Scaling Limitations of Flash Memory

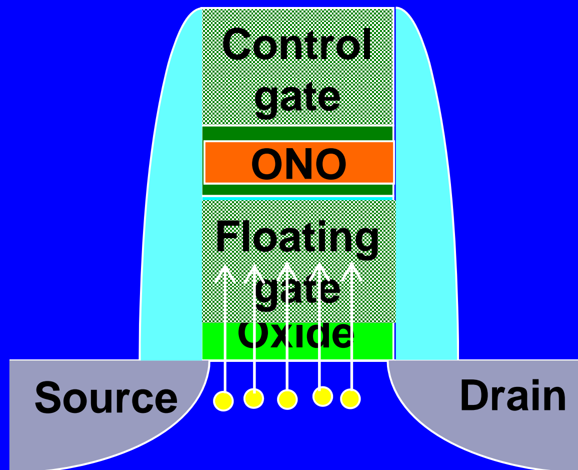
Rich Liu

Macronix International Co., Ltd.

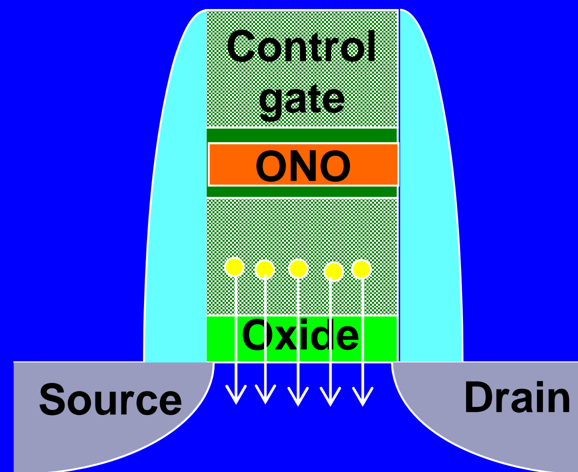
Hsinchu, Taiwan, R.O.C.

MXIC MACRONIX INTERNATIONAL CO., LTD.

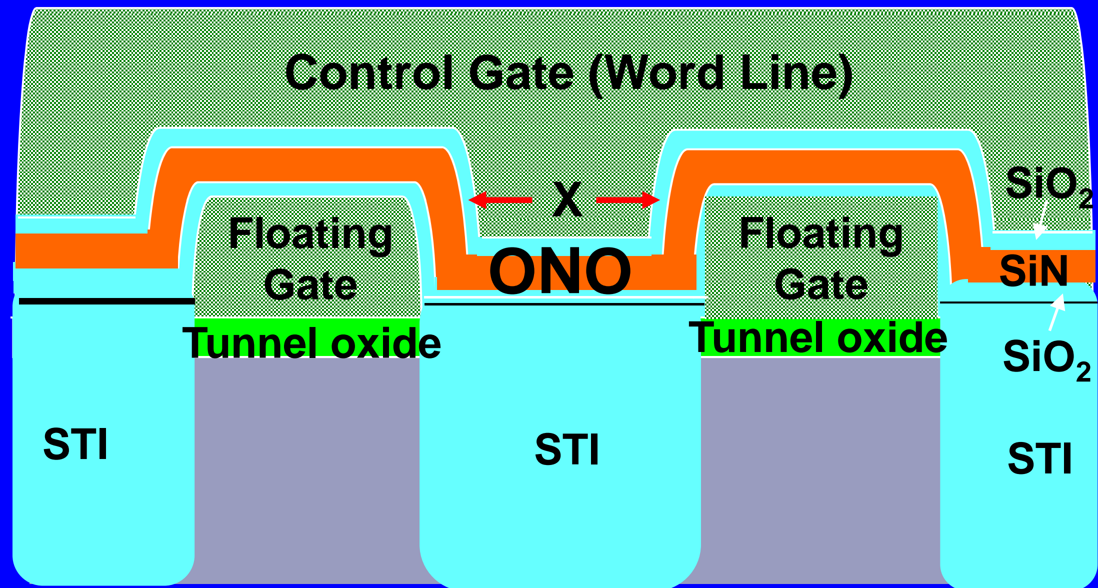
Floating Gate NAND Device 101



Program by channel FN tunneling



Erase by –FN



Gate Coupling Ratio (GCR)

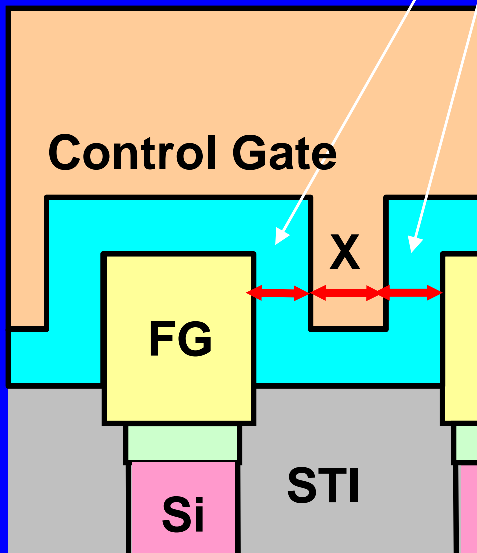
= C(CG to FG) / C(FG total)

Must be > 0.60

V(FG) = (GCR)·Vg

(Want most of the gate voltage drop across the tunnel oxide, not across the ONO.)

Physical Limit for Floating Gate NAND



1. FG must be tall enough to give good GCR.
2. At < 20nm node, there is no space (X < 0) left for control gate after IPD filling.

Not a physics limit.

A physical (geometrical) limit.

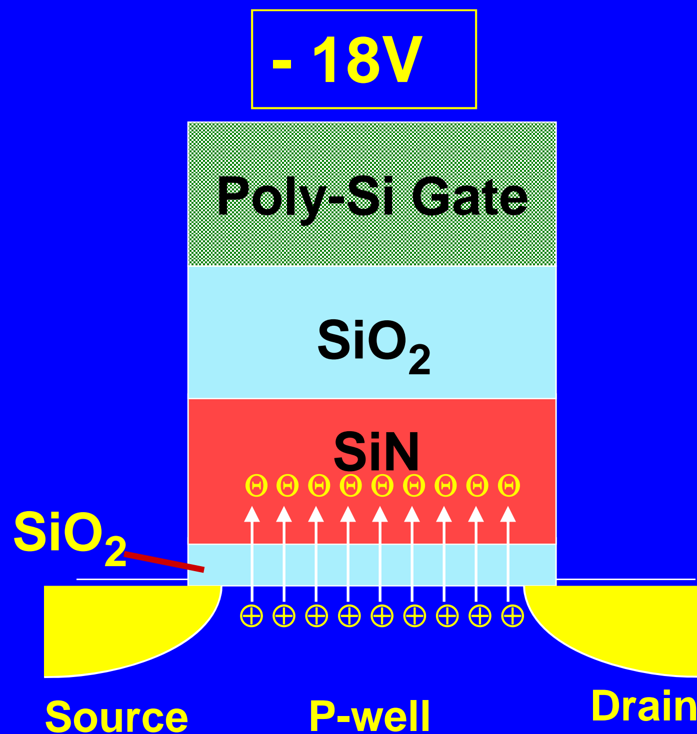
Device doesn't work w/o GCR.

→ Charge trapping device (planar).

or

→ Planar FG, resonant tunneling, or nano-crystal device with high-K/metal-gate

SONOS Has It's Own Problems



SONOS device:
Electrons are trapped in SiN. De-trapping is very slow. → Must use hole tunneling to erase (hard).

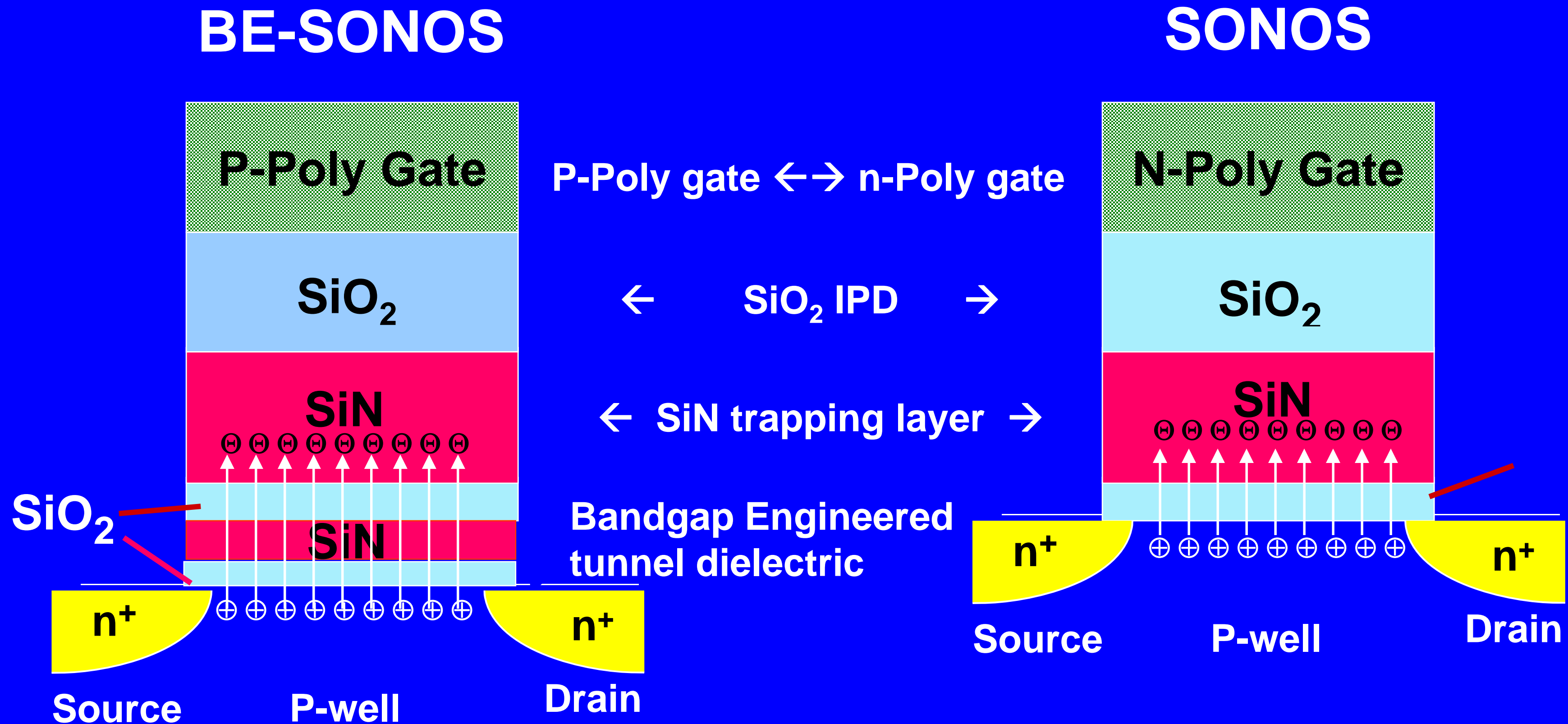
Hole tunneling needs very thin tunnel oxide.

Thin oxide cannot stop direct tunneling → poor retention.

SONOS is known for many years.

There is no "right" thickness of tunnel oxide that can satisfy both erase and retention requirements.

One Solution: Barrier Engineering → BE-SONOS



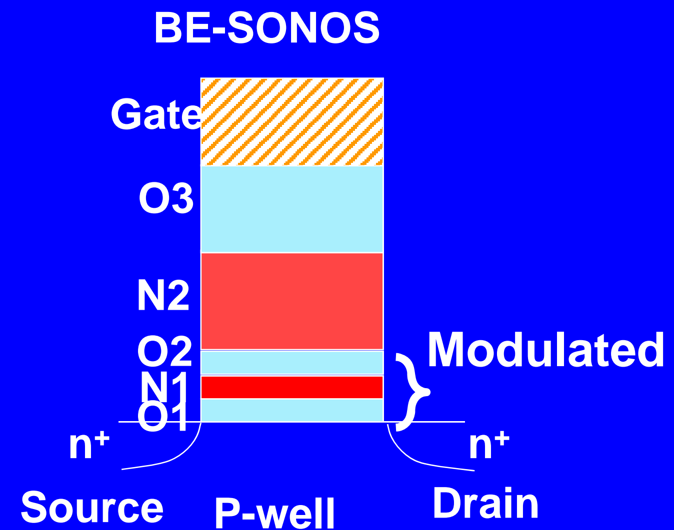
Difference between BE-SONOS and SONOS:

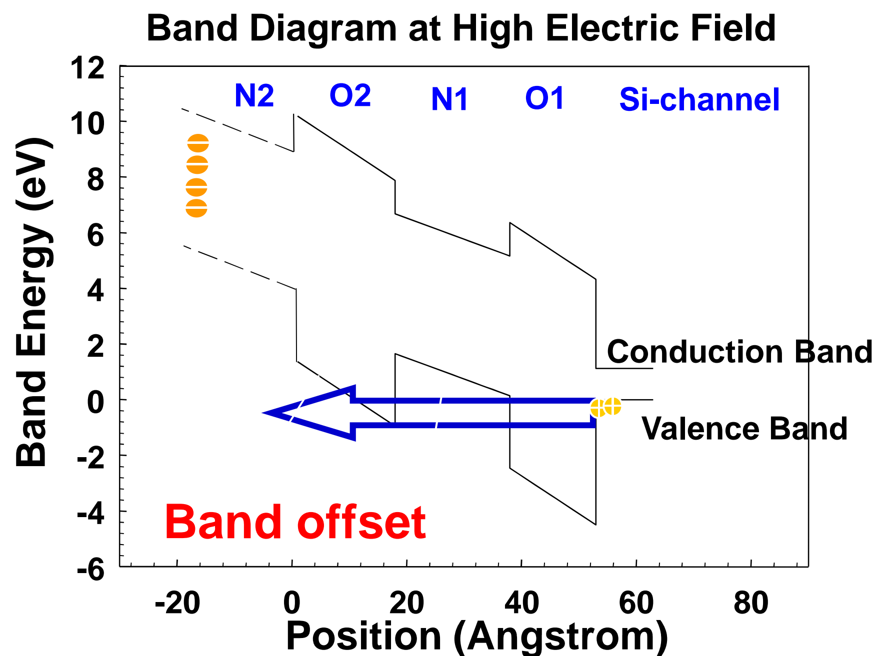
- ❑ Composite ONO tunneling barrier allows both fast hole erasing and good data retention
- ❑ P-poly gate to reduce gate injection

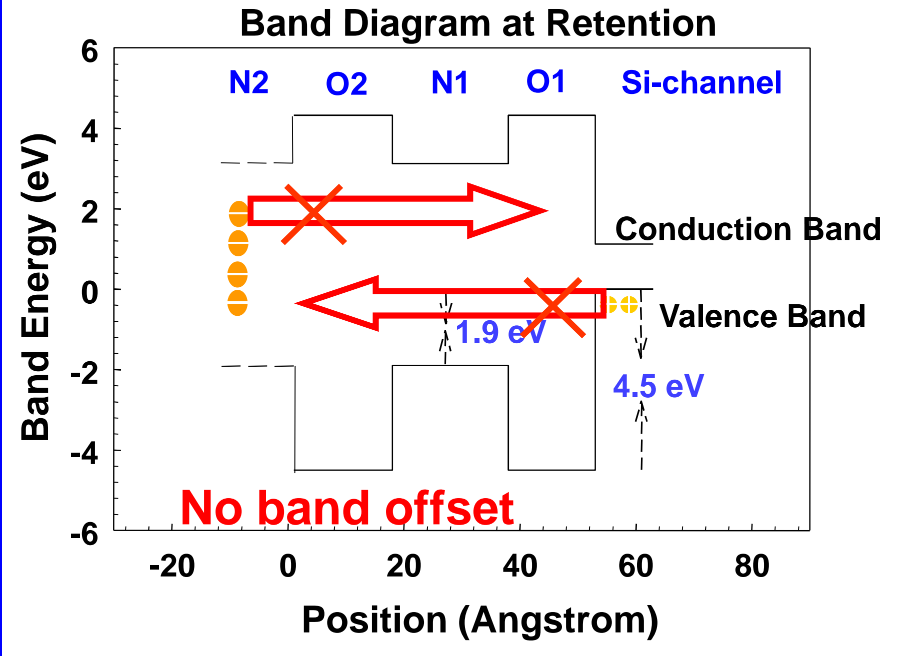
Barrier Engineering of Tunnel Oxide

SONOS → BE-SONOS

Modulated tunneling barrier → Achieves both erase and retention

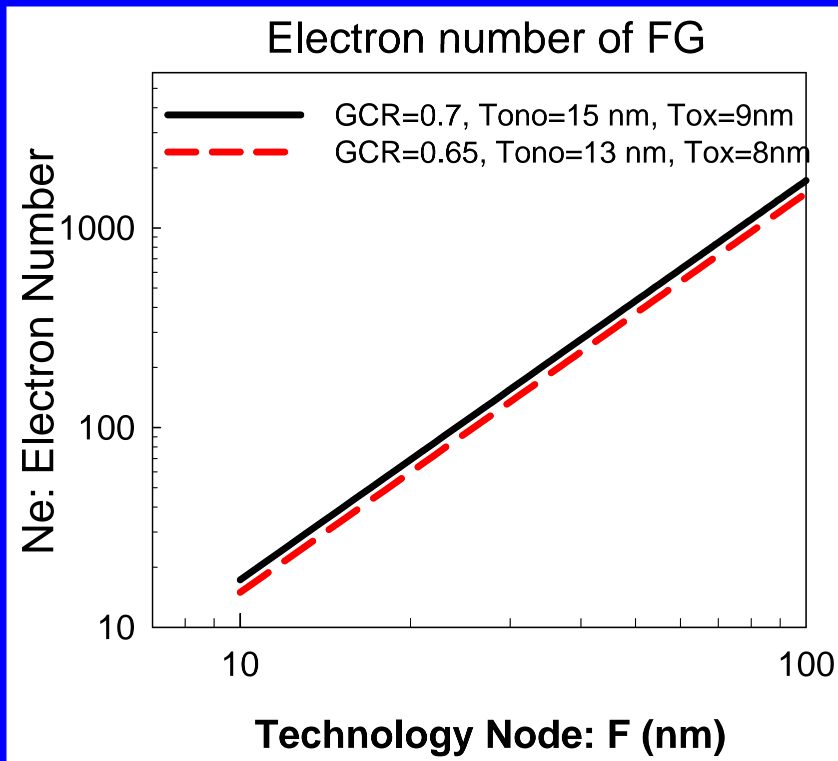




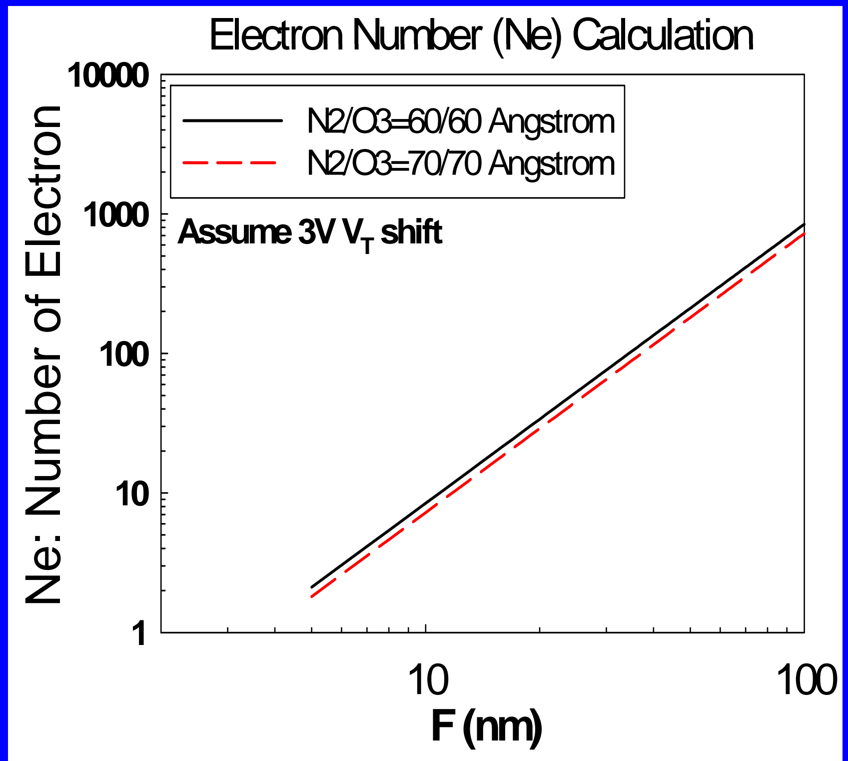


1x nm Nodes: Running out of Electrons

The ultimate scaling limit for both FG and CT is the small number of storage electrons.

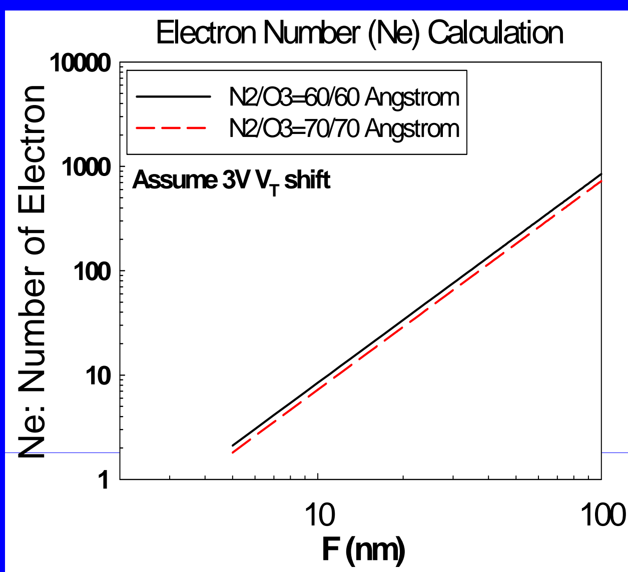


Number of electrons in FG device (~ 15 for 10nm device)

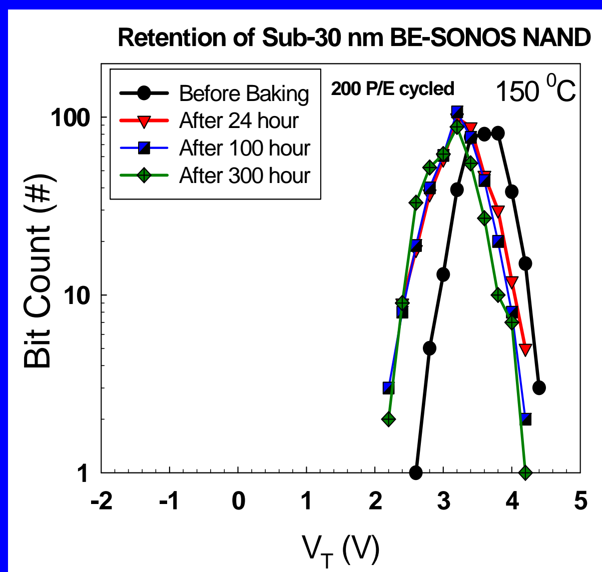


Number of electrons in SONOS device (~ 10 for 10nm device)

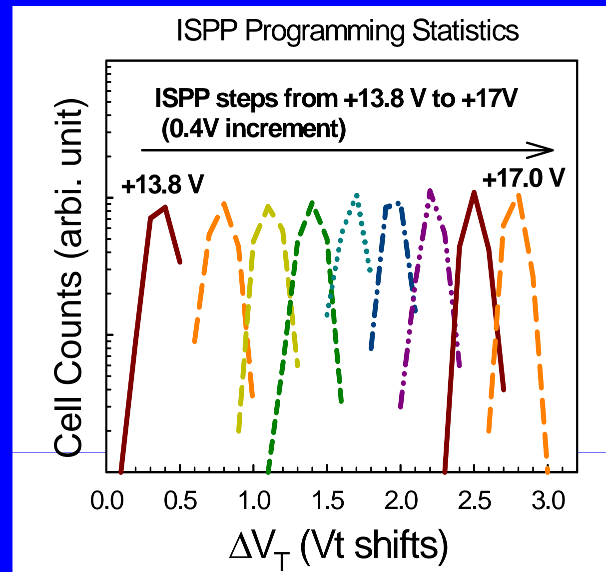
Running out of Electrons



Number of electrons in SONOS device (~ 50 for 25nm device)



Retention for sub-30nm BE-SONOS @ 150C (Quite good !)



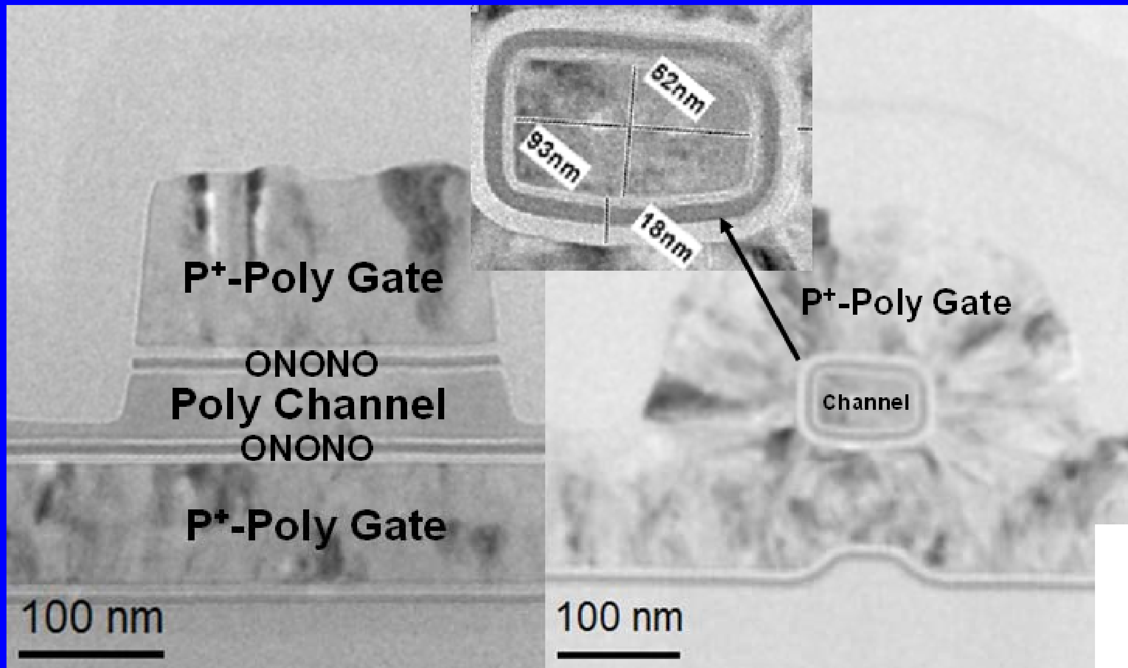
Program 20nm device Identical pulse gives different Vt

For CT devices, retention is still good even when Ne < 50.

Programming shots, however, have < 10 electrons.

→ Statistical limit for MLC first, eventually for SLC.

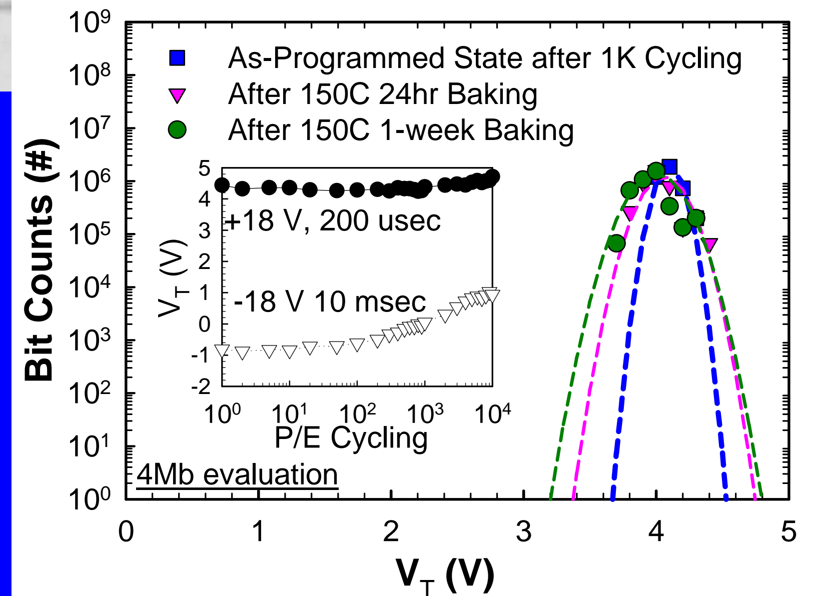
Beyond 1x nm Node – 3D Arrays



Surround gate device is suitable for 3D layer stacking integration.

BE-SONOS TFT device shows very good performance, approaching that for bulk device.

Both geometrical and physics limits still exist. But 3D layering uses large (40nm) devices.



Summary

❑ Floating gate NAND Flash faces physical (geometrical) limit at ~ 20nm node.

❑ Charge trapping device can go further (being planar)

❑ Number of electrons decreases rapidly with node, even CT devices face statistical limit at 10nm node.

❑ Only known solution is 3D layering.

❑ 3D does not solve physics and physical limitations. 3D by-passes these limits by using relatively large devices (~ 40nm).

❑ There is no perspective of using FG device for 3D. Must be CT devices.