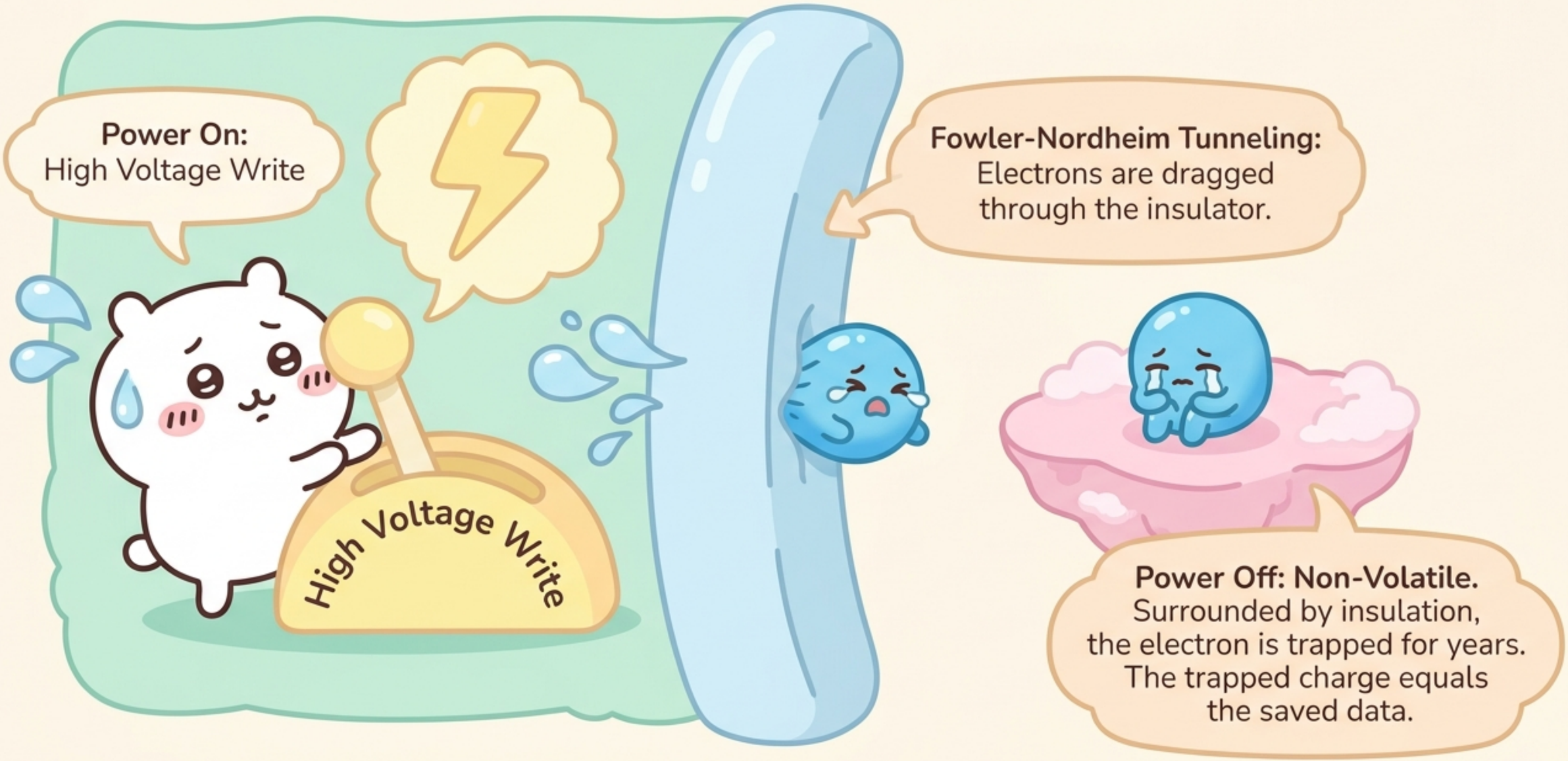
Permanence requires physical isolation.



Density is purchased with fragility.

COMPARISON MATRIX

SLC (Single-Level Cell)

1 Bit per Cell | 2 Voltage Levels | ~50k-100k P/E Cycles

Verdict: Reliable, thick margins, expensive.

QLC (Quad-Level Cell)

4 Bits per Cell | 16 Voltage Levels ~100-1,000 P/E Cycles

Verdict: Razor-thin margins, extremely fragile, cheapest per GB.

Reading is harmless, but writing destroys the cell over time. The industry's push to cram 16 voltage levels (QLC) into one cell slashed lifespan from 100,000 cycles to mere hundreds.

Scaling stopped shrinking and started climbing.

Surface level:
2D Planar NAND Wall
(Cells too close, bits corrupt each other)

Shallow depth:
2013: 24 Layers

Mid depth:
2024: 276 to 321 Layers
(Micron G9, SK Hynix)

Deepest drill point: 2030
Roadmap: 1,000+ Layers

The cost-per-gigabyte curve survived by changing axes. Cramming bits vertically means vastly more terabytes on the exact same silicon footprint.

The structural absurdity of block erasure.

Write Unit = Pages
(Small, ~a few Kilobytes).
Metaphor: Single Rooms.

Erase Unit = Blocks
(Massive, Hundreds of Pages).
Metaphor: Whole Buildings.

Flash Translation Layer (FTL) / Wear-Leveling in action.

You cannot overwrite a single page. To change one byte, the controller must read the whole block, move the good data elsewhere, and erase the entire block. This ‘Write Amplification’ means 1GB of saved data can cause 3GB of actual physical wear to the drive.

AI transforms the cold archive into a hot overflow.

Synthesis

The Arbitrage: AI inference is severely capacity-starved but overwhelmingly read-heavy.

The Perfect Match: Because reading NAND is physically gentle, the AI workload completely bypasses QLC's fatal weakness (write fragility) while exploiting its greatest strength (vertical density).

HBM: Hot, scarce, and wildly expensive. Wasted on idle AI context memory.

CMX / HBF Offloading: Spilling cold KV-cache data down the hierarchy.

NAND:

NAND: Cold, abundant, and incredibly cheap.

NAND hasn't changed. But by routing read-heavy AI workloads to the cheapest bucket, the industry is bypassing the memory wall.