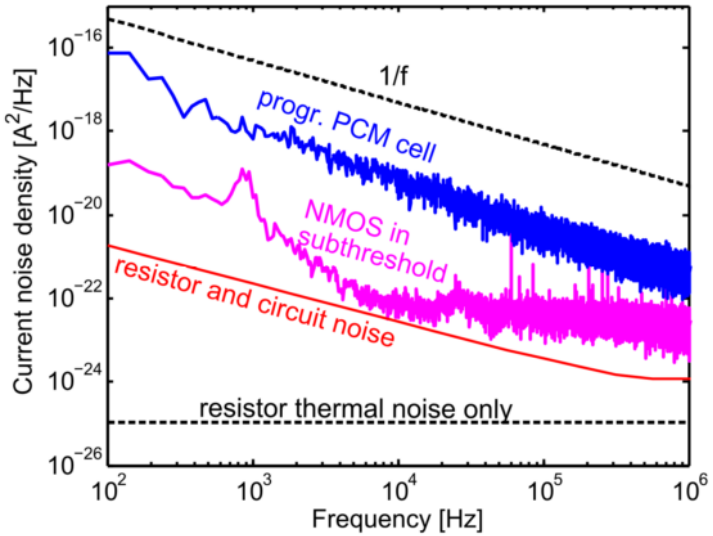
$$R(t) = R_0 \cdot (t/t_0)^{\gamma}$$

γ: drift coefficient



- **1/f noise**
  - PCM cell is a dominant noise source in the memory cell.

The signal-to-noise ($\bar{I}/\sigma$) is estimated to be 19 at 22nm node.

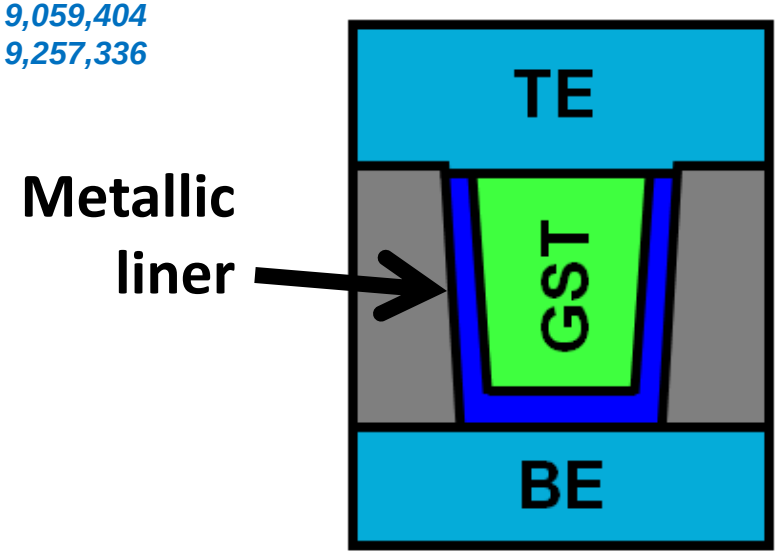


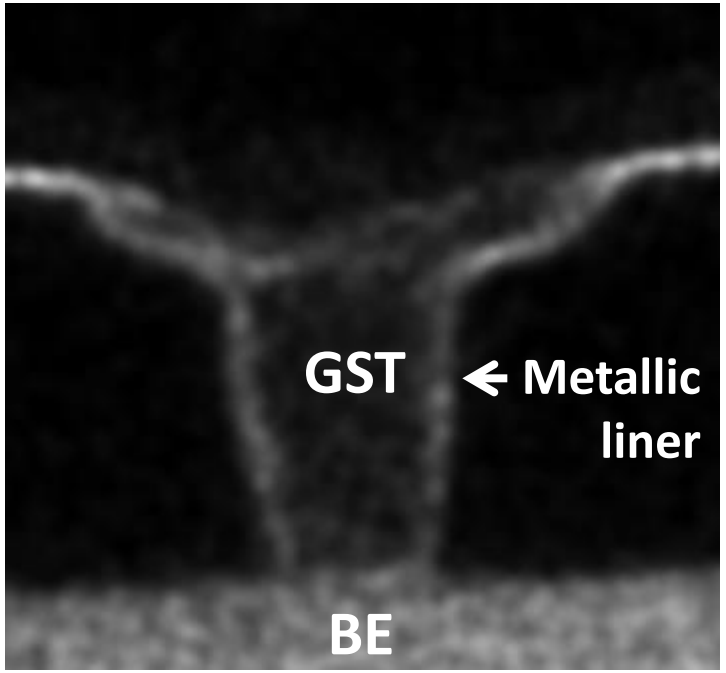
G. Close et al., IEDM 2010

G. Beneventi et al., IEEE EDL 2012

PCM cell with metallic liner

S. Kim et al., IEDM, 2013
M. BrightSky et al., IEDM, 2015
SangBum Kim et al., IEEE TED, 2016
W. Kim et al., IEDM, 2016
W. Koelmans et al., Nat. Comm., 2015
S. Kim et al., US Patent 9,059,404
S. Kim et al., US Patent 9,257,336

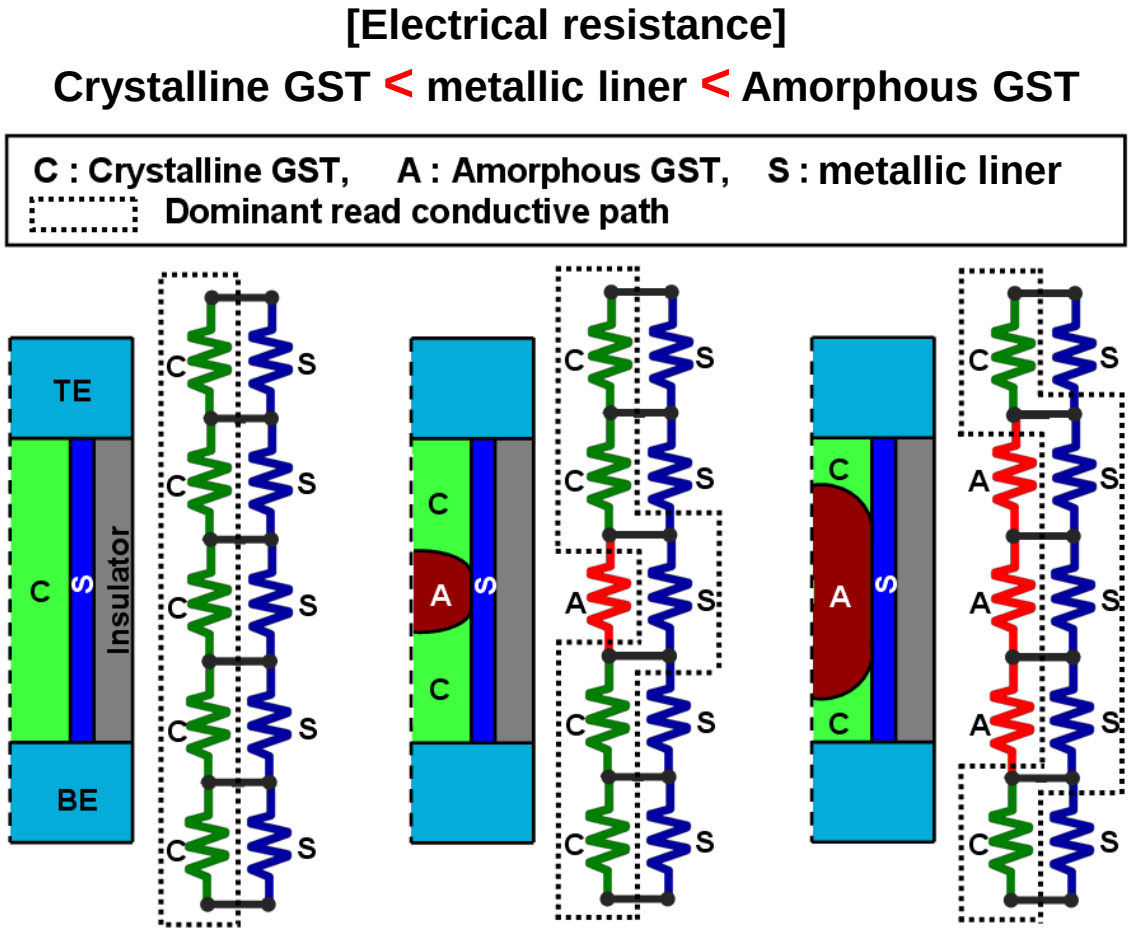


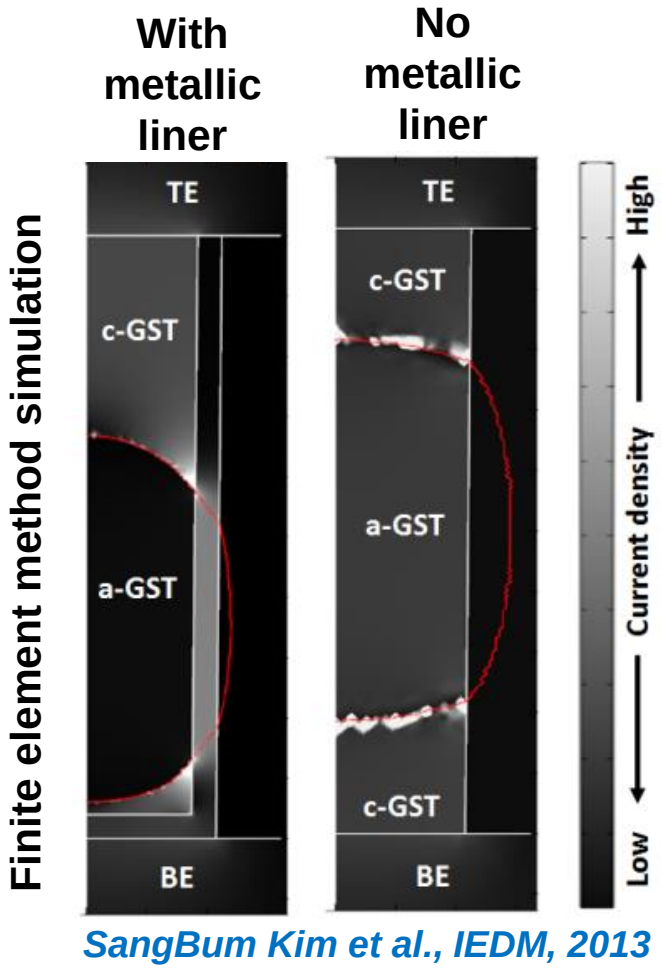


GST : Ge-Sb-Te

- Confined cell with metallic liner
  - *Enables reliability enhancement* : Drift mitigation and noise reduction
  - Cost effective integration

Electrical conduction path with metallic liner

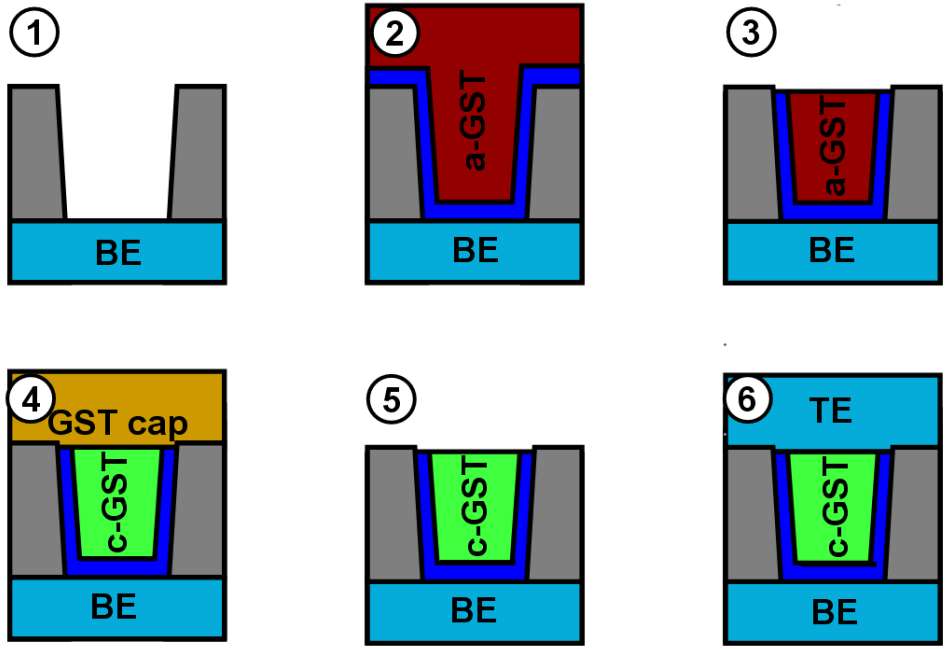




SangBum Kim et al., IEDM, 2013

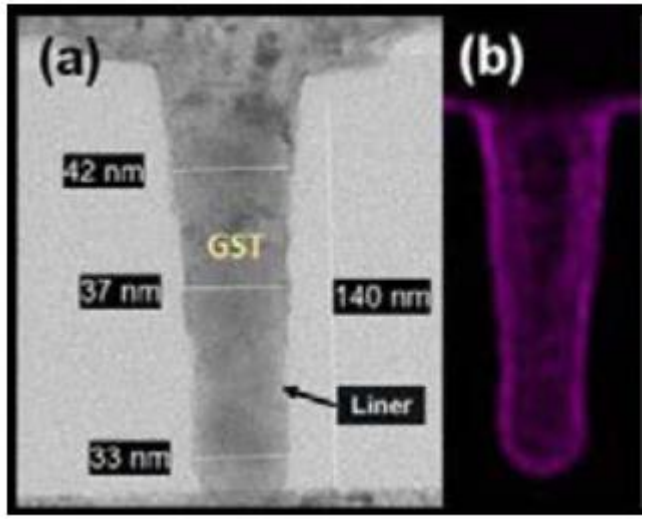
- The metallic liner provides *an alternative conductive path* to the amorphous region.
- The effective metallic liner resistance is modulated by the *amorphous region thickness*.

Cost-effective integration scheme



1. Pore formation
2. Metallic liner + GST deposition
3. Field GST removal
4. GST cap + anneal
5. GST cap removal
6. Top electrode formation

S. Kim et al., IEDM, 2013

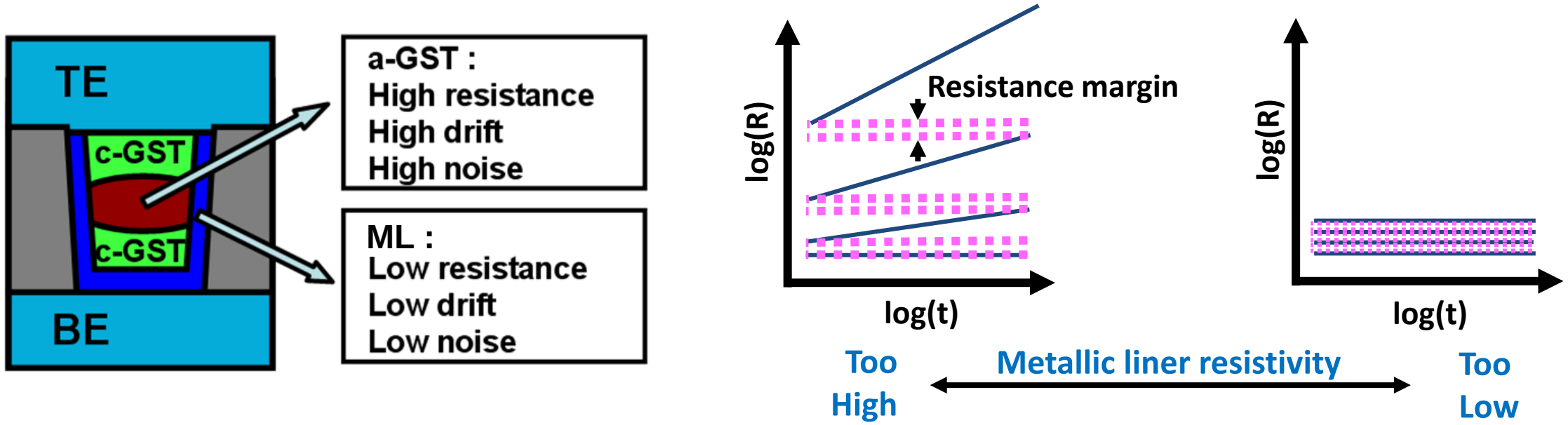


Void-free filling capability of ALD metallic liner and GST ($Ge_xSb_yTe_z$)

W. Kim et al., IEDM, 2016

- *Cost-effective integration scheme*
  - No additional mask
  - Small cell size (Metallic liner 2-8nm)
- ALD metallic liner and GST deposition
  - ALD deposition tool : ULVAC Entron^{TM}-EX W300

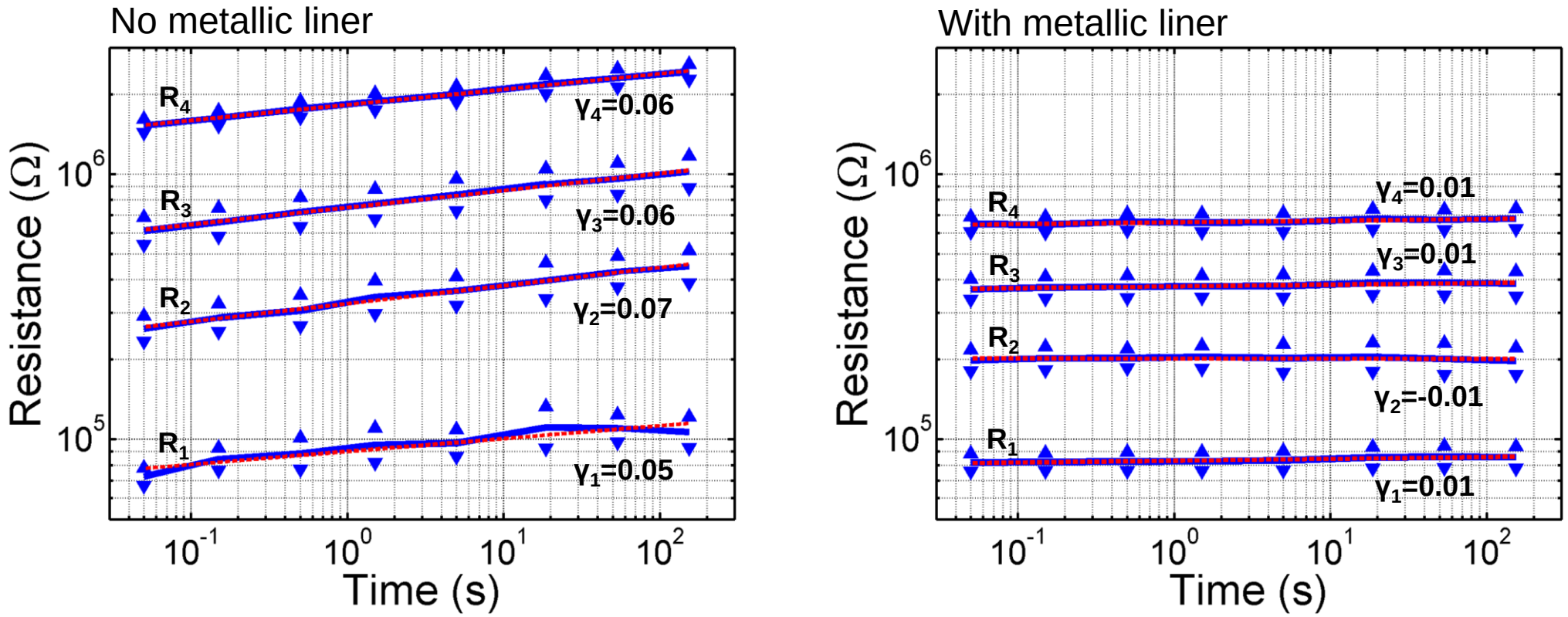
Requirement for metallic liner



- *Finding the right metallic liner* is the most important.
  - Material property requirement for metallic liner
    1. Electrical resistivity (Semi-metallic)
    2. Low noise and drift (Thermal stability)
    3. Wetting capability with the phase change material (Low interface energy)
    4. Conformal deposition (ALD deposition)

Drift mitigation with metallic liner

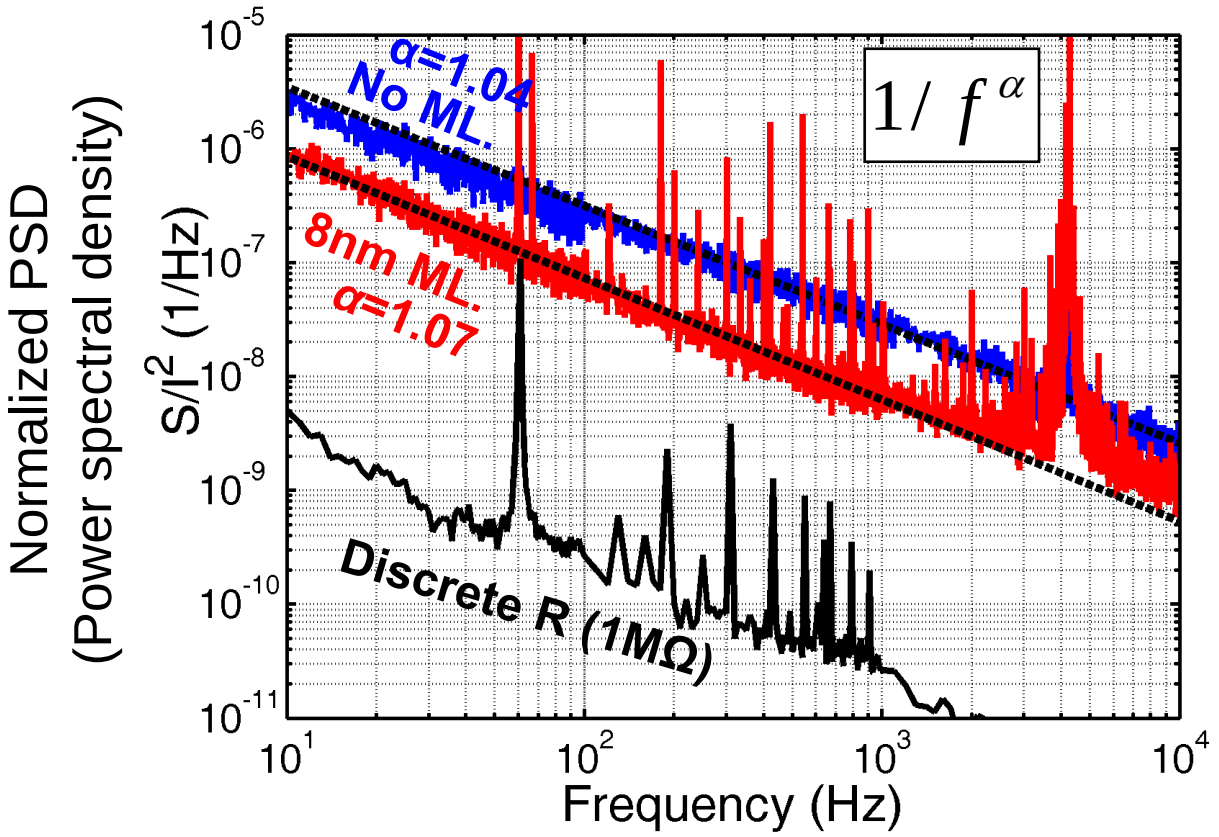
Drift model $R(t) = R_0 \cdot (t/t_0)^{\gamma}$



S. Kim et al., IEDM, 2013

- Drift is successfully mitigated with the metallic liner.
  - 0.06 vs 0.01: *6 times smaller* resistance drift coefficient is achieved.
  - 0.01 : 7% increase over 3 decades in time → Almost negligible

Noise reduction with metallic liner



Signal to noise ratio can be calculated from the normalized PSD

$$\sigma^2_{PSD} / I^2 = \int_{f_1}^{f_2} S(f)/I^2 \cdot df$$

S. Kim et al., IEEE TED, 2015

- Noise is successfully reduced with metallic liner.
  - The PCM+ML(8nm) cell shows *4x smaller noise PSD* than PCM without ML.
  - But the noise is still much higher than a discrete resistor.

Conclusion

- NVM neuromorphic chip with on-chip LIF and STDP
  - Neuromorphic core for spike-based machine learning
  - Non-volatile PCM as a synapse
  - Brain-inspired : LIF, STDP, spiking neural network
  - 64k (256-by-256) 2T-1R PCM synaptic cells
  - Verified LIF and STDP operation at the array level

- Resistance stabilizer for neuromorphic NVM
  - Resistance needs to be stabilized to reliably store synaptic weights
  - The confined PCM cell with metallic liner mitigates
    - Resistance drift
    - Noise

Backup