Master of Science's Thesis in

Micro and Nano Technologies for Integrated Systems

2025/2026

Politecnico di Torino

Multi-State Conduction in Y-36 Lithium Niobate: Electrical Characterization and NeuroSim Simulation for In-Memory Computing

Supervisors:

Prof. Carlo Ricciardi

Prof. Gianluca Piazza

Candidate:

Francesco Pio Minoia

Table of Contents

Abstract

The computational efficiency of artificial intelligence is becoming more and more constrained by the so called Von Neumann bottleneck causing the transition toward analog Compute-in-Memory (CIM) architectures. This requires synaptic devices capable of combining non-volatile storage, high resolution and as linear as possible conductance modulation. This thesis investigates the potential of ultra-thin (43 nm) Y-36 Lithium Niobate ($LiNbO_3$) films to address these requirements, through devices characterization and system-level benchmarking.

Through electrical characterization of Metal-Ferroelectric-Metal devices an optimized pulse protocol was developed to handle the switching dynamics. This approach was able to set 102 distinct conductance states (more than 6-bit precision) within an analog window operating at currents as low as $3\ \mu A$.

System-level benchmarking via NeuroSim software on an MNIST set classification task showed that high synaptic resolution and acceptable linearity is necessary for training stability. The optimized configuration achieved an accuracy of 77.61%. Hardware analysis highlighted a performance duality: while the device exhibits remarkable read energy efficiency (~435 μJ), the write energy is currently high due to long pulse duration and high device area. Theoretical projections on future works indicate that overcoming these limits by optimizing switching dynamics to faster regimes can reduce consumption to the femtojoule level, validating Y-36 $LiNbO_3$ *as a scalable platform for next generation neuromorphic computing.*

Introducion

Rapid ascent of Artificial Intelligence and Von Neumann bottleneck

The fast growth of artificial intelligence, especially deep learning, has sharply increased the need for computing memory and power[1]. Because modern neural networks (NN) involve billions of parameters, their training often depends on a huge amount of data along with powerful hardware implementation. Such demand has worsened existing limits in system design, especially constraints from the classic von Neumann model. In standard setups, processing units and storage stay apart; this forces constant movement of data across a common channel linking the processor and the memory through buses[2]. Although that structure once supported flexibility, expansion and scalability, it now struggles under AI workloads needing wide memory throughput and quick responses during computations.

The von Neumann bottleneck occurs since instructions and data move slowly due to limited connection capacity between chip (for example the CPU) and memory. When models grow larger, moving data takes more time and can drive most of the energy use instead of actual calculations. For neural networks, constantly transferring weights during trainings, outputs and temporary values may consume a big portion of total power, far exceeding what math operations require. Limited memory access doesn't just delay processing, it leaves processors in an idle state as they wait for inputs, wasting resources and time. While newer AI chips handle math and internal storage much quicker, moving data outside the chip remains slow and inefficient by comparison.

Also, the typical sequential flow of von Neumann architectures makes simultaneous and parallel computations harder to handle[3]. Instead, multi-core designs try reducing delays by adding more cores; however, problems such as energy use and growing shared memory demands are still present. Devices built for speed, like GPUs, reduce certain limits using tailored memory management and hierarchical caching, though their performance is still held back due to slow buses and access to the main memory block.[4]

Compute-in-memory

To address this issue, new design approaches, such as compute-in-memory (CIM) and neuromorphic computing, merge processing and calculations with storage, so information is handled right where it resides. Instead of moving data through buses, devices like memristors, resistive RAM and phase-change elements allow both storing and calculating within the same array structure. These methods use analog computations across array-like structures, running many calculations at once to cut down delays and power consumption[5]. Modern hardware solutions now take advantage of localized computing, showing major improvements in performance and energy savings when dealing with artificial intelligence tasks.

With AI expanding into areas like self-driving cars and health diagnostics, demand rises sharply for fresh and efficient hardware designs. Moving past the von Neumann limit sparks joint advances across material research, chip design development and computational methods, setting the stage for leaner learning machines.[4]

Academic effort in Neuromorphic Computing

Over the past ten years, studies in neuromorphic computing have grown fast, fueled by demand for hardware that uses less power while supporting advanced AI tasks. Taking inspiration from Neuroscience, these systems take cues from how neurons connect, using dense interconnectivity, dynamic learning patterns and continuous adaptation. Because of this shift, experts from electronics, materials science and cognitive research now work together, making joint efforts that push forward novel circuit designs and system layouts.

Major academic work has led to a variety of different hardware designs. Key cases involve digital systems such as IBM TrueNorth, Intel Loihi, or SpiNNaker, these use special chips to run thousands of artificial spiking neurons and synapses at once, enabling fast parallel computation.[4] At the same time, research community is testing novel materials and devices like memristors, phase-change memory units, spin-based tools and photonic circuits[3] to build analog and mixed signal architectures that mimic the behavior of real synapses.

Recent studies point out potential uses of neuromorphic computing. For instance, systems using event-based vision, robotics, sensors, brain implants, and compact AI for on-site

tasks requiring minimal energy and real time responsiveness[5]. Still, hurdles exist; one major issue is connecting scalable, software-friendly digital designs with the high efficiency of new neuromorphic chips.

Different types of Neuromorphic Computing devices

Neuromorphic computing uses different devices and physical methods to create hardware that mimics brain-like functions. One leading approach relies on electric charge, especially in materials with ferroelectric properties, thanks to stable data retention, fast switching, while possibly supporting multiple states storage per unit.

Ferroelectric RAM

Ferroelectric RAM (FRAM) is widely used especially for non-volatile digital memory applications relying on the bistable polarization of ferroelectric layers such as PZT or $HfZrO_2$ and typically operates with the two binary states 0 and 1. For example, FRAM devices show remanent polarization of around 20–100 $\mu C/cm^2$ and high endurance exceeding 10^{12} cycles, but their use in neuromorphic computing is generally limited by the sharp switching and binary nature of the response making it difficult to behave like a memristor, also due to non-linear polarization curves, although some architectures and advanced pulse schemes can achieve more than two conductance states in lab conditions.[4]

Non-binary ferroelectric systems are drawing more interest in neuromorphic computing, especially where or multi-level tuning supports synaptic behavior. Instead of full switching, using controlled pulses or engineered domains helps reach various resistive states reliably. As an example, thin-film MFM units built from $LiNbO_3$ achieved 40 conductance levels, equal to around 5 bits, while certain HfO_2 setups reached up to 8 levels (roughly 3 bits)[6]. Even so, differences between individual devices, data stability over time, and endurance during repeated use remain key challenges before wide integration.

Ferroelectric tunnel junctions

Ferroelectric tunnel junctions (FTJs) employ ultrathin ferroelectric barriers (typically less than 5 nm) where the resistance state is determined by the barrier's polarization orientation via quantum tunneling. Researchers have observed stable and robust ON/OFF ratios, often exceeding two orders of magnitude and have found that tunneling current can be almost continuously modulated through controlled polarization and pulse dynamics.[7] Although FTJs utilizing materials such $HfZrO_2$ or $BaTiO_3$ show impressive retention operating with very low energy (requiring only fJ to pJ per bit)[8], challenges remain regarding device uniformity and integration with well established CMOS logics.

Ferroelectric field effect transistors

Ferroelectric field effect transistors (FeFETs) incorporate a ferroelectric layer directly as the gate dielectric, in this way the transistor's channel conductance can be tuned by its own polarization state. FeFETs are well known for their extremely fast sub-nanosecond switching capabilities, multiple analog threshold states and, like FTJs, compatibility with CMOS logic. In the context of neuromorphic computing, FeFET synapses are particularly valuable as they can represent a decent dynamic range of weights while operating at high speeds and low voltages. Currently, state-of-the-art FeFET arrays have successfully demonstrated synaptic behavior with up to 32 analog states.[8]

Overall, charge-driven ferroelectric systems like FRAM, FTJs, MFM capacitors and FeFETs are becoming key for neuromorphic computing, offering analog tuning plus compact and scalable design. Although multi-level function is now more achievable thanks to better materials and fabrication methods, maintaining stable performance across large arrays continues being an active area of research development.

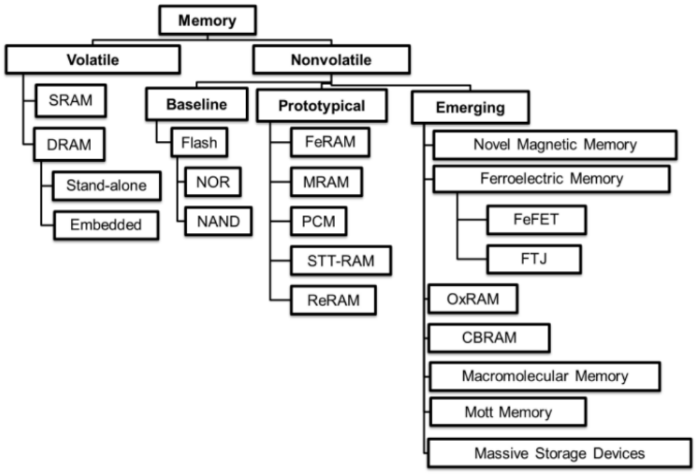


Figure 1. Classification of conventional and non-convention memories.

[Adapted from IRDS Beyond CMOS 2023]

Resistive based neuromorphic memories provide other options instead of charge-driven systems, each using distinct physical principles and design approaches. Among them, three main types stands out: RRAM, PCM (also called PcRAM), while magnetic variants include STT-RAM and MRAM.

Resistive RAM

Resistive RAM (RRAM), sometimes called memristor arrays, uses a thin insulating layer, made from materials like HfO_2, TiO_2, or Ta_2O_5 placed between two metal contacts. Instead of bulk changes, it works through tiny conductive paths that form or break; these filaments rely on missing oxygen atoms (oxygen vacancies) or moving metal ions to shift resistance levels between high resistance state (HRS) and low resistance state (LRS). From the performance standpoint, they support tight scaling, reaching about $4F^2$ per cell, along with rapid switching under 10 nanoseconds and efficient energy usage around 0.1 to 10 pJ per write cycle. By precisely adjusting voltage pulses, multiple conductance steps can emerge, one setup manages up to 128 levels (equaling 7 bits)[5], [8], with ON/OFF ratio commonly above 100. Advanced RRAM setups allow IMC using crossbar grids for efficient vector-matrix multiplication, this boosts appeal for brain-inspired systems. Still, issues like inconsistent performance, limited lifespan, or difficulty tuning weights accurately persist.

Phase-change memory

Phase-change memory (PCM or PcRAM) utilizes chalcogenide materials such as $Ge_2Sb_2Te_5$ (GST) that can be reversibly switched between amorphous (high resistance) and crystalline (low resistance) phases by controlled heating by using electrical pulses. Multi-level resistance is possible by controlling the crystalline volume fraction, allowing up to 120 stable conductance levels in experimental devices. Typical PCM features switching speeds in the microsecond range (around 6 μs) and write energy on the order of 100pJ per bit, with excellent scalability and retention (over 10 years at room temperature)[9]. PCM crossbars have demonstrated highly parallel VMM for neural networks but device-to-device variability, reset drifting, and relatively high programming currents compared to RRAM and MRAM pose constraints for large arrays.

Magnetic RAM

Magnetic RAM, particularly spin transfer torque RAM (STT-RAM), retains data through magnetic alignment in ferromagnetic layers divided by a slim insulating film, often MgO. Writing happens when electric current flips the magnetization direction, producing either low or high resistance based on whether alignments match or oppose, enabling natural digital storage. Instead of relying on charge, this method allows quick access times, (reaching about 10 nanoseconds), while keeping memory intact without power. Although some studies show promise for storing more than one bit per cell using incomplete switching or new compounds, achieving stable analog behavior at scale is difficult. Thanks to strong durability and CMOS integration, it works well for cache and non-volatile storage; yet data retention and scalability for brain-inspired computing need further exploration[10].

All these resistor-like storage units bring key benefits compared to standard charge-driven types: they allow gradual tuning, compute inside memory, work at lower voltage, also pack tightly together. While one system may last longer, another might scale better or use less power, each balances performance differently. Current studies aim to build stable multi-step control, reduce inconsistencies across cells, while linking them as artificial synapse nodes in real neural systems that speed up AI tasks.

Pitfalls of memory technologies: energy and latency efficiency limitations

Although neuromorphic computing offers major improvements in hardware efficiency, every type of new memory tech, whether charge-driven or resistance-driven, comes with some drawbacks.

Charge-based Memories

Ferroelectric RAM works well over time, uses low energy, provides non-volatility, yet only handles binary states because it flips quickly between them. That makes it less useful for brain-inspired computing, which needs smooth, graded changes in memory values[4]. Some efforts have been tried achieving levels beyond two by tweaking how domains form or using incomplete switches; these show promise, but controlling many stable steps consistently remains hard. Variability across chips, uneven response, and shifting stored values complicate accurate tuning. As more analog stages become available, data may fade faster reducing accuracy during prolonged use.

Ferroelectric Tunnel Junctions (FTJs) enable analog control using electron tunneling across extremely thin ferroelectric films, less than 5 nm thick. Although these junctions show strong ON/OFF performance, such as around 300 in BaTiO setups, scaling them down remains challenging because fabrication is complex; also, leakage currents and defect sensitivity interfere[7]. For instance, certain tests reveal that HfO_2-type devices, one widely used FTJ option, achieve just 8 distinct levels, equivalent to 3 bits.[6]

Ferroelectric Field Effect Transistors (FeFETs) show potential for compact, quick performance through multi-step voltage control. When used in neuromorphic settings these devices support broad range of memory weights variation. Recent FeFET grids demonstrate 32 continuous levels mimicking synapse activity[8]: however, precise analog adjustments typically last around 10^4 cycles.

Resistive-based Memories

Resistive RAM works fast, speeds reach nanoseconds, with energy use as low as 0.1 to 10 pJ; it also handles multiple resistance levels. Still, its filaments form stochastically: forming or breaking them varies between devices and repeated operations leading to substantial variability. Because of that, performance shifts over time and some states may blend together. For continuous analog operation, lifespan ranges from 10^4 up to 10^8 cycles[11], [12]. Even if certain tests report 128 distinct levels (about 7 bits), inconsistency remains a core challenge for reliable memory precision.

Phase Change Memory supports around 100–120 steady conductance steps, using materials like GST, by controlling its crystallization; it offers strong data retention lasting over a decade along with compact design. However, operation usually demands higher energy per bit (~100 pJ) making it less efficient compared to RRAM. Switching is also relatively slow, needing pulses of about 6 microseconds. Over time, small shifts in partially programmed states occur because of relaxation in the crystalline structure, which reduces accuracy[8].

Magnetic RAM, particularly STT-RAM, is known for lasting a long time, plus it employs low energy and fast switching (about 10 ns). Although usually seen as binary, in neuromorphic simulations such as NeuroSim, it’s treated as having multiple states to mimic synapses[13]. Still, real-world use in analog mode runs into issues: data doesn't stay stable long enough, reading can disrupt stored values, and differences between levels are often too wide.

Despite differences among types, challenges like inconsistent processes, restricted scalability, lack of consolidated industry standards for testing and benchmarking make implementation harder[11].

Promising Aspects of Y-36 $LiNbO_3$ Ferroelectric Devices

Thin-film $LiNbO_3$ devices, particularly with Y-36 orientation, have emerged as strong candidates for neuromorphic and memory applications due to their impressive material and switching properties. $LiNbO_3$ has high spontaneous polarization and robust ferroelectricity up to its high Curie temperature (around 1140–1210°C). Recent studies

on metal–ferroelectric–metal structures using 43 nm Y-36 $LiNbO_3$ thin films demonstrate remanent polarizations near 58 µC/cm² as shown in figure 3, with low coercive fields (Ec as low as 0.4-0.9 MV/cm. This allows polarization switching with voltages around 2 V[14]. These characteristics offer distinct advantages for low-power, high-density applications.

Endurance tests show consistent behavior beyond 10^9 cycles; meanwhile, data stays intact over brief periods, Pr remains stable after 100 seconds, which makes it suitable for dependable storage or adaptive analog circuits. Instead of binary mode, multiple states are now possible, a key step toward brain-inspired computation: experiments on the 43 nm structure reached 5-bit resolution (~40 levels)[6], achieving a conductance switch ratio near 6[14]. While nonlinear response and limited ON/OFF contrast still require investigation, the Y-36 variant enhances both ferroelectric output and mechanical sensitivity, while also easing incorporation into compact acoustic or light-based modules, useful in combined processing platforms and tunable oscillator designs.

Overall, Y-36 aligned $LiNbO_3$ films show strong polarization alongside reduced switching voltage; they also offer solid durability and enabling new multi-state analog control making them a compelling option for future neuromorphic systems.[14]

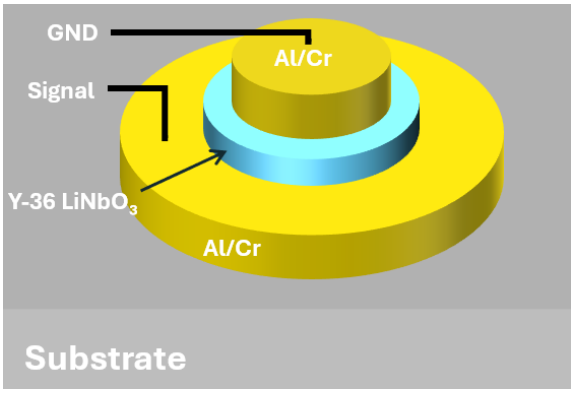


Figure 2 - MFM capacitor structure[14]

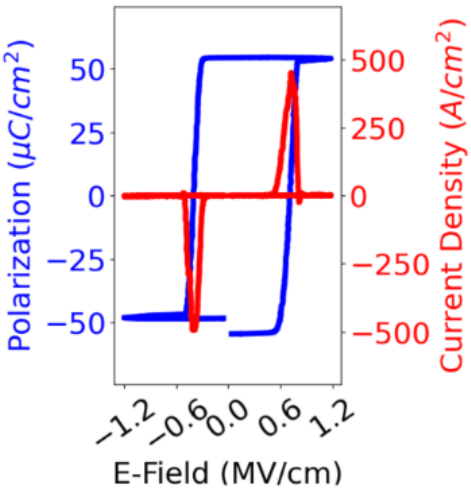


Figure 3 - Hysteresis loop of the characterized $LiNbO_3$ device[14]

Introduction to ferroelectric materials and $LiNbO_3$

Ferroelectric materials are intriguing solids showing natural electrical polarity. In other words, they carry an inner charge separation without needing outside voltage. The reason lies in their built-in dipole moment; polarity occurs only when ion patterns lack symmetry. Such asymmetry happens exclusively in crystals missing central balance.

This reversal happens naturally. A strong outside voltage can shift it from one state to another in a reversible manner. Common types are oxide materials with perovskite layout, like $BaTiO_3$ or PZT, also $LiNbO_3$.

$LiNbO_3$ forms a trigonal crystal structure (figure 4), often called rhombohedral, part of the R3c space group. Its ferroelectric behavior comes from lacking symmetry around a central point so positive and negative charges do not align inside the unit cell. This misalignment results in an overall electric dipole moment.

$LiNbO_3$'s structure relies on linked oxygen octahedra. These units connect face-to-face, orienting parallel to the crystal's polar c-direction, also called the trigonal axis. Inside this network, Nb ions sit within shared corners of NbO_6 groups. Meanwhile, Li ions take up spaces found between these clusters. Polarization happens when Li+ and Nb5+ shift along the c-axis, moving off-center inside their surrounding oxygen cages[15]. This setup forms a clear dipole along the c-axis. Being a uniaxial material, its natural polarization aligns strictly in one of two opposing directions on that axis.

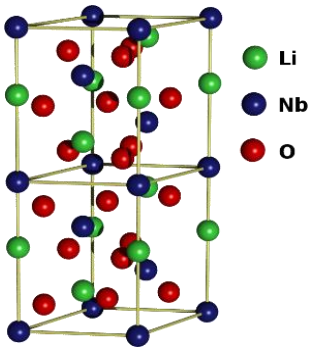


Figure 4, Crystalline structure of $LiNbO_3$ (Adapted from: Ahellwig, Wikimedia Commons, Licence CC-BY-SA-2.0)

A key feature of ferroelectric materials is the PvsE curve showing a hysteresis loop.[7] Like magnets, these materials remember past states, this shape reveals that trait. Following how polarization flips by sweeping voltage helps explain the loop's shape as shown in figure 5:

1. Saturation: As the strong positive E is applied, all internal dipoles align along the field direction, leading to a maximum or saturation polarization.
2. Remanent polarization (P_r): When the external field is removed, the polarization does not return to zero. The material remembers its previous state retaining a remanent polarization, P_r This non-zero polarization at zero field is the physical basis for non-volatile memory (data retention).
3. Coercive field (E_c): To erase this stored polarization, a negative (reversed) electric field must be applied. The specific field strength required to force the material's polarization back to zero is called the coercive field E_c. It represents the "coercive force" needed to flip the state, thus the dipoles.
4. Negative saturation and reversal: As the negative field increases, the dipoles align in the opposite direction reaching negative saturation. The cycle is completed by removing the negative field (leaving a negative P_r) and applying a positive field (requiring a positive E_c) to switch it back.

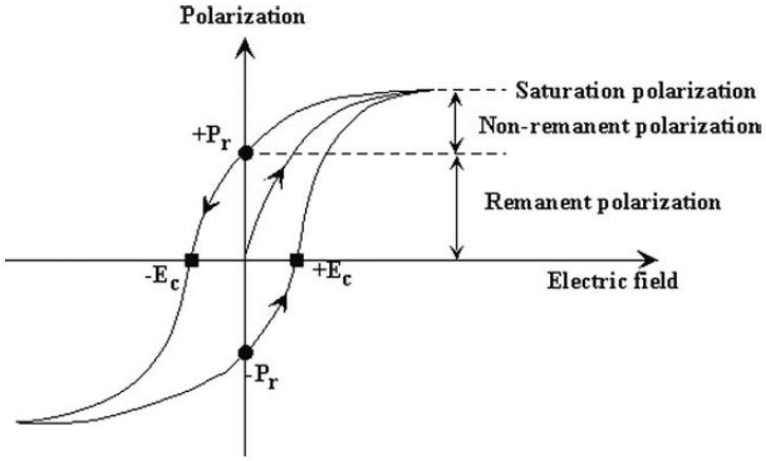


Figure 5 - Characteristic hysteresis loop of ferroelectric materials[16]

The presence of two separate polarization states, positive and negative, at zero electric field enables binary data storage that retains information without power consumption.

The way ferroelectrics switch inside depends on how domains move and influence each other, tiny areas where dipoles point uniformly. An electric field causes these zones to shift or expand resulting in sudden state changes and memory effects. Between differently polarized regions, domain walls exist; their shape, movement, and resistance to displacement strongly affect performance aspects like switching rate, wear-out, and stability.[17] For ultra-thin and nanoscale materials, controlling domain patterns is key when aiming for graded, continuous switching states.

This ferroelectric behavior changes with temperature. When heated beyond a threshold called the Curie point (T_c), the material no longer holds its built-in electric charge, shifting into a paraelectric phase. Such transition defines the highest usable temperature for these materials in applications. $LiNbO_3$ it withstands much higher temperatures (around 1140–1210°C) compared to $BaTiO_3$, which shifts at roughly 120°C.[18]

Ferroelectric materials show other various behaviors, like piezoelectric response, turning electric charges into motion; pyroelectric effect, producing voltage when heated or cooled; along with electro-optical control, applied in light-based data transfer. Such combined traits support many applications and a wide range of devices, from capacitors and oscillators to high-precision sensors, actuators and modulators. For storing data, FeRAM uses switchable polarization states to store information. At the same time, ongoing studies on quantum barriers, gate-controlled switches, and layered material systems seek to enable continuous-level behavior for neuromorphic computing and processing within memory structures.

Material synthesis along with integration continues to evolve. New ferroelectric materials, ranging from lead-free perovskites to 2D systems, are being discovered and designed, broadening options for scalable tech solutions. High-performance thin films are made using methods like CIS, PLD, or sputtering; these allow control over crystal alignment and domain layout depending on device requirements.[19]

Memristors

In neuromorphic systems, the memristor acts like an artificial synapse, enabling physical storage of synaptic weights while supporting adaptive changes. At its core, this element is a simple two-terminal device with resistance that shifts with applied stimulus instead of staying constant. Its value depends on prior electrical activity; past signals shape present performance. The process follows a modified version of Ohm's law tied to internal states. These states shift gradually when exposed to voltage or current inputs of varying strength and/or length[2].

Instead of just on-off levels, memristors handle many conductance steps, this supports dense analog data storage inside their physical setup. At one end, there's the Low Resistive State (LRS); at the other, the High Resistive State (HRS). The LRS, also called ON mode, comes from the SET operation and acts like a stronger neural connection with higher signal efficiency. On the opposite side, the HRS, or high resistance/OFF condition, is formed through RESET, mirroring a weaker link where transmission drops[5]. Because this component can shift between these two stable states accurately, it works as a reliable holder for information.

The way resistance changes works differently depending on the material. Although filament-based systems use moving ions or oxygen gaps to create and break a conductive route switching between low and high resistance states, devices using ferroelectric materials, often adjust Schottky barriers at interfaces or grow and spread polarized regions[5], [20]. Such control makes it possible for the device to mimic how synapses strengthen and support learning behaviors such as spike-timing-dependent plasticity.

Experimental Methodology

The earlier sections showed that neuromorphic computing can effectively address inefficiencies in traditional Von Neumann systems, while also highlighting ferroelectric materials as a strong option for building analog synapses. In this context $LiNbO_3$ stands out due to its unique properties.

The experimental work described here supports a larger project focused on building a new kind of CIM accelerator using adjustable piezoelectric acoustic resonators. Instead of separate components a Y36-cut thin film serves both as memory and as the active material in the resonator, enabling compact signal handling. Due to the resonator's high quality factor (Q), it can switch between two clear impedance levels: extremely high at parallel resonance, exceptionally low at series resonance. Because of these sharp contrasts, power consumption drops significantly during operations like vector-matrix multiplication (VMM), along with its reverse process (IVMM).

To show a tunable piezoelectric resonator, two key conditions shaped the design approach. Firstly, so it can work inside adjustable RF front-end filters, the thin film needs to handle switching voltages compatible with CMOS technology. This happens using a 43 thick, single-crystal Y36-cut layer, which switches fully at low voltage thanks to its low coercive field. Second, instead of just focusing on one aspect, both the useful electromechanical coupling coefficient (k_t^2) and quality factor (Q) of the device are improved, to create strong, reversible shifts in resistance near resonance, making frequency control via polarization more efficient. Also, picking Y36 as the orientation boosts both k_t^2 and Q.[14]

The setup includes a core active region, where vibrations occur, made by placing a layer between two tungsten (W) films. Because of this design, it can hit a target frequency in the FR3 range; small variations from manufacturing lead to a tuned resonance in the range 18 ± 2 GHz. On top and beneath the center part alternating Ti and W layers form Bragg reflectors that trap acoustic energy, boost k_t^2 and Q, while reducing unwanted signals.[21] For electrical contact, heavy aluminum coatings act as terminals, shaped simply to carry signals out.

Compared to standard designs, TBAW (Thin-film Bulk Acoustic Wave) enables vertical signal passage, also offering stable, adjustable impedance changes with little power

waste. Its ability to integrate memory functions alongside efficient computing in one unit, while working with CMOS processes, makes it a strong candidate for future analog CIM systems or adaptable filters.

Still, analyzing the device under AC conditions while developing it into a high-frequency resonator comes later, this falls outside the focus of the present thesis. Though relevant, that stage follows initial groundwork and isn't addressed here.

The work described here forms the basic work for that idea: measuring direct current behavior. Before changing the resonator state, it is essential to show the main part, the 43 nm LiNbO3 layer, which works like an analog memristor.

The next sections outline the test methods applied to confirm memristor operation. The goal was to show that sending repeated DC voltage signals adjusts the device's conductivity across several steady levels and measures how well it mimics biological synapses. The DC results later served as inputs for larger system models using NeuroSim, estimating efficiency metrics like size, power needs, response time, and precision within extensive crossbar networks built from such components.

Fabrication workflow

The fabrication of the LN TBAW starts with a smooth, Y36-cut thin film bonded directly to a silicon substrate, supplied by NGK. After preparing this initial substrate, material layers are added via sputtering onto the lithium niobate top side (fabrication flow depicted in figure 7). A 32nm tungsten layer goes on first, this helps build the cavity structure. Following that, multiple pairs of titanium (77nm) and tungsten (75nm) films are sputtered down one after another. On top, a 1700nm aluminum layer is deposited; above it comes a coating of silicon dioxide SiO_2. These stacked W-Ti layers work both as an acoustic chamber and as a reflector for resonance behavior. Meanwhile, the aluminum contributes to the electrode system assembly. This upper dielectric coat plays a key role: it protects metal parts during later bonding steps.

To build the device stack, benzocyclobutene (BCB) acts as an adhesion layer to move the full stack onto a sapphire substrate. Instead of direct growth, this method forms a BAW structure where the piezoelectric layer sits upright between two metallic contacts. Once transferred, the stack is flipped upside down so the silicon support can be taken off.[19]

A large portion of silicon gets stripped away via reactive ion etching with gases like CF_4 and O_2. What's left behind undergoes a finer removal using XeF_2[22]. Finally, an RCA-1 wash clears leftover residues, leaving a clean, particle-free interface.

The process goes on by depositing fresh layers of W, Ti, and Al, same thickness as before, onto the newly uncovered surface using sputtering (figure 6). Instead of building everything at once, these coatings set up the shape of the device along with its electrical connections via later lithography paired with ion milling, cutting away excess down to the lithium niobate boundary, shaping both upper contact and edge of the resonator. In a comparable way, another round of etching plus design work handles the lower section; here, ion-based removal runs until hitting the initial protective film, this step gives structure to the bottom electrode.

Electrical separation along with signal paths come from adding a SiO_2 spacer over the whole structure. This film isolates conductive parts while guiding sideways connections. Patterning this layer precisely uses reactive ion etching with gas mixtures, creating openings to touch the upper aluminum layer directly plus cutting down into material below so lower metal levels can electrically connect.

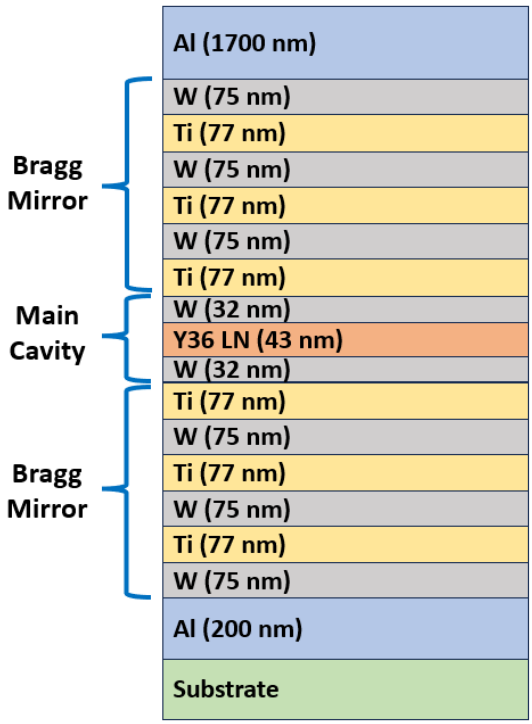
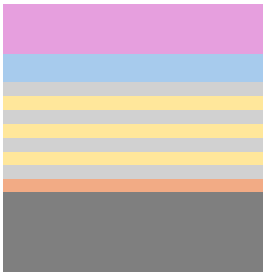


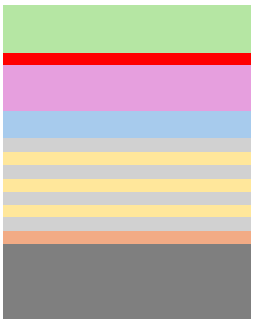
Figure 6 - Schematic of sputtered layers arount the lithium niobate film[23]

An RF sputter etching removes natural oxide layers from open contact zones. Following this, a thick layer of aluminum is deposited across the full surface to finish connections. A photolithography process shapes the metal, then reactive ion etching, using a mixture of Cl_2, BCl_3, and argon cuts precise paths ensuring electrical routing. Once done, the lithium niobate TBAW component, built with detailed acoustic and electronic traits, is complete and prepared for integration and test.

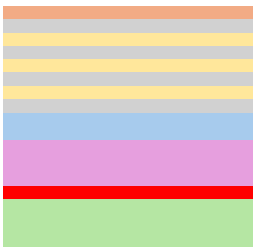
1. Lower Stack Deposition



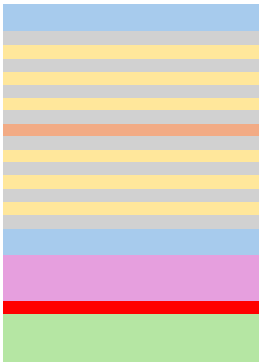
2. Flip Chip Bonding



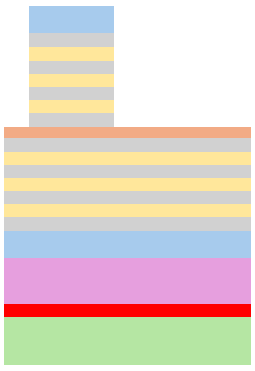
3. Si Substrate Etch



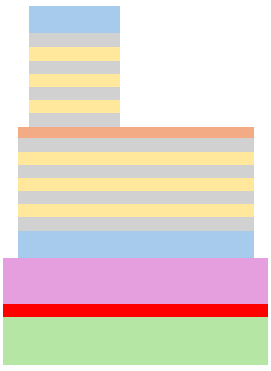
4. Upper Stack Deposition



5. Upper Mirror Etch



6. Bottom Mirror Etch



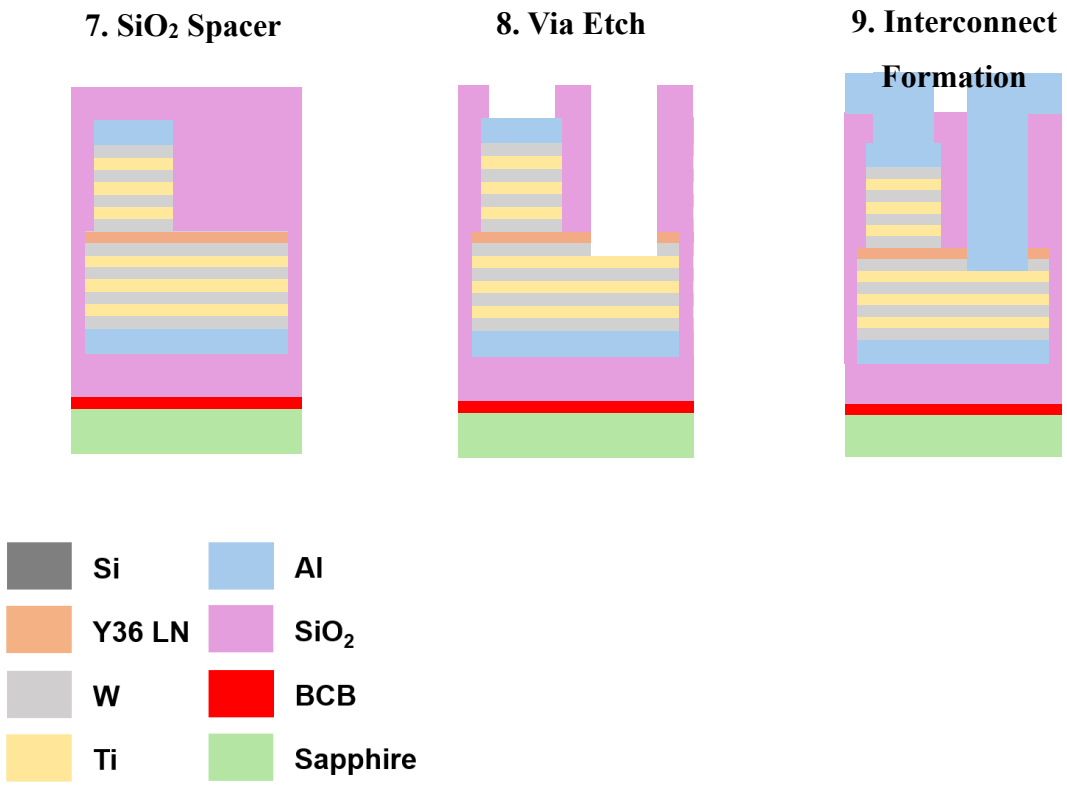


Figure 7 - Fabrication workflow of the entire resonator structure[23]

Conductance states measurements

To analyze the memristive behavior in $LiNbO_3$ thin film, a custom electrical test configuration was used. Measurements took place inside a LakeShore probe station so connections to electrode contacts stayed consistent. The core instrumentation consisted of a Zurich Instruments UHFLI 600 MHz lock-in amplifier for signal processing. Excitation voltages came from the UHFLI's built-in AWG section, delivered via probes to the sample's lower terminal. Its upper contact linked to a low noise transimpedance amp that turned current outputs into measurable voltage levels, then fed them back into the UHFLI for analysis.

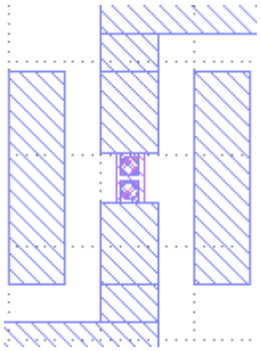


Figure 8 - Top view of a single device layout with metal contacts

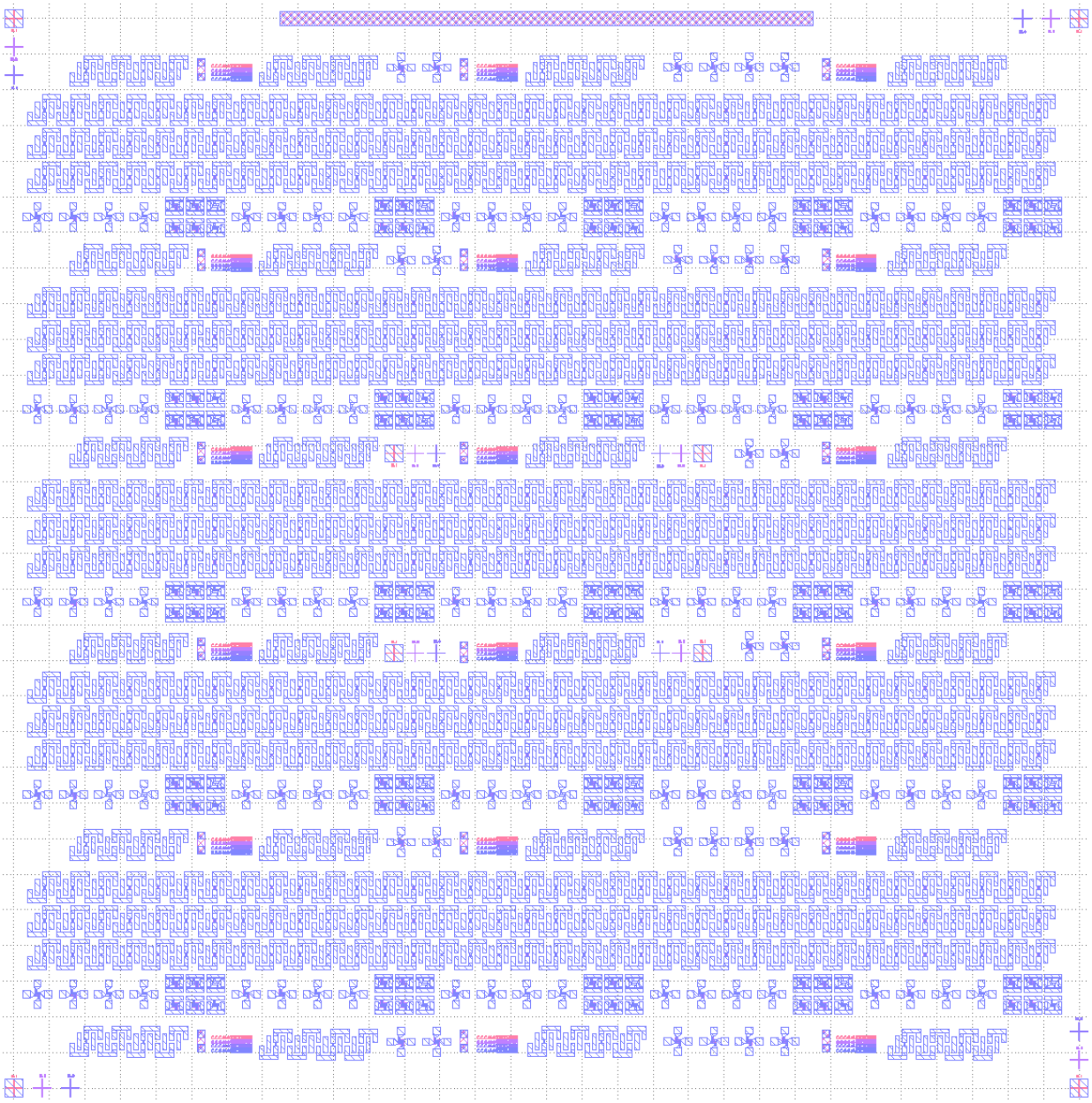


Figure 9 - Full chip layout showing hundreds of devices

Once the tools and setup were set, tests began to check how the device's memory-like behavior worked. To see how the $LiNbO_3$ layer responds across different conditions and capture its full performance span, we used voltage pulse trains with different amplitudes. We first tested write pulses from 0.25 V up to 5 V; later adjusted that range to fit needs for Neurosim modeling, shown in upcoming sections. The process started with potentiation mode, applying rising positive pulses, then shifted into depression mode using negative ones growing stronger (from -0.25 V to -5 V). This way lets us track changes in conductivity while confirming whether the device adjusts its connection strength smoothly, step by step.

The test process carefully alternated programming steps with measurement phases: changing the device state used varying write pulses, whereas a steady 10ms and 250 mV pulse measured conductance right after. A key equipment limitation appeared during these checks. Although writing allowed use of low transimpedance settings, offering wide bandwidth and minimal signal distortion, the reading step faced issues because of the device's high resistance when in the off state. Detecting tiny currents from the $LiNbO_3$ layer's HRS demanded extreme sensitivity; thus, the TIA had to operate at maximum amplification (10^8 V/A). This limited bandwidth imposes a theoretical restriction on how fast the signal can change before distortions start to show up. Applying the standard relationship between bandwidth and rise time

$$t_{rise,min} = \frac{0{,}35}{BW}$$

With this formula the theoretical minimum rise time required to avoid signal distortion would be approximately 29.2 µs. Using a rise time faster than that value could generate large transient displacement currents capable of driving the amplifier into saturation, thereby obscuring the resistive current of the device during the initial moments of the pulse.

To check how well this setup works in practice and find the best timing, a 47 pF capacitor served as a test load, chosen because it matches the capacitance of the real Y-36 $LiNbO_3$ MFM device. Triangular voltage signals with steady 2us rise durations were applied, while switching through every available TIA gain setting so that charging effects could

be separated from memristor-like responses. From these lab results, key values were taken to set up SPICE models comparing an ideal circuit's output with the actual limited-bandwidth performance of the HF2TA amplifier. Then it was found that a slower 50us ramp-up and ramp-down period for reading would be ideal.

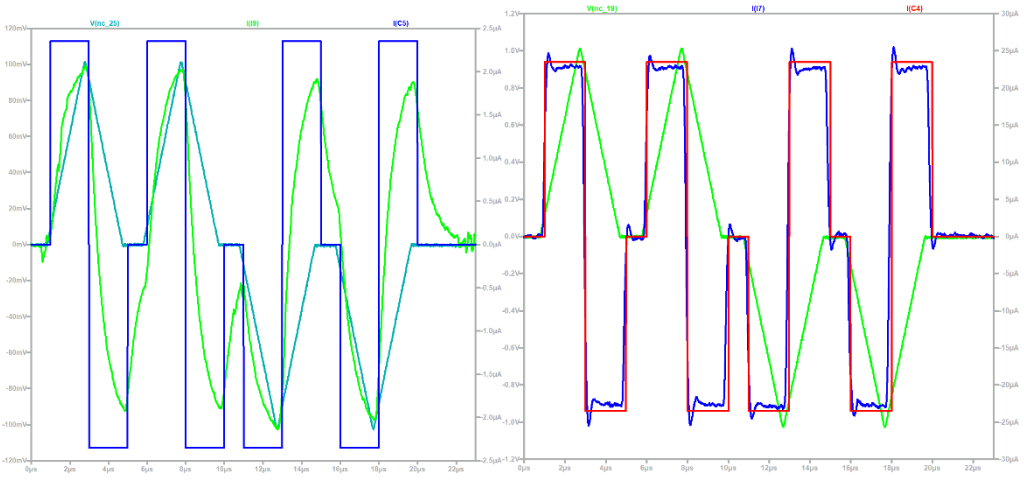


Figure 10 - Spice simulation showing measurements with 1M gain (left) and 100k gain (right). Referring to the left plot, the blue trace is the theoretical current, dark and light green traces are the triangular applied signal and the measured current respectively.

To examine how timing affects steady resistance changes, the study looked at different pulse lengths, specifically 100 microseconds, 1 millisecond, and 10 milliseconds. Instead of fine adjustments, a 0.25V step was used during early tuning to cover the full device range quickly.

Tests showed one specific time worked best for managing multiple states. The current-voltage graph reveals a key shift near the coercive field, marked by a sudden rise in current (more visible in log plot, figure 11 bottom). Still, what happens after depends strongly on how long the pulse lasts: just 10ms pulse duration allowed steady, smooth growth in current when moving from switch point toward higher voltages. Pulses such as 100us failed to stabilize the high conductance states, leading to meaningless outputs right after activation.

The response seen following polarization switching shows what could be a different variety of conduction processes at work. The slow rise in conductivity after the ferroelectric switching suggests additional effects shaped by voltage and duration. For

materials like lithium niobate and related ferroelectrics, researchers have proposed various explanations found across studies:

1. Schottky Barrier Modulation: Polarization charges from the ferroelectric layer affect the Schottky barrier height at the metal interface. When polarization flips, shifts in band bending occur and this can turn the contact from blocking into a conductive mode: as a result, current injection rises slowly over time[24], [25].
2. Oxygen vacancy movement and filament formation: longer pulses could allow enough time for charged defects, like oxygen vacancies, to shift position. Because of this, they might move through a high electric field, creating narrow conducting channels, or small-scale electron jumps[26], over time. As a result, material's inner conductivity slowly rises.[27]
3. Space-Charge-Limited Conduction (SCLC): With extremely thin layers, charge flow may depend on injected carriers surpassing thermally generated levels. A slow increase in current might reflect traps in the energy gap capturing charges, shifting from limited trapping toward full saturation[28].

With this improved setup (flat timing), the unit showed reliable switching, achieving ON/OFF conductivity levels between hundreds and beyond 10^4.

It should be noted that the data shown here, with a 0.25 V increment, is mainly meant to confirm the device's working range and physical response. However, for NeuroSim system simulations, another approach is applied. To boost synaptic precision, which strongly affects neural network performance, voltage steps will become finer within the consistent region found earlier. As a result, more LTP and LTD levels can be created, making better use of the device's continuous capability instead of relying on broad steps used during early tuning.

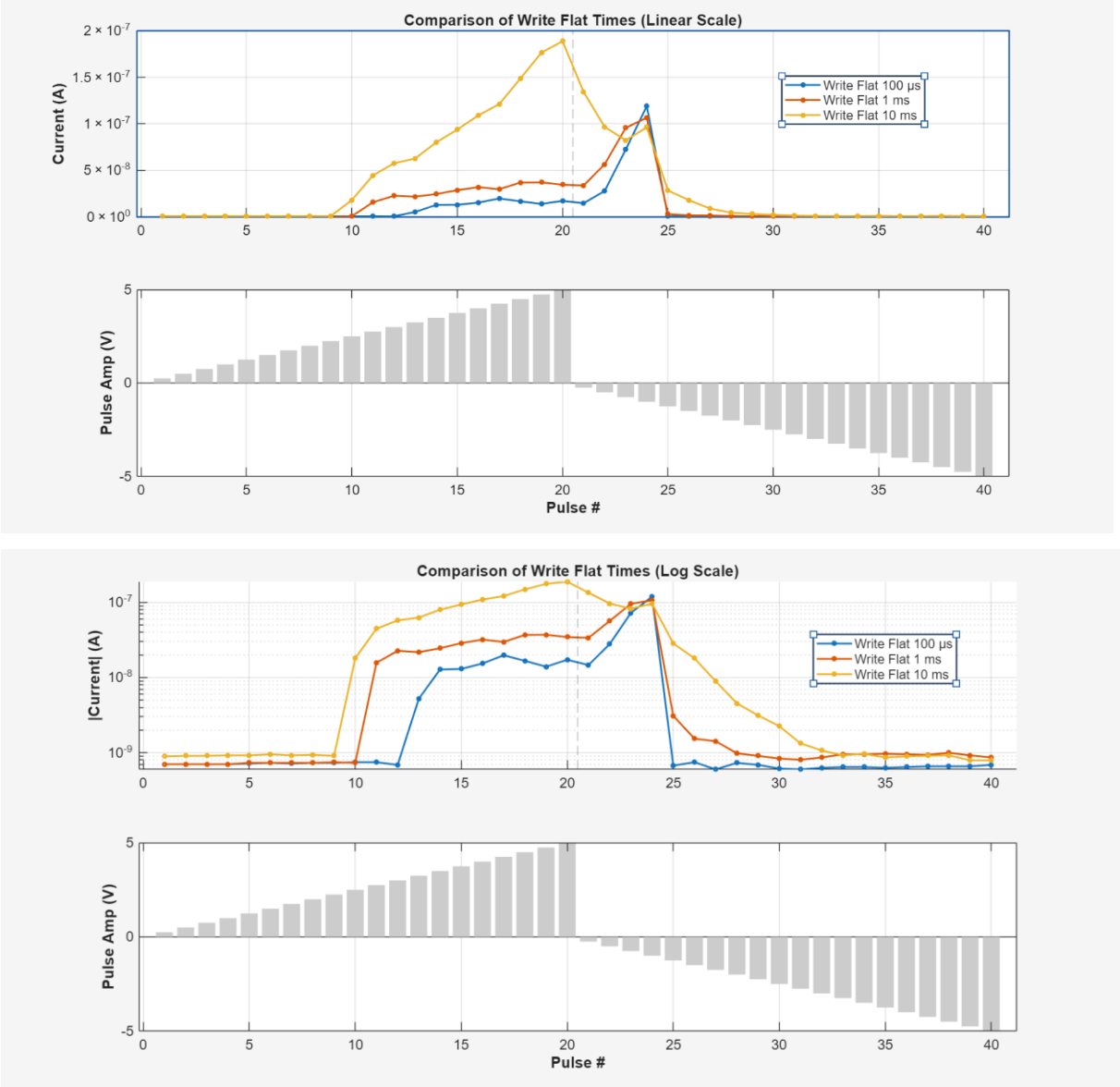


Figure 11 - I vs pulse # measurements at different flat times showing better memristor performance for a 10ms pulse. Linear (top) and log (bottom).

To determine the optimal signal amplitude for state readout, a preliminary screening was conducted on a single test device. Four distinct read voltage levels were evaluated: 10 mV, 50 mV, 100 mV, and 250 mV. Figure 12 shows a comparison between those read voltage amplitudes pinpointing how the 250mV determines the best ON/OFF ratio. However, this test has been performed on a single device and there is not a remarkable distinction between the different the read voltages amplitude tested. Further experiments on a bigger pool of devices in necessary to better understand the mechanism behind this choice.

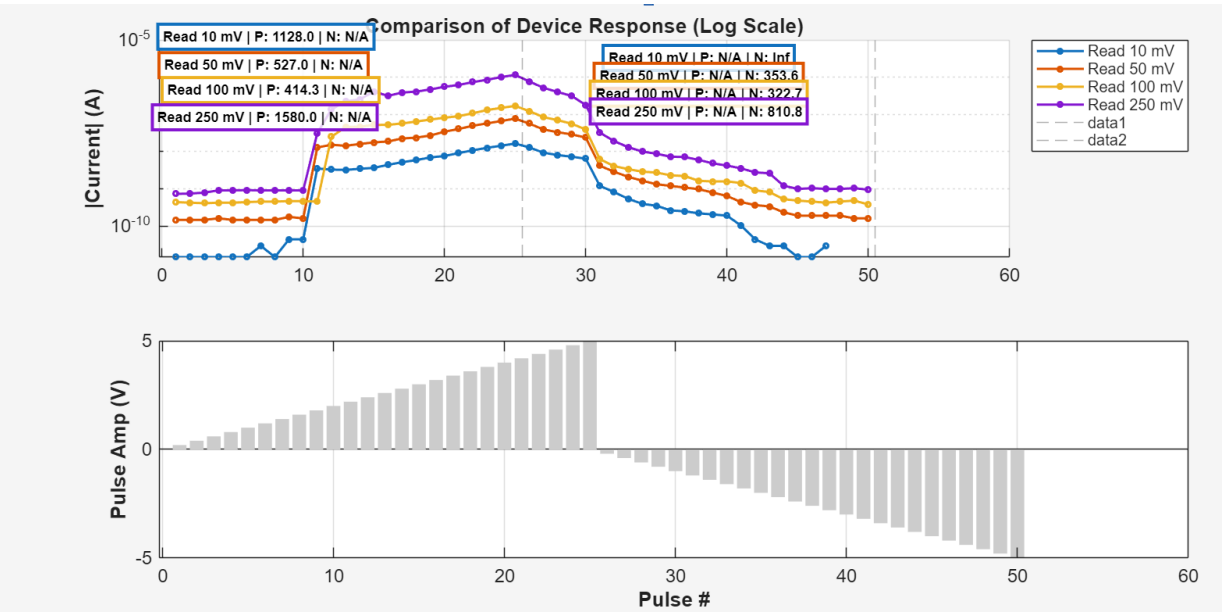


Figure 12 - Comparison between 4 read voltages magnitudes

Following the pulse timing optimization, a set of five distinct devices was characterized to maximize the dynamic range. These measurements employed the final optimized configuration: a 10 ms flat time for both write and read pulses, with a 1 µs rise time for the write pulse and steady increasing voltages amplitude, and a 50 µs rise time for the read pulse with a 250mV constant amplitude. To achieve high-resolution state mapping, the voltage step was refined up to 12 mV.

These revealed a dual operational nature enabling distinct applications based on the chosen voltage range. Digital Memory Regime: When utilizing the full voltage sweep range (2.2 V to 5 V), the device exhibits a massive Physical ON/OFF Ratio of around 4000–6000 (figure 13). This sharp contrast is driven by the abrupt ferroelectric switching event. Such a large resistance margin could be highly advantageous for digital Non-Volatile Memory (NVM) architectures, as it ensures a robust read noise margin, making the device a strong candidate for high-density binary storage applications[5]. Analog Synaptic Regime: For neuromorphic computing, where linearity is critical for training accuracy[8], the operational window is strictly restricted to the post switching region[17], [26]. In this mode the device yields an Effective Analog ON/OFF Ratio between 11 and 15. While smaller than the digital ratio, this range represents a notable improvement over direct literature precedents for this material. Specifically, recent work on identical 43 nm

Y-36 $LiNbO_3$ capacitors (only MFM structure, without the resonator stack) reported a usable analog ratio of only 6[6]. Our optimized pulse protocol has therefore doubled the dynamic range available for synaptic weight updates for $LiNbO_3$. Moreover, this range compares favorably to standard analog candidates used in neuromorphic benchmarks, such as HfO_2-based synapses (typically ratio ~4.4) or TaOx/HfOx stacks (ratio ~10)[8], [13], proving sufficient for neural network training when coupled with appropriate weight update algorithms.

Current levels and efficiency

A critical advantage of this device lies in its ultra-low operating current magnitude. The measured OFF-state current is 730 pA and the maximum ON-state current (at 5 V write and 250 mV read) is 3.1 µA. These values highlight the highly insulating nature of the ultra-thin film and offer a substantial benefit compared to other emerging technologies, for example filamentary devices like RRAMs often require compliance currents in the range of 50 µA to more than 1 mA to maintain stable filaments. Our device operates at currents more than an order of magnitude lower, significantly reducing the IR drop along the bit-lines of large crossbar arrays[29]. FTJs can exhibit ON-currents up to ~100 µA[7] while our device demonstrates better static current efficiency.

Despite the low currents, the experimental energy per write operation appears high due to the specific constraints of the characterization setup. The energy per spike (E_{write}) can be calculated as:

$$E_{\text{write}} = V_{\text{write}} \times I_{\text{on}} \times t_{\text{pulse}}$$

Using the experimental parameters ($V = 5\ \text{V}, I = 3.1\ \mu\text{A}, t = 10\ \text{ms}$), the energy consumption is:

$$E_{\text{exp}} = 5\ \text{V} \times 3.1\ \mu\text{A} \times 10\ \text{ms} = 155\ \text{nJ}$$

This value is significantly higher than mature technologies like Phase Change Memory (PCM), which consumes ~100 pJ/bit. However, this consumption is strictly an artifact of two experimental factors like the macroscopic device

area ($100\ \mu m^2$ since the device is a $10\mu m x 10\mu m$ square) and the 10 ms long pulse width used to stabilize the analog states.

To evaluate the technology's true potential, we project the energy consumption under scaled conditions: the current device area is $10 \times 10\ \mu m = 100\ \mu m^2$. Scaling the device down to nanometric dimensions typical of high-density arrays (e.g., 300×300 nm) reduces the active area to $0.09\ \mu m^2$. The scaling factor (SF) is ≈ 1111 and assuming a constant current density ($J \approx 3.1$ A/cm^2), the operating current drops from 3.1 μA to ~2.8 nA.

With this area scaling alone and still maintaining the 10ms pulse, the energy becomes:

$$E_{\text{area\_scaled}} \approx 5\ \text{V} \times 2.8\ \text{nA} \times 10\ \text{ms} \approx 140\ \text{pJ}$$

This value is already comparable to the energy consumption of PCM technology and demonstrates that the material's intrinsic high resistivity is a key factor for scaling. Hypothetical Temporal Scaling: while a 10ms pulse was necessary in this work in order to get stable conductance states, future optimization and engineering of the material stack and surfaces may allow for faster switching speeds. If the pulse width could be reduced to 100 ns (a standard for integrated memory), the energy consumption would decrease linearly:

$$E_{\text{fully\_scaled}} \approx 5\ \text{V} \times 2.8\ \text{nA} \times 100\ \text{ns} \approx 1.4\ \text{fJ}$$

This projection indicates that the high experimental energy is not a fundamental material limit since, with geometric scaling alone, the device becomes competitive with PCM. Furthermore, if future research accelerates the analog switching dynamics, the device has the potential to reach the femtojoule regime becoming competitive with the most efficient devices.

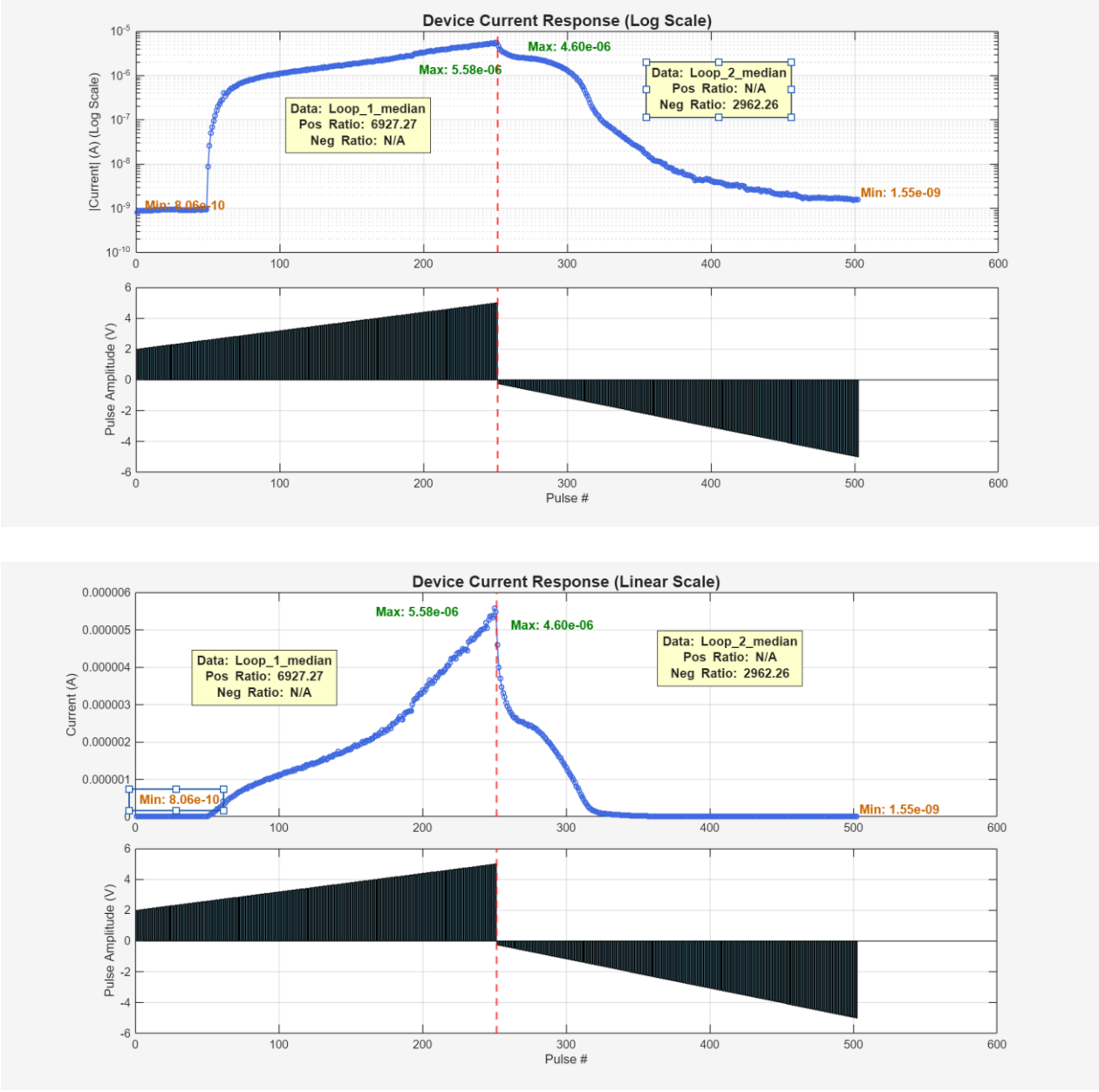


Figure 13 - Log plot (top) and linear plot (bottom) pinpointing the considerable ON/OFF ratio using the previously determined pulsing parameters

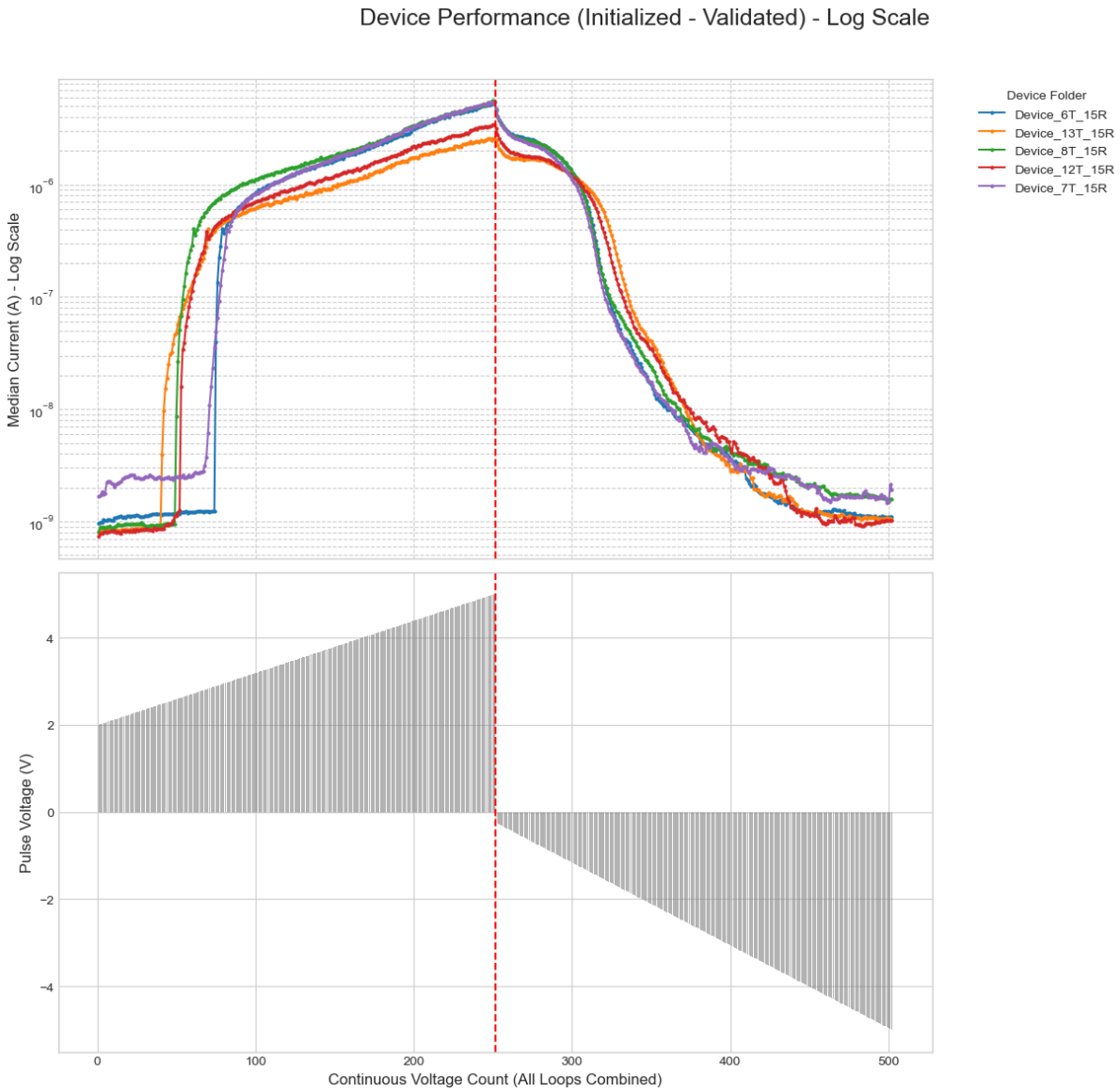


Figure 14 - Plots of all the 5 devices tests overlapped

To find the exact settings for NeuroSim system tests three distinct devices have been tested. This step aimed to record how conductivity changes in the linear range found during tuning. Voltage levels were set precisely to avoid sudden ferroelectric shifts and improve smoothness. During strengthening mode, signals went from 2,2 V up to 5 V while in weakening mode they swept between -0,25 V and -1,85 V in order to ensure the best linearity.

To assess how synaptic precision affects accelerator efficiency, measurements used two voltage steps, 25 mV or 50 mV. Such difference altered the number of available conductance levels noticeably. With the smaller step, resolution improved and this led to

102 LTP states alongside 61 LTD states. In contrast, when applying the larger step, fewer states emerged: only 57 for LTP yet 33 for LTD as shown in figure 15. Notably, limiting the voltage range to the one mentioned brings the ON/OFF ratio down to 11–14.This compromise is necessary, for enabling steady weight adjustments during network learning. Each of the three devices showed similar responses here, supporting the stability of chosen settings.

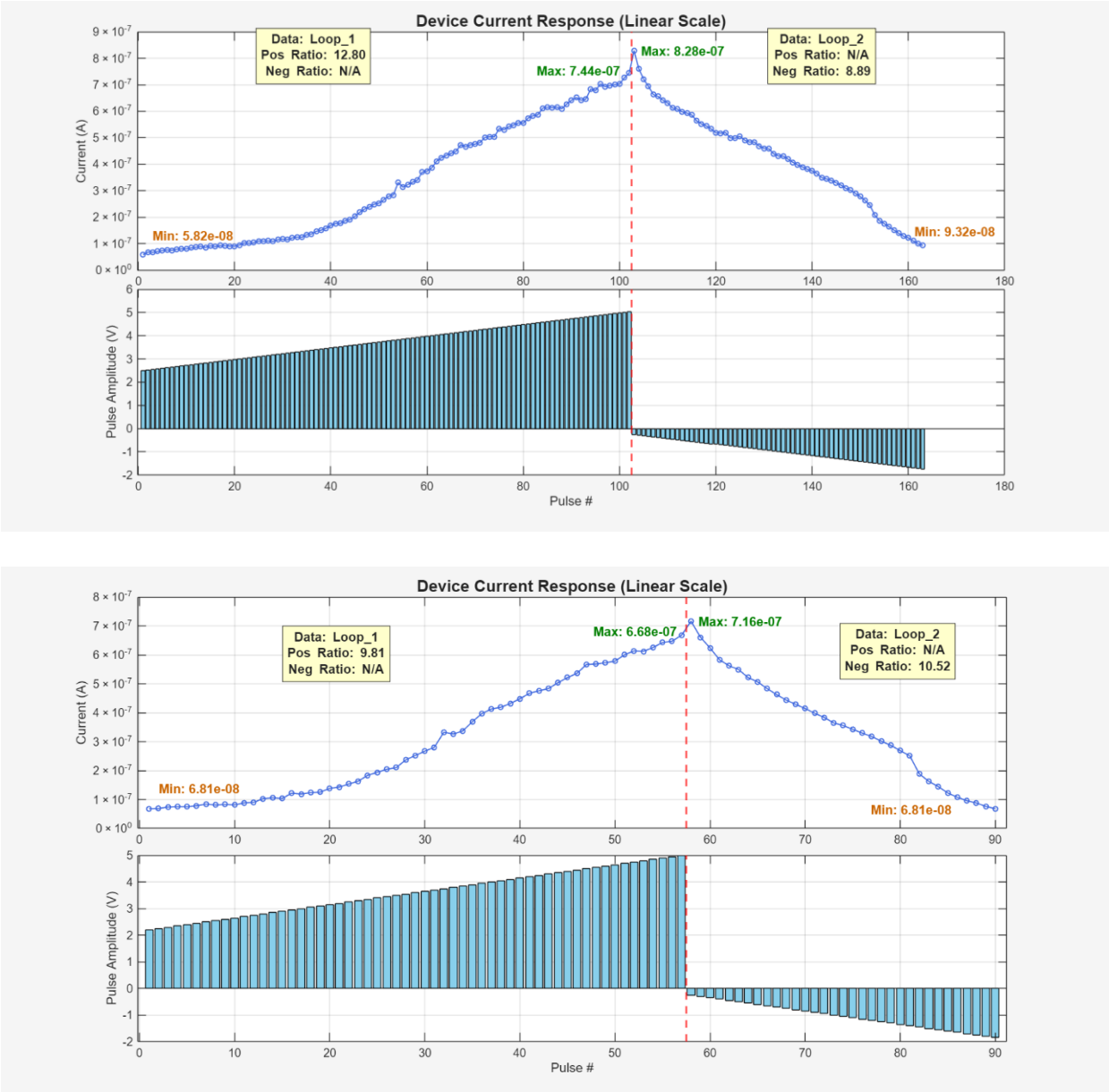


Figure 15 - Higher and lower number of states measurements (top and bottom plot respectively)

In figure 16, slightly displaced ON current values can be observed among the three tested devices even though the OFF current shows better consistency. This will be taken into account when calculating non idealities for the Neurosim measurements.

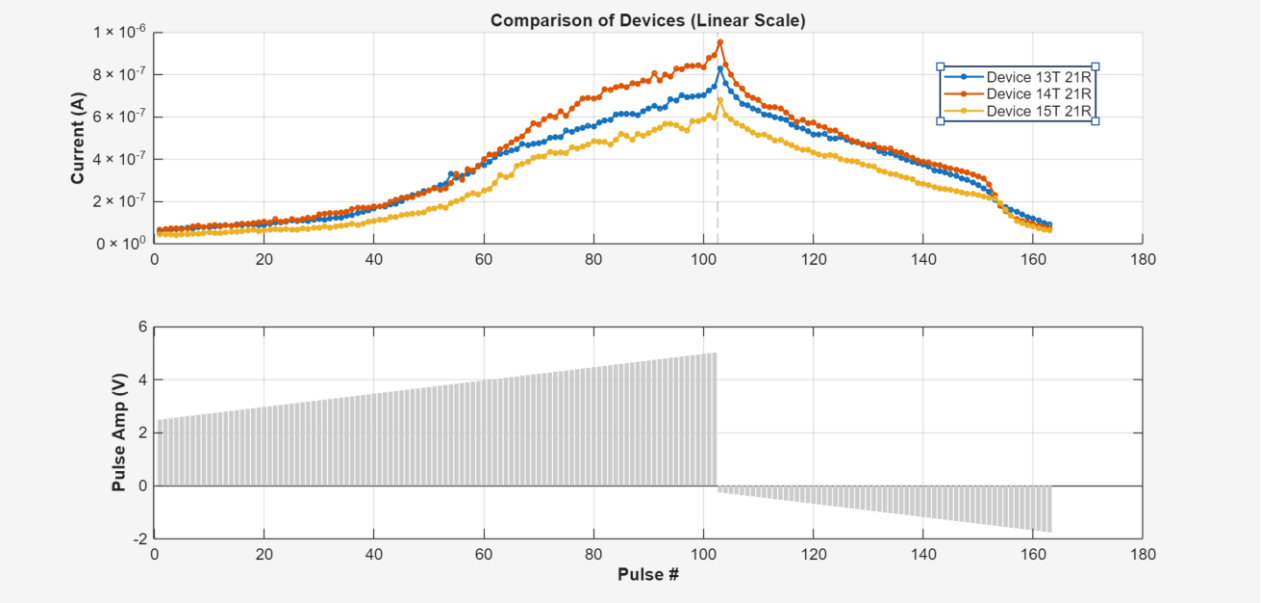


Figure 16 - Comparison between the higher number of states measurements of three devices

Neurosim simulations

Overview

NeuroSim is a tool made to assess how well in-memory and neuromorphic computing systems work at the hardware level. Rather than acting like typical AI simulators, this one looks closely at physical behavior inside the chips. It mimics memory blocks, such as RRAM, SRAM, eNVM, FeFET, and includes flaws that appear in reality, together with needed digital and analog peripheral circuits. Because of this detail, users gain solid estimates not only on speed or power use, but also pinpoint trouble spots like delays, chip area, energy, or side effects from imperfect components during actual neural network execution. For researchers, NeuroSim links lab-measured device data with projected full-system results, serving as an effective tool for initial design testing and practical evaluation of novel neuromorphic ideas.

How Neurosim operates

It accounts for non idealities in analog synapses by applying simple math models based on experimental data. Rather than relying on steady, straight-line changes in conductance values when the devices are excited, NeuroSim lets users add unique response patterns, often curved, that match what's seen in physical hardware, using data fitting.[29]

For LTP and LTD, changes in device conductance across programming pulses are modeled using these formulas:

- Potentiation (LTP):

$$G_{LTP}(P) = B(1 - e^{-\frac{P}{A}}) + G_{min}$$

- Depression (LTD):

$$G_{LTD}(P) = -B(1 - e^{\frac{P-P_{max}}{A}}) + G_{max}$$

where:

- G_{min}: minimum conductance
- G_{max}: maximum conductance
- P: number of applied pulses
- P_{max}: maximum number of pulses (i.e. the number of programmable states)
- A: fitting parameter controlling the curvature (nonlinearity)
- $B = \frac{G_{max} - G_{min}}{1 - e^{-P_{max}/A}}$: scale factor that ensures the conductance swings between the desired range.

Setting those properties requires enabling nonlinearWrite=true in the input code. The degree and asymmetry of nonlinearity are quantified by two parameters: NL_LTP for potentiation and NL_LTD for depression which are obtained from fitting experimental data with a MATLAB code provided by Neurosim.

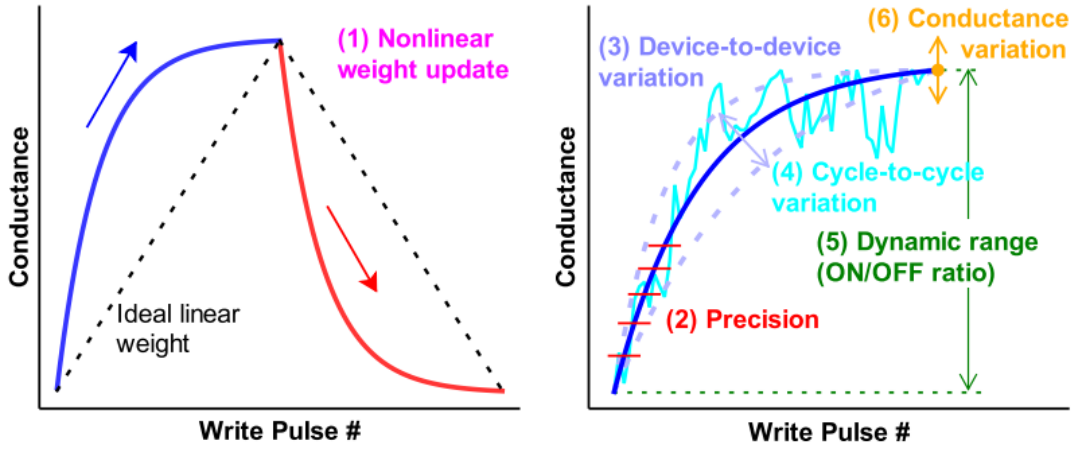


Figure 17 - Schematic of non idealities that is possible to take into account in the Neurosim environment[13]

In place of a linear update, NeuroSim allows extra non-ideal effects to be added; instead of perfect behavior, real world deviations can also be modeled by considering the following non idealities (figure 17):

- Limited accuracy: The total number of stable conductance steps affects how well a system learns. In NeuroSim, these ranges are defined separately, maxNumLevelLTP sets upward changes, whereas maxNumLevelLTD handles downward changes; when using digital hardware, the bit count specified in numWeightBit within Param.cpp (one of the source codes) determines weight detail.
- Device-to-device deviations: actual arrays show variability in conductance across cells, represented by sigmaDtoD: if it is zero, the variation is turned off; when it's not zero, it reflects measured, often normal-distributed discrepancies among cells. A higher value means greater divergence in performance between units. Experimental data usually determines the magnitude of this parameter.
- Cycle-to-cycle variation: the same device might show slightly different conductance after every pulse. This inconsistency within a single device, called intra-device variability, is set by sigmaCtoC, which stands for part of the full conductance range. Setting it to zero turns off such variations; higher values make updates more unpredictable. To stay on the safe side it is necessary to consider the higher value observed during either LTP or LTD phases (worst case scenario).
- Dynamic range (ON/OFF ratio) depends on maxConductance and minConductance, these reflect actual device limits. Since no additional setting is required, the range comes from those two values alone. However, if someone wants to skip this behavior, using minConductance = 0 gives a perfect, unlimited ON/OFF ratio.
- Conductance changes: To mimic state-dependent variations that occur at certain levels. The users can turn this by setting conductanceRangeVar=true while defining maxConductanceVar along with minConductanceVar, these reflect standard deviations at highest and lowest conductance points. When devices show high ON/OFF contrast, setting just maxConductanceVar tends to work well.
- Read noise refers to fluctuations in conductance measurements, especially noticeable with weak signal levels. Is enabled via readNoise=true and the extent

depends on sigmaReadNoise. This value sets the spread of the Gaussian distribution applied. Effects become more apparent under low-conductance conditions. Control requires adjusting this parameter carefully.[13]

The fitting process in Matlab extracts device settings from measurements weight changes for use in simulation. Instead of using raw measurements, the script nonlinear_fit.m adjusts them to match NeuroSim's model equations, based on recorded conductance shifts during strengthening (LTP) or weakening (LTD). Before analysis, data needs some adjustments: align both LTP and LTD starts at pulse zero, flip LTD values left-to-right per fit requirements (figure 18). Although Pmax defines upper state limits, it's scaled to one here, this normalization lets the A factors adjust curve shape accurately when plotted against relative conductance steps.

The fitting procedure happens in two stages: initially, optimal A values for LTP and LTD are identified by ignoring fluctuations, update variability is set to zero; next, actual variation levels, which mimic cycle-to-cycle changes via pseudorandom seeds, are added to refine accuracy. Within the script, these parameters appear as A_LTP and A_LTD. Afterward, the final values, along with conductance range and state count, are extracted and set as an input into NeuroSim as device settings.[13]

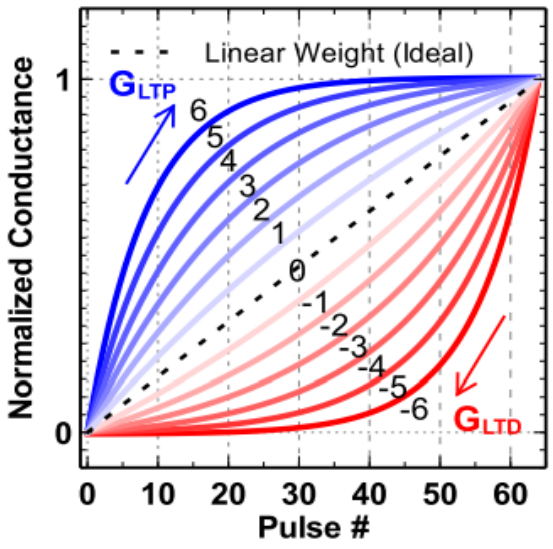


Figure 18 - Visualization of how the A parameter models nonlinearity.[13]

NeuroSim allows multiple methods to model real-world memory setups in synaptic hardware, enabling users to define not only perfect crossbars but also mixed designs such as 1T1R (a type of pseudo-crossbar). Configuration relies on settings like cmosAccess, while resistanceAccess accounts for transistor effects when simulating 1T1R units. Such detailed modeling matters, since physical circuits face issues from unwanted parasitic resistances and switching elements that alter current paths, delay responses, and reduce effectiveness. Different array layouts along with operational approaches are explained further in the following section.

Most significant here is NeuroSim's ability to model varying programming pulses. Enabled via the nonIdenticalPulse setting, it allows separate configuration, such as VinitLTP, VstepLTP, PWinitLTP, for up and down states. These settings adjust amplitude or width differently per pulse as shown in figure 19. Realistic coding often uses such varied signals.

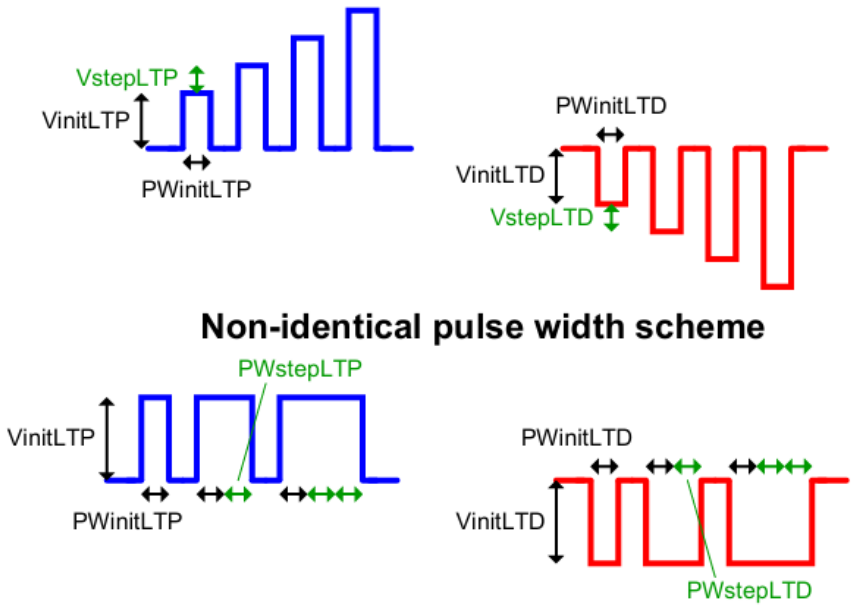


Figure 19 - Non identical pulse scheme[13]

Synaptic Arrays in Neuromorphic Hardware: Crossbar and Pseudo-Crossbar Architectures

In neuromorphic systems, memory and processing rely on synaptic arrays, structured grids linking inputs (axons) to outputs (neurons) using adjustable memory units as weights. Various setups can arrange these components, including dynamic CMOS RAM, content-addressable storage, FeFET-driven blocks However, the crossbar layout along with its 1T1R variant stand out as key models explored for in-memory operations.

Analog eNVM Crossbar Array (1R)

The analog eNVM crossbar array offers a dense, straightforward design for building synaptic weights in neuromorphic systems. At every junction between a word line (WL) and a bit line (BL), one memory unit, like RRAM or PCM is placed (figure 20). Because of this layout, space use is optimal, reaching just $4F^2$ per cell, with F standing for the smallest fabrication dimension. During matrix-vector operations, voltage inputs go across all WLs at once; meanwhile, output currents form on BLs in a parallel fashion. Instead of sequential steps, multiple calculations happen together through analog signals inside memory units.

Yet without separation between memory units, missing selectors or transistors at junctions, issues arise, particularly when updating values. When voltage signals are sent to adjust weights, inactive cells might still face accidental shifts due to leakage currents or stray writes. To lower such risks, a specific voltage setup is applied: unselected WLs and BLs usually rest at half the supply level (V/2), instead of zero, limiting disturbances on idle components. During every update phase, one entire line gets activated, with controlled pulses delivered via bit lines, enabling simultaneous adjustments across multiple locations within that line, the system detects this matrix layout automatically if cmosAccess=false.

Analog eNVM Pseudo-Crossbar Array (1T1R)

To tackle interference from unwanted paths and unintended writes, the 1T1R setup uses a transistor linked in sequence with every storage unit (figure 21), allowing individual selection. This design links the transistor's control terminal to a word line (WL), its input side to a source line (SL), while the upper contact of the memory element attaches to a bit line (BL). Instead of direct routing, the output end connects via an underlying vertical link under the cell. Now, chip space depends more on the transistor footprint, often exceeding $6F^2$ per unit, particularly when greater programming current demands broader switching elements.

A standard 1T1R setup cannot run fully parallel analog computations since access transistors disrupt the inherent grid balance. Instead, the modified crossbar design shifts BLs by 90 degrees ash depicted in figure 20, this activates every transistor along an addressed WL at once, applying inputs through BLs while outputting summed currents via source lines (SLs) in unison.

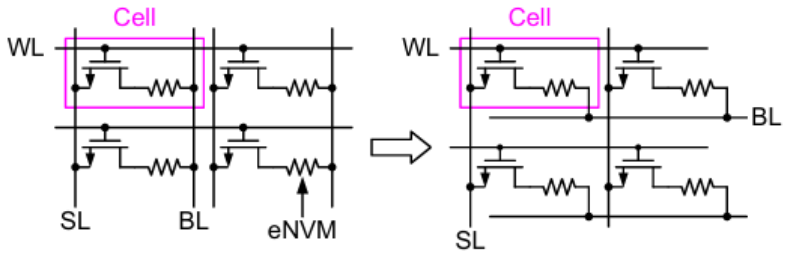


Figure 20 - Rotated BLs ensures a better paralelization of VMMs operations[13]

With weight programming, just the WLs tied to specific rows turn on, activating certain transistors so only designated cells obtain programming voltage. This setup greatly reduces unwanted interactions during writes while boosting energy efficiency over basic crossbar designs. During simulation, selecting cmosAccess=true in the surce code enables this configuration. Other supported architectures in NeuroSim include digital eNVM arrays, SRAM-based arrays and FeFET-based arrays.[29]

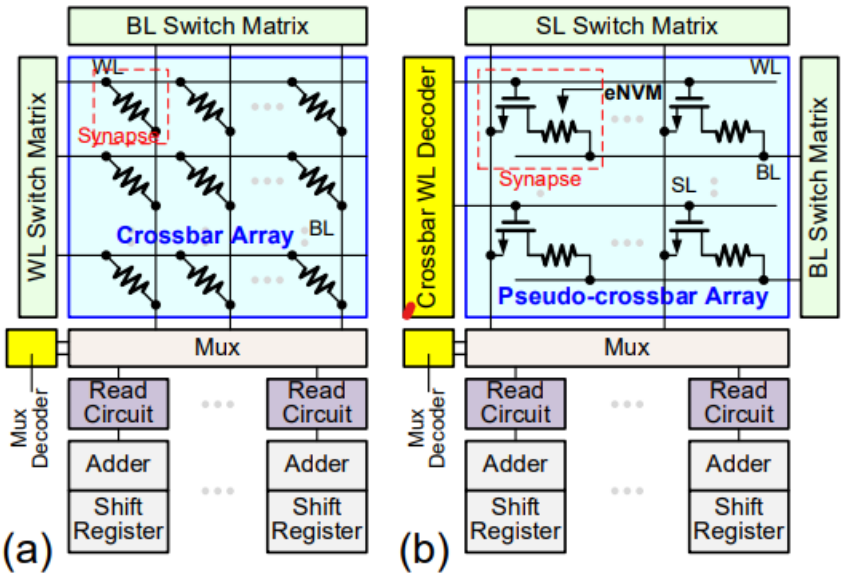


Figure 21 - Schematic of crossbar array (a) and pseudo-crossbar array (b) architectures along with the peripherals.[13]

Array Peripheral Circuits

The peripheral circuits in NeuroSim help run real tasks in synaptic arrays, like crossbar or pseudo-crossbar setups, by making them work efficiently. These components support functionality through control and signal handling across different array types, depending on design needs.

- Switch matrix: In the bit line switch setup, transmission gates, guided by register-stored signals, link BLs either to read voltage or ground. When computing weighted sums, input patterns become control commands, turning on array sections at once. If a higher bit accuracy is needed, inputs roll out across several clock phases instead of relying on analog levels. Source line, bit line, and word line switching units come from the SwitchMatrix template; meanwhile, digital blocks use a unique WL-BL variant for simultaneous data reading.
- Crossbar WL Decoder: it adjusts the classic word line design. Instead of just one path, it uses separate routes to turn on specific lines via address input. When full concurrency matters, like during analog computation, a signal called ALLOPEN triggers every line together. Its layout pairs a standard row decoder with added

follow-up stages. These additions allow dynamic adjustment depending on task needs. During testing, this unit combines specialized decoding blocks with output modules.

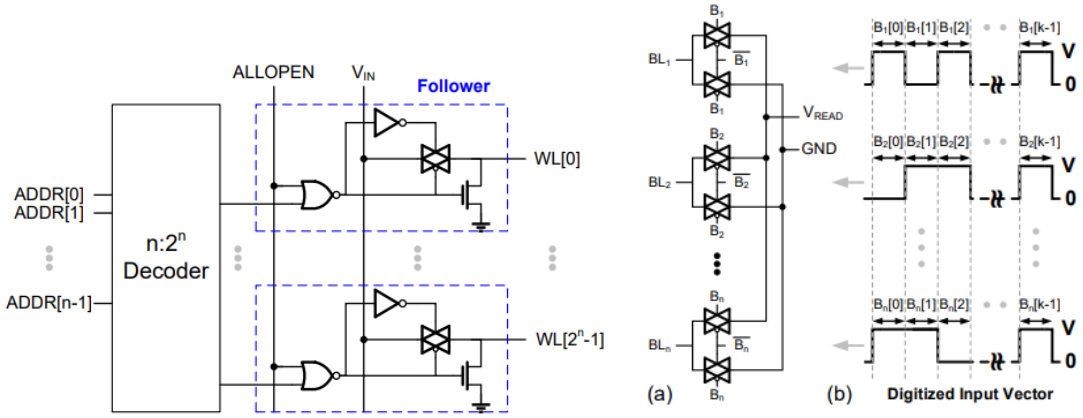


Figure 22 - Crossbar WL decoder (left) and switch matrix (right)[13]

Other Peripherals:

- Multiplexer with Mux Decoder: shares costly readout circuits across array columns reducing space use but adding delay from time-sharing needs. It is possible to tune the degree of resource sharing from the source codes.
- Analog-to-Digital Read Circuit: Once analog processing finishes, a spike-driven read unit converts summed current into digital form, spike frequency reflects input magnitude, setting ADC resolution through dynamic response.
- WL/Column Decoders & Drivers: Standard decoding units pick rows or columns, while driver stages allow accurate voltage setup when writing data.
- Adder, Register, or Shift Register: these components gather, hold, besides manage weighted sums, key when working with multi-bit inputs.

MLP Neural Networks overview

After highlighting the Neurosim's hardware capabilities, in these subsequent paragraphs are presented theory concepts regarding neural networks and how the software assesses training.

Multilayer perceptron (MLP) is a basic type of neural network that includes an input layer, at least one hidden layer, and an output layer, each made up of linked processing units called neurons (figure 23). In this setup, every neuron gets signals from all neurons in the previous layer, where those inputs are scaled by weights; then a bias value is included before combining them. Instead of just adding values, the total passes through a nonlinear function like sigmoid. Because of this transformation step, the model gains the ability to capture intricate patterns, not limited to straight-line trends.[29]

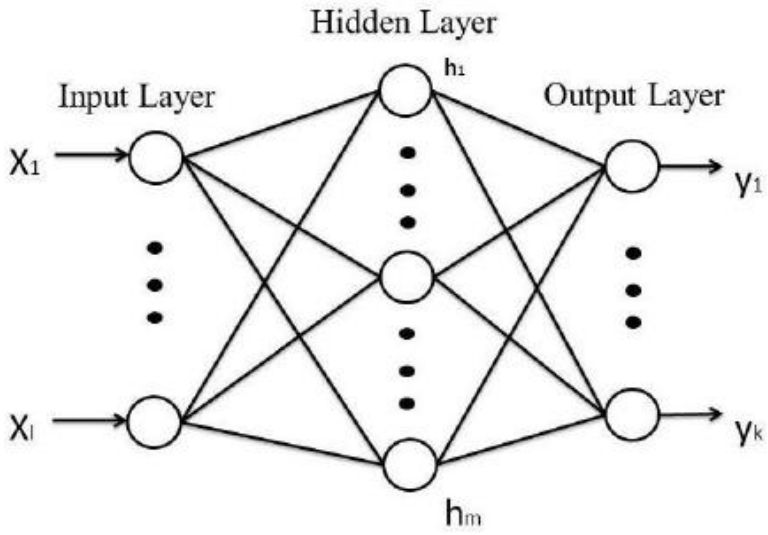


Figure 23 - Simple schematic of an MLP[30]

The functioning of one neuron inside a MLP takes a feature vector, computes weights multiplied by inputs, then adds a bias before passing the result through a non-linear activation function.

More formally, given inputs $x_1, x_2, \ldots, x_n$, each associated with a weight w_i, the neuron calculates the pre-activation output as in the formula:

$$z = \sum_{i=1}^{n} w_i x_i + b$$

where b is the bias. This linear combination is then fed through an activation function $\sigma(z)$ to produce the neuron's final output:

$$a = \sigma(z)$$

Common choices for $\sigma(z)$ include the sigmoid function, which squashes the output to the open interval (0, 1)

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Other activation methods like tanh or ReLU appear frequently, based on how deep the network is and what kind of task it manages.

In MLPs, the simple neuron action scales up through various layers. While one layer works, its neurons take inputs from every output of the prior level, creating full connectivity. By linking many layers that use nonlinear activation, the model can mimic intricate patterns, making it far more capable than linear approaches.

After defining a neuron’s math, those steps easily apply to practical jobs like spotting what digit appears in a picture using collections sets such as MNIST. This dataset holds 70,000 black and white pictures of numbers written by hand; every measuring 28 by 28 pixels and carries a label from 0 up to 9 that represent the “true” value that is the expected to be recognized from the neural network.[8] To get an image ready for use in a basic neural net, the flat square grid turns into a straight-line list holding 784 entries, one per dot on screen. That line-up feeds data into earlier explained processing: individual dots multiply their assigned weights, add offsets afterward, then shift results via simple curves.

In a standard MNIST setup, the MLP uses ten output neurons, one per output digit. These outputs show how strongly each digit is predicted. As data moves forward through layers, predictions are compared to correct labels using a loss measure, then errors flow

backward to adjust weights and biases. Repeating this on many samples helps the model link pixel patterns to actual digits.

MNIST's layout, consistent image size and straightforward labels make it well suited for showing how matrix-driven networks function. Through the math outlined earlier, the MLP turns basic pixel data into correct number identifications, linking theoretical steps to a real-world, familiar task.

The learning process in an MLP involves adjusting the weights and biases in an iterative manner to minimize the error between the predicted output and the actual target label associated to every MNIST dataset training picture. This process typically consists of two phases:

1. Forward Propagation: Input data is fed through the network and layer by layer produces an output. A loss function (for example the Mean Squared Error or Cross-Entropy) calculates the difference between this output and the true target.
2. Backpropagation: The error is propagated backward from the output layer to the input layer. Using the chain rule, the algorithm computes the gradient of the loss function with respect to each weight while these gradients indicate the direction and magnitude by which each weight should be adjusted to reduce the error.

In hardware-based neuromorphic computing, learning methods fall into two types:

- Online learning: the neural net trains straight on the chip. As fresh data comes in, synaptic strengths adjust instantly. Hardware synapses, like memristors, must allow balanced, steady changes. These adjustments follow LTP and LTD rules so training stays stable over time.
- Offline Training (Classification Only): Another option is to train the network separately with precise software tools. After finding the best weights, these values get transferred and set into the physical chip. From that point, the system runs just forward passes to label incoming data. Since weights stay fixed while running, this method reduces demands on how linear or durable the components must be[8].

Once the basic ideas of MLP are set, we turn to how NeuroSim shifts theory into real-world hardware tests. This tool evaluates memory-focused computing systems built for neural nets like MLPs, by mimicking software logic alongside physical device behavior.

In practice, NeuroSim uses a two-layer MLP, counting only hidden and output layers, for adjusted versions of datasets such as MNIST. To improve speed and match real-world devices, it resizes input images to 20×20 pixels and turns them into black and white leading to 400 input neurons instead of 784. Normally, the setup follows 400-100-10: 400 inputs link to 100 hidden nodes; these connect to 10 outputs standing for digit categories[8].

NeuroSim works with both online learning and offline (classification-only) operation. When using online training, it acts like a chip learning in real time, images come in one by one at random, while the system adjusts connections step by step through basic optimization rules. It does not just simulate neurons; it includes physical limits like low-precision values and pulse-driven changes that convert abstract negative-or-positive weights into actual non-negative hardware equivalents. For pre-set use, connection strengths are calculated ahead of time in code, then loaded directly into the model. Once set, these do not change, the circuit runs recognition tasks only, which eases requirements on stability and accuracy.

A main part of NeuroSim's training involves epochs, each meaning a full cycle through the whole dataset, so forward and back propagation. With every round, the model sees all samples one time adjusting its weights step by step. For NeuroSim, users can set how many cycles to run and how many images to use per cycle. These settings appear in files like Param.cpp, where they’re open to changes. This lets users manage training time along with batch setup, while watching how the network improves step by step during the simulation. For instance, using standard values like totalNumEpochs = 125 and numTrainImagesPerEpoch = 8000 (adding up to 1 million images), the tool outputs per-epoch results, giving clear insights into model development and overall system behavior[13].

MATLAB Fitting

The extraction of synaptic device parameters for the simulation was performed using the standard behavioral model integrated into the NeuroSim framework. This model mathematically describes the weight update curve shape (conductance change vs pulse number) using the exponential function seen in the previous paragraphs, defined by the nonlinearity parameters α_{LTP} and α_{LTD}. A fundamental mathematical property of this specific model is that it enforces a constant concavity throughout the entire range; in analytical terms the sign of the second derivative of the fitting function $\frac{d^2G}{dP^2}$ remains constant.

Unfortunately, the experimental characterization of the $LiNbO_3$ devices revealed complex conductance curve dynamics that deviate from this ideal mono curvature behavior. Specifically, the measured LTP and LTD curves exhibit inflection points: regions where the curvature transitions from convex to concave or vice versa. Because the standard NeuroSim model lacks the degrees of freedom characteristic of higher-order functions like polynomial equations necessary to map these inflection points, a perfect fit across the entire synaptic window was hard to obtain, especially for LTP. Consequently, the fitting curves generated by the script represent a "best-fit" approximation that intersects the experimental data at multiple points, capturing the average non-linear trend while inevitably deviating in regions where the intrinsic device curvature inverts as shown in figure 24. Despite this limitation, the extracted parameters provide a representative estimation of the device's average update behavior suitable for the statistical benchmarking of the accelerator.

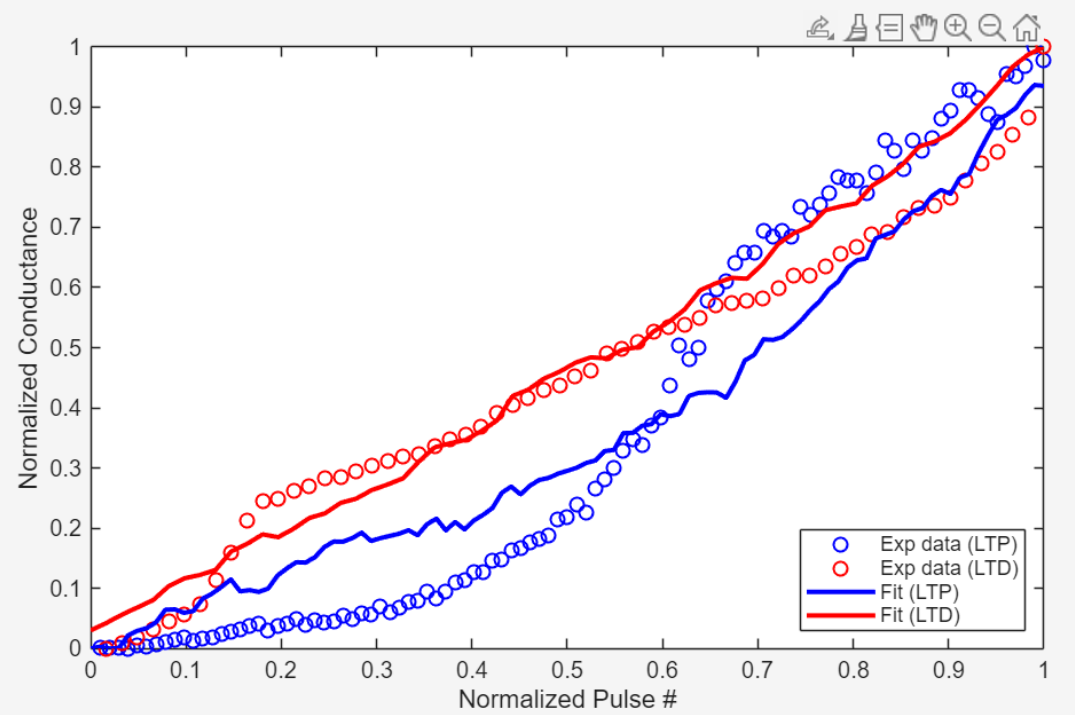


Figure 24 - The performed fitting clearly shows the inflection points and the displacement from the experimental data, especially for LTP

The previously discussed parameters extracted from the experimental data using the MATLAB fitting procedure, are summarized in the following tables showing both the higher and lower number of states measurements tested.

Parameter	Description	Our Value
Nonlinear weight update	**nonlinearWrite=True** (to turn on) Provide **NL_LTP** and **NL_LTD**	**NL_LTP = -3.36 , NL_LTD = -1.58**
Limited precision	Number of conductance states **maxNumLevelLTP** and **maxNumLevelLTD**	**maxNumLevelLTP = 57** **maxNumLevelLTD = 33**
Device-to-device weight update variation	**sigmaDtoD** (standard deviation of d-2-d variation) Set to 0 if not considered	**sigmaDtoD =0.81**
Cycle-to-cycle weight update variation	**sigmaCtoC** (standard deviation of c-2-c variation) Multiplied with **maxCond – minCond** User encouraged to select higher	Var_amp = 0.005 from Matlab
Dynamic range (ON/OFF ratio)	**maxCond** and **minCond** are used, **minCond** = 0 for inf ON/OFF ratio	**Max_G (avg) = 2.69e-6, Min_G (avg) = 2.43e-7** **ON/OFF = 11.07**
Conductance variation	**ConductanceRangeVar** = true, then provide values for **maxConductanceVar** and **minConductanceVar** Standard deviation of conductance at max and min conductance states	**maxConductanceVar (sigma)= 0.41** **minConductanceVar (sigma)= 0.47**
Read noise	**readNoise** = True, then provide the standard deviation of read noise in gaussian distribution	/

Table 1 - Lower number of states measurements parameters

Parameter	Description	Our Value
Nonlinear weight update	**nonlinearWrite=True** (to turn on) Provide **NL_LTP** and **NL_LTD**	**NL_LTP = -1.5 , NL_LTD = -1.29**
Limited precision	Number of conductance states **maxNumLevelLTP** and **maxNumLevelLTD**	**maxNumLevelLTP = 102** **maxNumLevelLTD = 61**
Device-to-device weight update variation	**sigmaDtoD** (standard deviation of d-2-d variation) Set to 0 if not considered	**sigmaDtoD =0.325**
Cycle-to-cycle weight update variation	**sigmaCtoC** (standard deviation of c-2-c variation) Multiplied with **maxCond – minCond** User encouraged to select higher	Var_amp = 0.005 from Matlab
Dynamic range (ON/OFF ratio)	**maxCond** and **minCond** are used, **minCond** = 0 for inf ON/OFF ratio	**Max_G (avg) = 2.98e-6, Min_G (avg) = 2.26e-7** **ON/OFF = 13.19**
Conductance variation	**ConductanceRangeVar** = true, then provide values for **maxConductanceVar** and **minConductanceVar** Standard deviation of conductance at max and min conductance states	**maxConductanceVar (sigma)= 0.59** **minConductanceVar (sigma)= 0.46**
Read noise	**readNoise** = True, then provide the standard deviation of read noise in gaussian distribution	/

Table 2 - Higher number of states measurements parameters

To accurately model the stochastic behavior of the considered devices, three key variability parameters have been calculated: Device-to-Device (DtD) variation, Cycle-to-Cycle (C2C) variation, and conductance variation. All these parameters are represented by the standard deviation (σ) of the experimental data.

Since the experimental dataset consists of a limited number of samples (3 devices), the sample standard deviation formula (using $N-1$) was employed instead of the population standard deviation. This approach is known as Bessel's correction, is essential to correct the bias in the estimation of the population variance when using a small sample size. The formula used is the following:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

where:

- x_i represents each individual data point (for example a nonlinearity value or a conductance value).
- $\bar{x}$ is the sample mean.
- N is the total number of samples.

As an example, σ_{DtoD} was calculated by analyzing the fitting parameters (α_{LTP} and α_{LTD}) extracted from 3 distinct devices. First, the sigma of the nonlinearity parameter was calculated for the LTP curves of the 3 devices (σ_{LTP}) and similarly, the standard deviation was calculated for the LTD curves (σ_{LTD}). Since NeuroSim V3.0 accepts a single input for DtD variation, the final σ_{DtoD} was computed as the average of these two values:

$$\sigma_{DtoD} = \frac{\sigma_{LTP} + \sigma_{LTD}}{2}$$

Simulations results

To evaluate the impact of the experimental device characteristics on neural network, two distinct simulation scenarios were defined and tested on NeuroSim. The first one, denoted as "high states," used the optimized configuration (25 mV step) which yielded 102 LTP and 61 LTD states characterized by moderate nonlinearity ($\alpha_{\mathrm{LTP}} = -1.5$) and controlled variability ($\sigma_{\mathrm{DtD}} = 0.325$). The second scenario, "low states" represented a coarser programming regime (50 mV step) resulting in reduced resolution (57 and 33 states for LTP and LTD respectively), higher nonlinearity ($\alpha_{\mathrm{LTP}} = -3.36$) and higher device-to-device variability ($\sigma_{\mathrm{DtoD}} = 0.81$).

To streamline the process and ensure reproducibility, the standard NeuroSim workflow, which typically requires manual editing of C++ source code for parameter definition was significantly enhanced. A custom Graphical User Interface (GUI) was developed within the Linux environment to directly interface with the simulator core. These tables allow for the rapid setting of all critical simulation parameters without navigating the underlying codebase. Through this tool, physical device metrics like number of conductance states, ON/OFF conductance values, non-linearity coefficients and cycle-to-cycle variability as well as algorithmic parameters (for example number of training epochs, array size, and learning rate) were systematically configured for each scenario.

All simulations were conducted with the cmosAccess parameter set to true, effectively configuring the synaptic array in a 1T1R architecture rather than a passive crossbar 1R. This architectural choice was critical for accurate performance estimation. While passive 1R arrays offer higher theoretical density, they suffer from severe sneak path currents,

unintended leakage paths through unselected cells that degrade read margin and increase power consumption[29].

The images below shows the customized GUI used for the presented simulations.

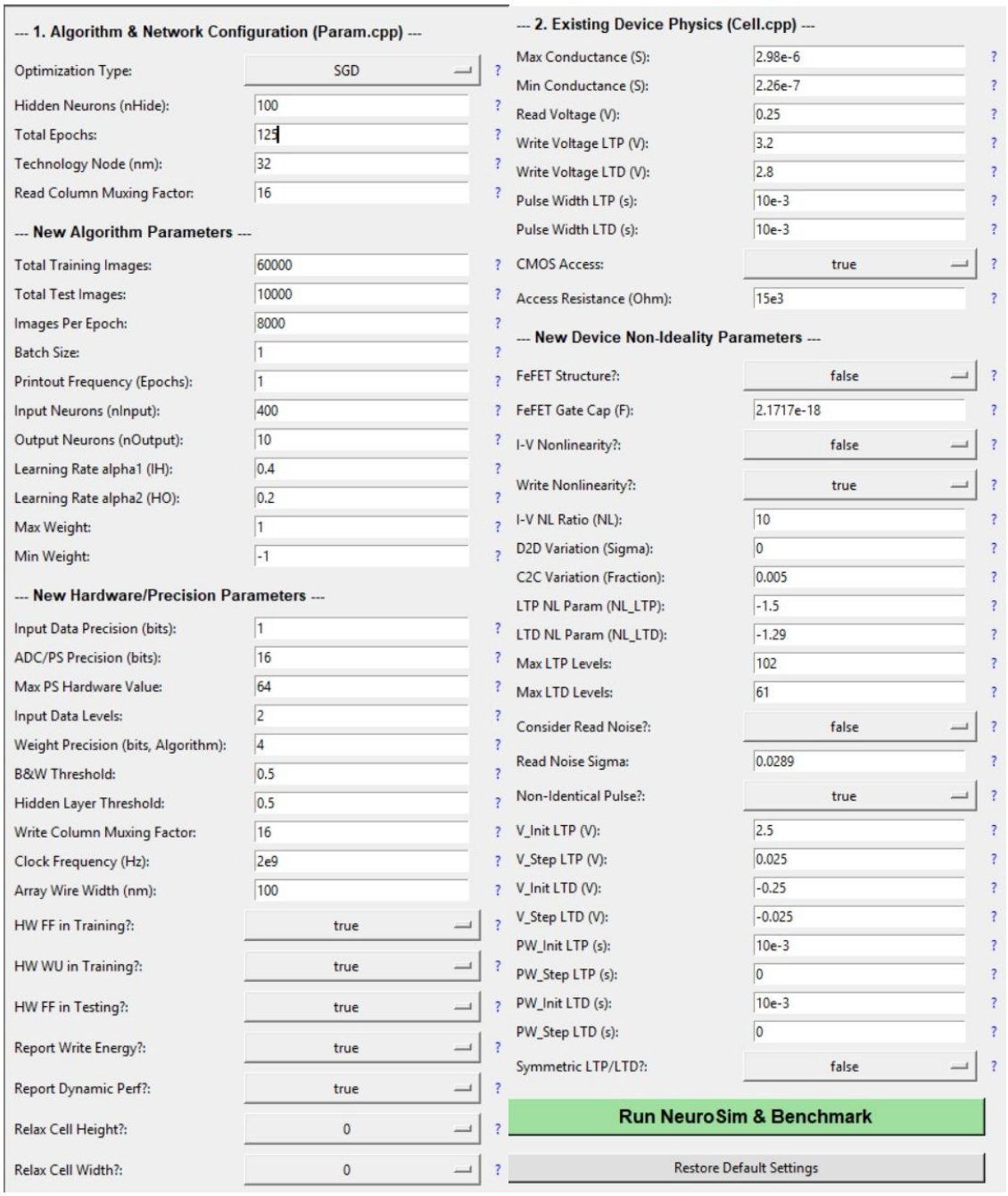


Figure 25 - Tables allowing a fast and error-free parameter setting.

A comparative analysis of the learning curves reveals a critical difference between the higher and lower states measurements. The high states configuration demonstrated robust convergence, with accuracy steadily increasing to a peak of 77.61% at the final epoch. This result confirms that the device's analog modulation, when optimized for linearity ($\alpha_{\mathrm{LTP}} = -1{,}5$) is sufficient to support effective learning. Although a performance gap remains compared to state-of-the-art $LiNbO_3$ benchmarks reaching ~95% of accuracy[26], this may be primarily attributed to the residual nonlinearity rather than a lack of precision. Conversely, the low states configuration exhibited classic signs of training instability driven by hardware nonidealities. While the network initially reached a promising peak accuracy of 67.17% (at epoch 36), the severe non-linearity ($\alpha_{\mathrm{LTP}} = -3{,}36$) and high device-to-device variability ($\sigma_{\mathrm{DtoD}} = 0.81$) caused catastrophic forgetting in subsequent epochs, driving the final accuracy down to about 38%. This drastic divergence underscores that minimizing α and σ is as critical as maximizing the number of states to ensure that weight updates remain bounded and converge towards a global minimum over time.

A key factor enabling the superior performance of the high states scenario is its high synaptic resolution, which merits a specific comparison with existing literature. The optimized configuration yielded 102 distinct conductance states for LTP and 61 for LTD. In terms of digital precision ($\log_2 N$), this corresponds to approximately 6.7 bits for potentiation and about 6 bits for depression. The coarser configuration provided 57 states (LTP) and 33 states (LTD), corresponding to approximately 5.8 bits and 5 bits, respectively.

This level of precision places the 43 nm Y-36 $LiNbO_3$ device in a highly competitive position within the landscape of emerging memories versus standard analog memories: The achieved 6–7 bit precision in the High States mode significantly outperforms typical ferroelectric and resistive devices used in neuromorphic benchmarks. For instance, standard HfO_2-based RRAMs and FeFETs are often limited to 40 states (~5.3 bits) and 32 states (5 bits), respectively.

State-of-the-Art: The resolution is comparable to high-performance Ag:a-Si memristors (97 states) and mature Phase Change Memory (PCM) technology, which typically offers 100-120 states[8].

Comparing instead with previous $LiNbO_3$ Work it can be stated that, most importantly, this result represents a substantial improvement over previous characterizations of the same 43 nm Y-36 $LiNbO_3$ material, which reported a maximum of 40 states (5 bits).

This analysis leads to a fundamental conclusion for system design: while the device possesses a capacity of 6-7 bits required for high-accuracy online learning (which typically demands at least 6 bits), the usability of this precision is strictly gated by the update linearity[8], [31]. The failure of the low states case—despite having a decent almost 6-bit resolution, proves that high precision alone cannot compensate for severe non-linearity and variability. Therefore, the high states configuration is superior not merely because it has more states, but because it successfully pairs high resolution with a linearized update rule that allows the neural network to effectively utilize those states.

The total estimated area for the synaptic core, including the 1T1R array and all peripheral circuitry (e.g., ADCs, Muxes, Shift Adders), is approximately 7.44×10^{-9} m^2 ($7441\ \mu m^2$) for a 128×128 crossbar array. This compact result is a direct consequence of the high-density integration potential of the 1T1R architecture. In standard digital neuromorphic architectures, a 6-bit synapse is typically realized using multiple SRAM cells (e.g., 6–8 cells per weight), occupying a significantly larger area. Benchmarks in the literature (e.g., Chen et al., 2018) indicate that SRAM-based synaptic cores for similar network sizes often consume $> 2 \times$ the area of equivalent eNVM-based cores due to the large cell size ($\sim 150F^2$) of 6T-SRAM compared to the dense 1T1R stack (~ 4–$12F^2$)[29].

Considering digital eNVM: While digital RRAM implementations (using multiple binary cells per weight) are more compact than SRAM, storing information in the multi-level conductance of a single device theoretically offers superior density, limited primarily by the peripheral circuitry required to manage the analog signals.

Concerning write latency and energy constraints, the simulation quantified the high cost of the conservative experimental parameters used for characterization. The total write latency for the full training process was substantial ($\sim 2.7 \times 10^5$ s), and the total write energy was approximately 103 mJ. The root cause can be that these elevated values are not intrinsic to the material but are a direct deterministic consequence of the long 10ms pulse width employed to stabilize the ferroelectric switching during DC testing. Since the total latency is the sum of all write pulses, using a pulse 10^5 times longer than standard

(100 ns) linearly inflates the total time and energy. For comparison, in standard benchmarks (for example Luo et al., 2019), optimized RRAM and PCM architectures typically report training energies in the range of 1–20 mJ and latencies in the order of seconds or minutes. The gap between these benchmarks and the results presented here is partly attributable to the unoptimized pulse width. This suggests that if the device could be engineered to switch reliably at microsecond or nanosecond speeds, the performance would align with or exceed state-of-the-art accelerators.

In contrast to the write metrics, the total read energy remained exceptionally low at approximately 435 µJ for the entire training duration. This metric is critical because read operations (feed-forward passes) constitute the vast majority of operations in deep learning inference. This low read energy confirms the advantage of the device's high resistance and low operating currents ($< 3\ \mu$A). It is significantly lower than the read energy of many filamentary RRAMs, which often suffer from high current read-out, and is competitive with low-power FTJ implementations. This distinct asymmetry between write and read energy consumption may suggests that while the current iteration of the device faces challenges for real-time on-chip training it is an immediate and highly promising candidate for inference-only applications (where weights are programmed once and read frequently) or for architectures where the pulse width can be successfully scaled down to the microsecond regime, potentially reducing the energy-delay product by several orders of magnitude.

	Analog eNVM synapses						Analog FeFET	Digital synapse		
Device type	**Ag:a-Si [1]**	**TaOx/HfOx [2]**	**PCMO [3]**	**AlOx/HfO$_2$ [4]**	**GST PCM [5]**	**EpiRAM (Ag:SiGe) [6]**	**HZO FeFET [7]**	**6-bit SRAM**	**6-bit STT-MRAM**	
# of conductance states	97	128	50	40	100-120	64	32	--	--	
Nonlinearity (weight increase/decrease)	2.4/-4.88	0.04/-0.63	3.68/-6.76	1.94/-0.61	0.105/2.4	0.5/-0.5	1.75/1.46	--	--	
R_{ON}	26 MΩ	100 KΩ	23 MΩ	16.9 KΩ	4.71 KΩ	81 KΩ	559.28 KΩ	--	3.5 KΩ	
ON/OFF ratio	12.5	10	6.84	4.43	19.8	50.2	45	--	2.3	
Weight increase pulse	3.2V/300µs	1.6V/50ns	-2V/1ms	0.9V/100µs	0.7V (avg.)/ 6µs	5V/5µs	3.65V (avg.)/ 75ns	--	1V/10ns	
Weight decrease pulse	-2.8V/300µs	1.6V/50ns	2V/1ms	-1V/100µs	3V (avg.)/ 125ns	-3V/5µs	-2.95V (avg.)/ 75ns	--	1V/10ns	
Cycle-to-cycle variation (σ)	3.5%	3.7%	<1%	5%	1.5%	2%	<0.5%	--	--	
Online learning accuracy	~72%	~80%	~33%	~20%	89%	92%	88%	~94%	~94%	~94%
Area	6292.3 µm^2	8663.1 µm^2	6292.4 µm^2	21760 µm^2	46565µm^2	9144.3µm^2	7032.6µm^2	65728 µm^2	70254 µm^2	66632 µm^2
Latency (optimized)	31997s	10.15s	12218s	470.42S	203.0s	229.6s	2.73s	5.98 s	6.9s (parallel)	90.1s (row-by row)
Energy (optimized)	13.44mJ	4.01mJ	2.53mJ	15.26mJ	35.0mJ	31.01mJ	1.9mJ	15.56 mJ	0.1467J	0.1462J
Leakage power	105.65µW	105.65µW	105.65µW	105.65µW	105.65µW	105.65µW	105.65µW	2.80 mW	124.8 µW	84.0 µW

Table 3 – Benchmark table comparing different technologies[8]

```
Total SubArray (synaptic core) area=5.6126e-09 m^2
Total Neuron (neuron peripheries) area=1.8285e-09 m^2
Total area=7.4410e-09 m^2
Leakage power of subArrayIH is : 7.8864e-05 W
Leakage power of subArrayHO is : 1.7306e-05 W
Leakage power of NeuronIH is : 2.3451e-05 W
Leakage power of NeuronHO is : 3.6160e-06 W
Total leakage power of subArray is : 9.6170e-05 W
Total leakage power of Neuron is : 2.7067e-05 W
Accuracy at 1 epochs is : 67.25%
        Read latency=3.2676e-02 s
        Write latency=1.8345e+03 s
        Read energy=3.4785e-06 J
        Write energy=7.0567e-04 J
Accuracy at 2 epochs is : 63.80%
        Read latency=6.5351e-02 s
        Write latency=3.8847e+03 s
        Read energy=6.9575e-06 J
        Write energy=1.4908e-03 J
```

```
Accuracy at 125 epochs is : 77.61%
        Read latency=4.0844e+00 s
        Write latency=2.7295e+05 s
        Read energy=4.3491e-04 J
        Write energy=1.0309e-01 J
```

```
Total SubArray (synaptic core) area=5.6126e-09 m^2
Total Neuron (neuron peripheries) area=1.8285e-09 m^2
Total area=7.4410e-09 m^2
Leakage power of subArrayIH is : 7.8864e-05 W
Leakage power of subArrayHO is : 1.7306e-05 W
Leakage power of NeuronIH is : 2.3451e-05 W
Leakage power of NeuronHO is : 3.6160e-06 W
Total leakage power of subArray is : 9.6170e-05 W
Total leakage power of Neuron is : 2.7067e-05 W
Accuracy at 1 epochs is : 60.34%
        Read latency=3.2676e-02 s
        Write latency=2.1873e+03 s
        Read energy=3.4793e-06 J
        Write energy=7.6479e-04 J
Accuracy at 2 epochs is : 50.47%
        Read latency=6.5351e-02 s
        Write latency=4.3056e+03 s
        Read energy=6.9593e-06 J
        Write energy=1.4680e-03 J
```

```
Accuracy at 125 epochs is : 38.04%
        Read latency=4.0844e+00 s
        Write latency=2.2388e+05 s
        Read energy=4.3505e-04 J
        Write energy=6.5206e-02 J
```

Figure 26 - Snapshots showing neurosim logs after simulating both the high states (top) and low states case (bottom)

Comparing the hardware metrics of the two simulation scenarios reveals a direct trade-off between synaptic resolution and computational efficiency. The low states configuration, characterized by coarser voltage steps (50 mV) and fewer conductance states (57 LTP / 33 LTD), resulted in a lower total write energy (~65.2 mJ) and shorter write latency (~2.24e5 s) compared to the optimized scenario. Specifically, the high states configuration (102 LTP / 61 LTD) required approximately 58% more energy (about 103.1 mJ) and 22% more time to complete the training workload. This increase is intrinsic to the higher resolution: finer voltage steps (25 mV) imply that a larger number of discrete pulses is required to drive the synaptic weight across its full dynamic range. Consequently, while the high states device presents a higher cost in terms of latency and energy, this is strictly necessary to provide the finer weight updates required for the network to converge stably (77.61% accuracy). In contrast, the energy savings of the low states case are effectively invalidated by its failure to maintain learning stability. Conversely, the read energy and leakage power remained identical across both scenarios (~435 uJ and 123 uW, respectively), as these metrics are determined by the array architecture and read voltage, which were kept constant during simulation.

Simulations conclusions

The simulations conducted via NeuroSim provide a comprehensive validation of the 43 nm Y-36 $LiNbO_3$ device as a competitive synaptic element for neuromorphic computing. The benchmarking results highlight the success of the optimized high states configuration, which demonstrated decent on-chip learning capabilities by achieving a stable final accuracy of 77.61%. This performance confirms that the experimentally optimized pulse protocol successfully accesses a synaptic window capable of supporting effective weight convergence preventing the training divergence observed in the coarser low states case. Crucially, the achievement of 102 distinct LTP states represents a significant advancement in synaptic resolution, comparing favorably with state-of-the-art analog memories such as Ag:a-Si (about 97 states) and Phase Change Memory (about 100 states)[8], and substantially outperforming previous reports on identical $LiNbO_3$ capacitors which were limited to 5 bit precision[6]. From a hardware perspective, the study establishes a distinct difference between the device's write and read performance

that underscores its potential for specific applications. A standout feature is the ultra-low read energy consumption (about 435 uJ total) driven by operating currents in the low microampere range (<3 uA). This efficiency significantly outperforms standard filamentary RRAMs, which typically require higher currents for stable operation, and positions the $LiNbO_3$ synapse as an ideal candidate for inference-heavy workloads where read operations dominate the power budget[5]. While the current write energy (about 103 mJ) is high this is strictly a consequence of the conservative 10ms pulse width used for characterization rather than a fundamental material limit. Theoretical scaling projections indicate that by scaling the device dimensions to the nanometer regime and optimizing the switching speed to nanosecond regimes, the write energy can be reduced by orders of magnitude potentially reaching the femtojoule regime. Overall, this work confirms that the 43 nm Y-36 LiNbO3 successfully combines the high endurance of ferroelectrics with the multi-level tunability required for neuromorphic computing, holding the potential to evolve into a high-density, ultra-low-power solution that rivals current state-of-the-art in-memory computing technologies.

During the preparation of this work, the author used generative AI tools for grammar refinement[32]

Conclusions

The exponential growth of data-centric workloads, driven primarily by Artificial Intelligence (AI) and Machine Learning (ML), has exposed the fundamental inefficiencies of the traditional von Neumann architecture. The physical separation between processing units and memory results in a data movement bottleneck that severely constrains both latency and energy efficiency. In this context, Compute-in-Memory (CIM) has emerged not merely as an optimization, but as a necessary change in basic assumptions, requiring novel hardware substrates capable of merging storage and computation. This thesis has explored the potential of ferroelectric materials to address this challenge, specifically validating ultra-thin (43 nm) Y-36 Lithium Niobate $LiNbO_3$ as a high-performance analog synaptic element.

Summary of Contributions and Findings

The primary objective of this work was to experimentally validate the memristive behavior of Y-36 $LiNbO_3$ in the DC regime and to benchmark its system-level potential. Through a rigorous experimental characterization using a Metal-Ferroelectric-Metal (MFM) architecture, this study demonstrated that polarization of 43 nm $LiNbO_3$ films can be modulated to achieve a significant amount of multi-level conductance states suitable for analog computing. The experimental results highlighted several key achievements. By optimizing the pulse programming protocol specifically through the use of incremental amplitude modulation with fine voltage steps (25 mV), it was possible to stabilize 102 distinct conductance states during potentiation (equivalent to >6-bit precision). This resolution represents a significant advancement over previous reports on similar materials, which were typically limited to approximately 40 states (5-bit). Furthermore, the device exhibited exceptionally low operating currents $< 3\,\mu$A and a robust physical ON/OFF ratio exceeding 10^3. While the effective dynamic range for analog operation was restricted to a ratio of about 15 to ensure linearity, this value proved sufficient for neural network inference tasks, comparing favorably with standard oxide-based analog memristors. To bridge the gap between device physics and system-level application, these experimental parameters were integrated into NeuroSim benchmarking framework. The simulations offered critical insights into the requirements for on-chip

learning and the comparative analysis revealed that synaptic resolution alone is insufficient; the linearity of the weight update trajectory is the decisive factor for training stability. The optimized high states configuration (102 states) achieved a stable learning accuracy of 77.61% on the MNIST dataset, whereas the coarser low states configuration suffered from catastrophic forgetting despite initial convergence. This underscores that future device engineering must prioritize the linearization of the potentiation and depression. From an energy perspective, the study identified a distinct division. The device demonstrated ultra-low read energy consumption, benefiting from the material's high intrinsic resistance. This makes the platform immediately competitive for inference-heavy edge applications. Conversely, the write energy and latency were found to be elevated, a result strictly attributable to the conservative 10ms pulse width used to ensure experimental stability. However, theoretical scaling projections indicate that this is not an intrinsic material limitation. By scaling the device area to the nanometer regime and optimizing interface dynamics to support nanosecond pulses, the write energy has the potential to decrease by orders of magnitude, reaching the femtojoule regime.

Future Work

The validation of the DC memristive properties of Y-36 $LiNbO_3$ presented in this thesis constitutes the foundational step for the broader FEMA (FErroelectric Memory embedded in Acoustic resonators) project. Future research efforts should focus on some specific directions: a priority for immediate follow-up work is the reduction of the programming pulse width. Investigating the transient switching dynamics at the microsecond and nanosecond scales is essential to validate the theoretical projections of low-energy writing and to enable real-time on-chip training. At the same time, scaling of the device active area via advanced lithography will be crucial to verify that the current density remains stable at deep sub-micron dimensions. Linearity Engineering: To bridge the gap between the achieved accuracy (77%) and the ideal software baseline (>96%), the non-linearity of the weight update must be improved. This could be addressed through interface engineering to homogenize the field distribution or by implementing advanced write verification schemes in the peripheral circuitry.

Bibliography

[1] A. Lu, X. Peng, W. Li, H. Jiang, e S. Yu, «NeuroSim Simulator for Compute-in-Memory Hardware Accelerator: Validation and Benchmark», *Front. Artif. Intell.*, vol. 4, giu. 2021, doi: 10.3389/frai.2021.659060.

[2] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, e W. Lu, «Nanoscale Memristor Device as Synapse in Neuromorphic Systems», *Nano Lett.*, vol. 10, fasc. 4, pp. 1297–1301, apr. 2010, doi: 10.1021/nl904092h.

[3] M. Parto *et al.*, «Ultrafast neuromorphic computing with nanophotonic optical parametric oscillators», 28 gennaio 2025, *arXiv*: arXiv:2501.16604. doi: 10.48550/arXiv.2501.16604.

[4] C. G. Kibebe e Y. Liu, «LiNbO3-based memristors for neuromorphic computing applications: a review», *Front. Electron. Mater.*, vol. 4, mar. 2024, doi: 10.3389/femat.2024.1350447.

[5] T. Shi, R. Wang, Z. Wu, Y. Sun, J. An, e Q. Liu, «A Review of Resistive Switching Devices: Performance Improvement, Characterization, and Applications», *Small Struct.*, vol. 2, fasc. 4, p. 2000109, 2021, doi: 10.1002/sstr.202000109.

[6] L. D. Hurtado e G. Piazza, «43nm Ferroelectric Y-36 $LiNbO_3$ Multi-State Conductance with Low Coercive Field for In-Memory Computing», in *2025 IEEE International Symposium on Applications of Ferroelectrics (ISAF)*, Graz, Austria: IEEE, lug. 2025, pp. 1–4. doi: 10.1109/ISAF61233.2025.11234978.

[7] A. Chanthbouala *et al.*, «A ferroelectric memristor», *Nat. Mater.*, vol. 11, fasc. 10, pp. 860–864, ott. 2012, doi: 10.1038/nmat3415.

[8] Y. Luo, X. Peng, e S. Yu, «MLP+NeuroSimV3.0: Improving On-chip Learning Performance with Device to Algorithm Optimizations», in *Proceedings of the International Conference on Neuromorphic Systems*, in ICONS '19. New York, NY, USA: Association for Computing Machinery, lug. 2019, pp. 1–7. doi: 10.1145/3354265.3354266.

[9] X. Peng, S. Huang, Y. Luo, X. Sun, e S. Yu, «DNN+NeuroSim: An End-to-End Benchmarking Framework for Compute-in-Memory Accelerators with Versatile Device Technologies», in *2019 IEEE International Electron Devices Meeting (IEDM)*, dic. 2019, p. 32.5.1-32.5.4. doi: 10.1109/IEDM19573.2019.8993491.

[10] P.-Y. Chen, X. Peng, e S. Yu, «NeuroSim: A Circuit-Level Macro Model for Benchmarking Neuro-Inspired Architectures in Online Learning», *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 37, fasc. 12, pp. 3067–3080, dic. 2018, doi: 10.1109/TCAD.2018.2789723.

[11] M. Lanza *et al.*, «Recommended Methods to Study Resistive Switching Devices», *Adv. Electron. Mater.*, vol. 5, fasc. 1, p. 1800143, 2019, doi: 10.1002/aelm.201800143.

[12] M. Lanza *et al.*, «Standards for the Characterization of Endurance in Resistive Switching Devices», *ACS Nano*, vol. 15, fasc. 11, pp. 17214–17231, nov. 2021, doi: 10.1021/acsnano.1c06980.

[13] P.-Y. Chen, X. Peng, Y. Luo, e S. Yu, «User Manual of MLP simulator NeuroSimV3.0».

[14] L. Hurtado e G. Piazza, «Characterization of ferroelectric switching in 43 nm Y-36 lithium niobate films», *Appl. Phys. Lett.*, vol. 127, fasc. 8, p. 082904, ago. 2025, doi: 10.1063/5.0285779.

[15] A. M. Kislyuk *et al.*, «Electrophysical properties, memristive and resistive switching of charged domain walls in lithium niobate», *Mod. Electron. Mater.*, vol. 9, fasc. 4, pp. 145–161, 2023.

[16] «FIG. 3. Typical polarization vs. electric field (P-E) hysteresis loop...», ResearchGate. Consultato: 27 novembre 2025. [Online]. Disponibile su: https://www.researchgate.net/figure/Typical-polarization-vs-electric-field-P-E-hysteresis-loop-of-ferroelectrics_fig2_243405640

[17] J. P. V. McConville *et al.*, «Ferroelectric Domain Wall Memristor», *Adv. Funct. Mater.*, vol. 30, fasc. 28, p. 2000109, 2020, doi: 10.1002/adfm.202000109.

[18] F.-C. Chiu, «A Review on Conduction Mechanisms in Dielectric Films», *Adv. Mater. Sci. Eng.*, vol. 2014, fasc. 1, p. 578168, 2014, doi: 10.1155/2014/578168.

[19] S. Huang *et al.*, «Resistive Switching Effects of Crystal-Ion-Slicing Fabricated LiNbO3 Single Crystalline Thin Film on Flexible Polyimide Substrate», *Adv. Electron. Mater.*, vol. 7, fasc. 9, p. 2100301, 2021, doi: 10.1002/aelm.202100301.

[20] X. Pan *et al.*, «Rectifying filamentary resistive switching in ion-exfoliated LiNbO3 thin films», *Appl. Phys. Lett.*, vol. 108, fasc. 3, p. 032904, gen. 2016, doi: 10.1063/1.4940372.

[21] R. H. Olsson *et al.*, «A high electromechanical coupling coefficient SH0 Lamb wave lithium niobate micromechanical resonator and a method for fabrication», *Sens. Actuators Phys.*, vol. 209, pp. 183–190, mar. 2014, doi: 10.1016/j.sna.2014.01.033.

[22] G. Jia e M. J. Madou, «MEMS Fabrication», in *MEMS*, CRC Press, 2005.

[23] J. Baek, L. Hurtado, J. Duncan, e G. Piazza, «18 GHz Y36 Lithium Niobate Ferroelectric Tunable Bulk Acoustic Wave Resonator», in *2025 IEEE International Ultrasonics Symposium (IUS)*, set. 2025, pp. 1–4. doi: 10.1109/IUS62464.2025.11201860.

[24] X. Liu, Y. Wang, J. D. Burton, e E. Y. Tsymbal, «Polarization-controlled Ohmic to Schottky transition at a metal/ferroelectric interface», *Phys. Rev. B*, vol. 88, fasc. 16, p. 165139, ott. 2013, doi: 10.1103/PhysRevB.88.165139.

[25] A. Zaman *et al.*, «Experimental Verification of Current Conduction Mechanism for a Lithium Niobate Based Memristor», *ECS J. Solid State Sci. Technol.*, vol. 9, fasc. 10, p. 103003, ott. 2020, doi: 10.1149/2162-8777/abc3ce.

[26] J. Wang *et al.*, «Analog Ion-Slicing LiNbO3 Memristor Based on Hopping Transport for Neuromorphic Computing», *Adv. Intell. Syst.*, vol. 5, fasc. 10, p. 2300155, 2023, doi: 10.1002/aisy.202300155.

[27] J. Wang *et al.*, «Reliable resistive switching and synaptic plasticity in Ar+-irradiated single-crystalline LiNbO3 memristor», *Appl. Surf. Sci.*, vol. 596, p. 153653, set. 2022, doi: 10.1016/j.apsusc.2022.153653.

[28] Y. Lee e S. Lee, «High-Performance Memristive Synapse Based on Space-Charge-Limited Conduction in LiNbO3», *Nanomaterials*, vol. 14, fasc. 23, p. 1884, gen. 2024, doi: 10.3390/nano14231884.

[29] P.-Y. Chen, X. Peng, e S. Yu, «NeuroSim: A Circuit-Level Macro Model for Benchmarking Neuro-Inspired Architectures in Online Learning», *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 37, fasc. 12, pp. 3067–3080, dic. 2018, doi: 10.1109/TCAD.2018.2789723.

[30] «Fig 1: Schematic representation of multilayer perceptron The figure 1...», ResearchGate. Consultato: 27 novembre 2025. [Online]. Disponibile su: https://www.researchgate.net/figure/Schematic-representation-of-multilayer-perceptron-The-figure-1-shows-the-architecture-of_fig1_259560023

[31] X. Peng, S. Huang, H. Jiang, A. Lu, e S. Yu, «DNN+NeuroSim V2.0: An End-to-End Benchmarking Framework for Compute-in-Memory Accelerators for On-Chip Training», *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 40, fasc. 11, pp. 2306–2319, nov. 2021, doi: 10.1109/TCAD.2020.3043731.

[32] Google, *Gemini*. (2025). [LLM]. Disponibile su: https://gemini.google.com