

From Device Physics to Circuit Co-Design: Multi-Physics Modeling of Monolithic 3D CMOS–NEM Hybrid Circuits

(Invited Paper)

Woo Young Choi*

Department of Electrical and Computer Engineering and the Inter-university Semiconductor Research Center (ISRC),
Seoul National University, Seoul 08826, Republic of Korea

*E-mail: wooyoung@snu.ac.kr

Abstract—Nanoelectromechanical (NEM) switches offer essentially zero off-state leakage, switching abruptness beyond the thermionic limit, nonvolatility based on contact adhesion, and tolerance to radiation and high temperature. Their most practical use is not as a replacement for the transistor but as a complement fabricated in the back-end-of-line (BEOL) above a CMOS substrate: monolithic three-dimensional (M3D) CMOS–NEM hybrid circuits. Realizing such circuits is essentially a modeling problem, because electrostatic, mechanical, surface, fluidic, and process physics are coupled over six orders of magnitude in length scale and the benefit is observable only at the circuit level. In this paper, we review the device physics of NEM switches, organize a four-stage modeling chain consisting of coupled-field finite-element analysis, physics-based analytics, Verilog-A compact models, and hybrid SPICE co-simulation, and connect it to BEOL process co-design. Fabricated M3D demonstrations—sub-1.2-V reconfigurable logic on a standard 65-nm Cu BEOL, NEM nonvolatile memory, and NEM physically unclonable functions—anchor the framework, which projects 4.6× higher density, 2.3× higher frequency, and 9.3× lower power at the tile level. Contact aging, stochastic variability, and EDA integration are discussed as open issues.

Keywords—CMOS–NEM hybrid; nanoelectromechanical switch; monolithic 3D integration; multi-physics modeling; compact model; design-technology co-optimization (DTCO); hardware security

I. INTRODUCTION

Supply-voltage scaling has effectively stalled while the integration density keeps increasing, so that off-state leakage and static power have become the dominant constraints in advanced CMOS. Because thermionic injection limits the subthreshold swing of any field-effect transistor to 60 mV/dec at room temperature, lowering V_{DD} at a fixed overdrive degrades the on/off current ratio. In always-on edge nodes, aerospace and automotive electronics, and dense reconfigurable fabrics, the energy dissipated while doing nothing is now comparable to that of computation.

A nanoelectromechanical (NEM) switch removes this constraint mechanically rather than electronically. Its electrodes are separated by an air gap in the off state, so that the off-state current is determined by the gap rather than by carrier statistics and remains below a typical measurement floor. The switch turns on through an electromechanical pull-in instability, which is more abrupt than the thermionic limit allows, and when the contact adhesion force is engineered to exceed the elastic restoring force, the on state is retained at zero bias so that the switch becomes a nonvolatile mechanical memory [1]. Mechanical switches are also tolerant of radiation and high temperature [2]. Their drawbacks are equally intrinsic: the mechanical motion is slower than carrier transport, the

actuation voltage is set by the stiffness and the gap, and the physical contact wears out [3].

These complementary characteristics call for hybridization rather than replacement, in which CMOS performs computation while NEM switches gate the power, program the routing, store the configuration, and secure the hardware [4]. Monolithic three-dimensional (M3D) integration makes such a hybrid almost free in area, because NEM devices are fabricated within the BEOL metal layers directly above the CMOS substrate and are addressed through conventional vias, as shown in Fig. 1 [5], [6]. Since BEOL processing must remain below $\sim 400^\circ\text{C}$ and the switch couples electrostatic, elastic, surface, fluidic, and process physics, the central obstacle is computational rather than conceptual, since only a calibrated modeling chain from process to device to circuit can quantify whether the hybrid wins. In this manuscript, we review the device physics that any model must capture (Section II), organize a four-stage modeling framework (Section III), discuss process co-design for M3D integration (Section IV), and present circuit-level results and open issues (Sections V and VI).

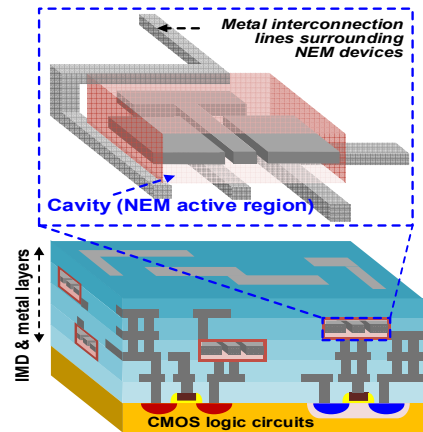


Fig. 1. Conceptual cross-section of a monolithic 3D CMOS–NEM hybrid. NEM switches are fabricated in BEOL tiers ($\leq 400^\circ\text{C}$) above CMOS logic, encapsulated by ALD Al_2O_3 , and addressed through standard vias.

II. DEVICE PHYSICS OF NEM SWITCHES

A. Electrostatic–Mechanical Actuation and Pull-In

Consider a conductive beam with an effective stiffness k suspended over a fixed actuation electrode of overlap area A with an initial air gap g_0 . An applied voltage V generates the attractive force

$$F_{elec}(x, V) = \epsilon_0 A V^2 / [2(g_0 - x)^2] \quad (1)$$

which is balanced by the linear restoring force $F_{mech} = kx$, where x is the beam displacement. Because the electrostatic force

grows superlinearly while the restoring force grows linearly, the equilibrium becomes unstable at one third of the gap and the beam snaps into contact.

$$x_{PI} = g_0/3, \quad V_{PI} = (8kg_0^3 / 27\epsilon_0 A)^{0.5} \quad (2)$$

Equations (1) and (2) assume a lumped parallel-plate capacitor and a single-degree-of-freedom spring. At sub-100-nm gaps, fringing fields, distributed beam compliance, and anchor rotation shift V_{PI} appreciably, so that the analytical form must be calibrated against the coupled-field solutions of Section III-A [7]. Within this limitation, the cubic dependence on g_0 and the linear dependence on k are the two principal levers on the actuation voltage. Scaling the as-fabricated gap, which is set in practice by the sacrificial-layer thickness, into the sub-100-nm regime brings V_{PI} below the logic supply voltage, which is what enabled sub-1.2-V operation on a 65-nm platform [6]. Slingshot actuation additionally exploits the transient momentum of the beam to induce pull-in below the static instability voltage without shrinking the gap [8]. Since the same scaling amplifies the relative contribution of surface forces, the actuation design is directly coupled to the contact physics.

B. Hysteresis, Release, and Nonvolatility

Once the beam is in contact, the gap at the actuation electrode is nearly zero and the electrostatic force is greatly magnified, so that the beam is released at V_{RL} well below V_{PI} and the quasi-static transfer curve becomes hysteretic, as shown in Fig. 2. Analytical release- and switching-voltage models reproduce V_{RL} and the hysteresis window directly from the device geometry [9], [10]. Surface adhesion, which originates from van der Waals attraction and metallic bonding at the contact, adds a force F_{adh} . If

$$F_{adh} \geq kg_0 \quad (\text{at } V = 0) \quad (3)$$

the beam remains latched after the voltage is removed and the switch operates as a nonvolatile memory element, *i.e.*, the NEMory cell [1]. Equation (3) treats adhesion, usually regarded as a stiction failure mode, as a design variable: volatile logic switches require $F_{adh} < kg_0$, whereas nonvolatile memory and configuration switches require the opposite. Both regimes have been demonstrated within the same BEOL technology by tuning the stiffness, contact area, and contact material [2], [5], [6].

C. Contact, Dynamics, and Reliability

The on-state resistance R_{on} is determined by the contact mechanics. Current flows through a few asperity junctions whose combined area is far smaller than the apparent contact area, which gives a Sharvin- or Holm-type constriction resistance. The switching dynamics follow

$$m\ddot{x} + b\dot{x} + kx = F_{elec}(x, V) + F_{surf}(x) \quad (4)$$

where the squeeze-film damping coefficient b governs the switching delay and the contact bounce [11]. The reliability is dominated by the contact, at which wear, micro-welding, oxidation, and contamination change F_{adh} and R_{on} during cycling [3]. The contact material is therefore decisive: soft Al contacts in encapsulated BEOL switches survived only ~ 11 hot-switching cycles [12], whereas hard refractory contacts such as W, Ru, and TiN extend the endurance by orders of magnitude [3].

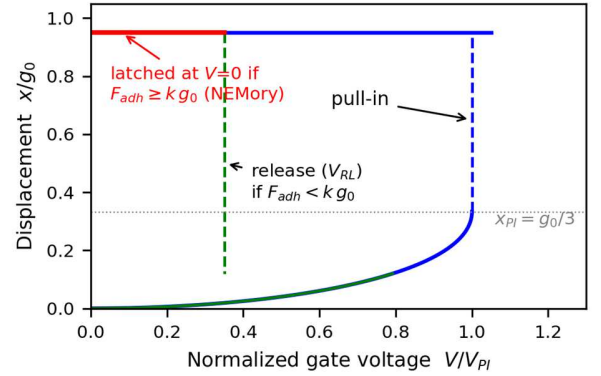


Fig. 2. Quasi-static electromechanical characteristic computed from (1)–(3). Pull-in occurs at $x = g_0/3$, and the release branch is hysteretic. When $F_{adh} \geq kg_0$ the beam stays latched at $V = 0$, providing the nonvolatile NEMory state.

III. MULTI-PHYSICS MODELING FRAMEWORK

No single tool covers the six decades of length scale between an asperity junction (nanometers) and a logic tile (millimeters). We therefore organize the modeling as the four-stage chain shown in Fig. 3, in which physical fidelity is traded for evaluation speed and each stage is calibrated against the one above it.

A. Stage 1 — Coupled-Field FEM and Process-Aware TCAD

The chain begins with finite-element method (FEM) solutions of the coupled electrostatic–elastic problem, including fringing fields [7], anchor compliance, residual stress and stress gradients inherited from deposition, and the topography of the BEOL patterning sequence. Process-aware TCAD is essential because the device is literally made of interconnect: the sacrificial-layer thickness sets g_0 , damascene dishing sets the contact planarity, and the film stress sets the beam curvature, each of which maps directly onto V_{PI} through (2). The sensitivities extracted at this stage, $\partial V_{PI}/\partial g_0$ and $\partial V_{PI}/\partial \sigma$, define the corner space handed to the compact model.

B. Stage 2 — Energy-Based Analytics with Surface Physics

A reduced-order energy functional $U(x, V)$ reproduces the pull-in and release voltages within a few percent of the FEM results at negligible cost, once the effective k , A , and g_0 are calibrated [9]. This analytical layer is augmented by an adhesion energy that determines F_{adh} and hence the volatile/nonvolatile boundary of (3), a multi-asperity contact-resistance model, and squeeze-film damping for the transient of (4). Since the released cavity is hermetically sealed (Section IV), b is set by the pressure trapped inside the seal rather than by the fab ambient, so that the sealing recipe and the switching transient are not independent design choices. Measured hysteresis and retention data are assimilated here.

C. Stage 3 — Verilog-A Compact Modeling

For circuit simulation, the device is described by a Verilog-A module which solves the mass–spring–damper equation self-consistently with the voltage-dependent capacitance $C(x, V)$ and a state-dependent contact resistance, following the modeling practice established for MEMS relay logic [13]. The model must be numerically robust across the pull-in discontinuity, for which continuation and regularization of the contact force are required, and it must expose the process

corners extracted in Stage 1. Since the mechanical transient is several decades slower than CMOS switching, the model also has to tolerate event-driven or multi-rate partitioning in which the mechanical state is advanced only while the beam is in motion.

D. Stage 4 — Hybrid Circuit Co-Simulation

Finally, CMOS transistors from a foundry PDK and NEM compact models are placed in a single netlist, which enables the timing, energy, retention, and variability of a complete hybrid tile to be analyzed. Because a NEM switch toggles rarely while CMOS toggles continuously, the standard digital-flow assumptions break down, and co-simulation quantifies where the mechanical delay is hidden. Only at this stage can technology-level claims such as those in Fig. 4 be made.

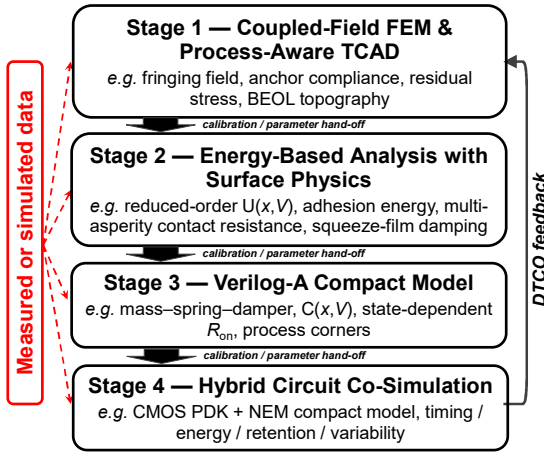


Fig. 3. Four-stage multi-physics modeling chain. Detail decreases and evaluation speed increases downward. Measured data and process simulation calibrate each hand-off, while circuit-level sensitivities feed back to device and process design (DTCO).

IV. M3D INTEGRATION AND PROCESS CO-DESIGN

The M3D CMOS–NEM platform imposes three process constraints: the NEM tier must remain within the BEOL thermal budget (≤ 400 °C), the movable structure must be released without damaging the surrounding dielectrics, and the released cavity must be sealed hermetically at the wafer level.

The 65-nm demonstration in [6] shows that no exotic step is required. Cu NEM switches were formed by the standard dual-damascene interconnect process itself, so that the NEM tier is a byproduct of the existing metallization with zero added mask cost, and the resulting reconfigurable logic operated from the 1.2-V supply of its host node. The encapsulation study in [12] quantified the release-and-seal module: a ~ 20 -nm ALD Al_2O_3 passivation protects the structure, vapor-phase HF at 40 °C for 15 min removes the sacrificial oxide through etch holes, and the sealing is reliable when the etch-hole aspect ratio exceeds ~ 1.14 , which yields an actuation gap of ~ 85 nm. The same scheme has recently been extended to a 22-nm baseline process, in which a NEM memory switch is placed at M7 above a metal–insulator–metal capacitor formed in M2–M4 [14].

Each of these numbers is a process parameter and a device parameter at the same time, since the ALD liner thickness shifts both the gap and the contact interface while the sealing profile perturbs the beam stress. Process–device co-simulation rather than sequential optimization is therefore the natural

methodology, in which the sensitivity chain of Section III-A carries a ± 5 -nm variation of the sacrificial thickness through to a circuit-level V_{PI} distribution.

V. CIRCUIT CO-DESIGN AND APPLICATIONS

A. Reconfigurable Logic

The flagship application is reconfigurable logic, in which NEM switches implement the configuration memory and the programmable interconnect above a CMOS logic fabric [5], [6]. The nonvolatile latch of (3) stores the configuration with zero standby power, and the mechanical switch in the routing path provides a low, voltage-independent on-resistance without the leakage of a pass transistor. Because the configuration bits are written rarely, the mechanical delay is architecturally invisible. Circuit co-simulation of tri-state-buffer tiles projects $4.6\times$ higher logic density, $2.3\times$ higher operating frequency, and $9.3\times$ lower power than the CMOS-only counterpart for tile-to-tile operation, as summarized in Fig. 4 [6], [15]. These gains originate jointly from the removal of SRAM configuration cells, the shortening of the vertical interconnect, and the elimination of routing leakage.

B. Nonvolatile Memory

The NEMory cell [1] exploits the same adhesion latch as a scalable nonvolatile memory. Its design and scaling analysis showed that the beam geometry cannot be optimized simultaneously for the write voltage, the retention margin F_{adh}/kg_0 , and the disturb immunity, since stiffening the beam improves the disturb immunity and the write speed but erodes the retention margin. The cell must therefore be sized against a target duty cycle rather than a single figure of merit. The same latch underpins nonvolatile in-memory search primitives such as ternary content-addressable memories and nearest-neighbor search, which shrink the unit cell and cut the match-line capacitance through BEOL placement [14], while multi-layer switches extend the concept to multi-path routing [16]. Unlike charge-based memory, the mechanical state is immune to radiation-induced upset [2].

C. Hardware Security

The stochastic character of the pull-in instability, which is a variability nuisance elsewhere, is used here as an entropy source, because nominally identical switches latch into different states and thus form a physically unclonable function (PUF). Multilevel NEM states extend the challenge–response space [17], and in situ XOR encryption performed by NEM-PUF cells during data transfer has been demonstrated for lightweight edge security with the PUF fabricated in the CMOS BEOL [18]. Security is therefore a third M3D payload alongside logic and memory, whose figure of merit is the same stochastic device physics that Section III must model.

D. The Co-Design Loop

In all three applications, the decisive quantities—configuration density, retention margin, and PUF entropy—are circuit-level observables of device- and process-level parameters, so that the DTCO loop of Fig. 3 runs in both directions: circuit sensitivity analysis identifies which process tolerances matter, and process simulation prunes the design space before fabrication.

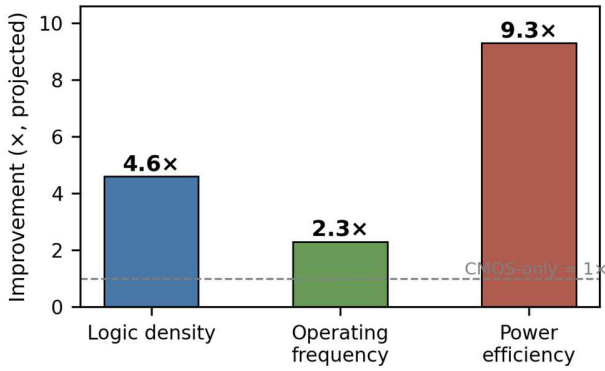


Fig. 4. Projected tile-level benefits of M3D CMOS-NEM reconfigurable logic over the CMOS-only implementation (tri-state buffer tile, tile-to-tile operation) [6]. The values are simulation-based projections anchored to fabricated 65-nm devices.

VI. CHALLENGES AND OUTLOOK

Contact aging remains the first-order open problem. The reported endurance ranges from ~ 11 cycles for soft contacts [12] to $>10^9$ cycles for optimized hard contacts [3], but no predictive physics-based aging model links the material, ambient, and switching energy to the lifetime. Kinetic Monte Carlo and molecular-dynamics studies of asperity evolution, upscaled into the compact-model layer, are a natural contribution from this community, together with a shared reporting convention, since published endurance data are obtained under incompatible loads and switching conditions.

Second, the variability is intrinsically stochastic, since adhesion, roughness, and release dynamics fluctuate device-to-device and cycle-to-cycle. Compact models therefore require stochastic extensions such as random F_{adh} processes and correlated corner sampling, so that the PUF entropy and the logic yield can be co-optimized rather than separately estimated.

Third, EDA integration lags. Hybrid CMOS-NEM design lacks the library characterization, static-timing abstraction, and layout automation that CMOS enjoys. Foundry-compatible NEM processes now exist, including SOI-platform relays [19] and the dual-damascene Cu approach [6], but the missing layer is still the PDK. Machine-learned surrogates trained on Stage-1 FEM data and physics-informed networks that respect the pull-in bifurcation offer a practical route to variability-aware library generation. Finally, NEM memory switches have already been used for in-memory Boolean operations such as XNOR for binary neural networks [20], so that coupling the NEM layers with compute-in-memory fabrics is a concrete near-term direction.

VII. CONCLUSION

M3D CMOS-NEM hybrid circuits convert the physics that makes NEM switches difficult—pull-in instability, adhesion, and contact mechanics—into functionality, i.e., abrupt switching, nonvolatile latching, and stochastic entropy. Demonstrations on standard BEOL platforms establish the feasibility, and the projected tile-level gains of $4.6 \times / 2.3 \times / 9.3 \times$ establish the payoff. Converting this feasibility into a technology is now mainly a modeling and co-design problem, for which the four-stage chain closed by process-aware DTCO defines a concrete agenda.

ACKNOWLEDGMENT

The author thanks the students and collaborators of Three-Dimensional Integration Device Lab. (TIDL) at Seoul National University, whose work is reviewed here. This work was supported by the MSIT under Grant RS-2026-25536380 and RS-2025-02653611 (Intelligent Semiconductor Technology Development Program).

REFERENCES

- [1] W. Y. Choi, T. Osabe, and T.-J. K. Liu, "Nano-electro-mechanical nonvolatile memory (NEMory) cell design and scaling," *IEEE Trans. Electron Devices*, vol. 55, no. 12, pp. 3482–3488, Dec. 2008.
- [2] S. Rana *et al.*, "Nanoelectromechanical relay without pull-in instability for high-temperature non-volatile memory," *Nat. Commun.*, vol. 11, p. 1181, Mar. 2020.
- [3] O. Y. Loh and H. D. Espinosa, "Nanoelectromechanical contact switches," *Nat. Nanotechnol.*, vol. 7, pp. 283–295, 2012.
- [4] H. S. Kwon and W. Y. Choi, "Nanoelectromechanical (NEM) devices for logic and memory applications," *J. Semicond. Technol. Sci.*, vol. 22, no. 3, pp. 188–197, Jun. 2022.
- [5] W. Y. Choi and Y. J. Kim, "Three-dimensional integration of CMOS-NEM hybrid reconfigurable circuits," *IEEE Electron Device Lett.*, vol. 36, no. 9, pp. 887–889, Sep. 2015.
- [6] H. S. Kwon, S. K. Kim, and W. Y. Choi, "Monolithic three-dimensional 65-nm CMOS-nanoelectromechanical reconfigurable logic for sub-1.2-V operation," *IEEE Electron Device Lett.*, vol. 38, no. 9, pp. 1317–1320, Sep. 2017.
- [7] B. Han, J. Y. Song, and W. Y. Choi, "Influence of fringe field on nano-electromechanical (NEM) memory cells," *IEEE Trans. Nanotechnol.*, vol. 13, no. 6, pp. 1102–1106, Nov. 2014.
- [8] W. Y. Choi and H. S. Kwon, "Slingshot pull-in operation for low-voltage nano-electromechanical (NEM) memory switches," *IEEE Trans. Electron Devices*, vol. 66, no. 4, pp. 2040–2043, Apr. 2019.
- [9] J. W. Lee, G. T. Park, and W. Y. Choi, "Analytical release voltage model of monolithic 3-D integrated nanoelectromechanical memory switches," *IEEE Trans. Electron Devices*, vol. 71, no. 5, pp. 3150–3155, May 2024.
- [10] H. M. Lee *et al.*, "Switching voltage analysis of nanoelectromechanical (NEM) memory switches for monolithic three-dimensional (M3D) CMOS-NEM hybrid reconfigurable logic circuits," *IEEE Trans. Electron Devices*, vol. 65, no. 9, pp. 3780–3785, Sep. 2018.
- [11] M. Bao and H. Yang, "Squeeze film air damping in MEMS," *Sens. Actuators A, Phys.*, vol. 136, no. 1, pp. 3–27, 2007.
- [12] J. Jo and W. Y. Choi, "Encapsulation of NEM memory switches for monolithic-three-dimensional (M3D) CMOS-NEM hybrid circuits," *Micromachines*, vol. 9, no. 7, Art. no. 317, Jul. 2018.
- [13] H. Kam *et al.*, "Design, optimization, and scaling of MEM relays for ultra-low-power digital logic," *IEEE Trans. Electron Devices*, vol. 58, no. 1, pp. 236–250, Jan. 2011.
- [14] J. W. Lee *et al.*, "Ternary content-addressable memory using one capacitor and one nanoelectromechanical memory switch for data-intensive applications," *Adv. Intell. Syst.*, vol. 8, no. 3, Art. no. 2500586, Mar. 2026.
- [15] J. W. Ko, G. Baek, and W. Y. Choi, "Scaling trends of monolithic 3D CMOS-nanoelectromechanical reconfigurable logic circuits," *IEEE Trans. Electron Devices*, vol. 67, no. 9, pp. 3861–3867, Sep. 2020.
- [16] J. Yoon, H. S. Kwon, and W. Y. Choi, "Multi-layer nanoelectromechanical (NEM) memory switches for multi-path routing," *IEEE Electron Device Lett.*, vol. 43, no. 1, pp. 162–165, Jan. 2022.
- [17] K.-M. Hwang *et al.*, "Multilevel states of nano-electromechanical switch for a PUF-based security device," *Small*, vol. 15, Art. no. 1803825, 2019.
- [18] D. Kim *et al.*, "In situ XOR encryption for lightweight security using nanoelectromechanical physically unclonable functions," *Adv. Intell. Syst.*, Art. no. 2400805, 2025.
- [19] Y. Li *et al.*, "Volatile and non-volatile nano-electromechanical switches fabricated in a CMOS-compatible silicon-on-insulator foundry process," *Microsyst. Nanoeng.*, vol. 11, Art. no. 140, 2025.
- [20] G. T. Park *et al.*, "XNOR operation of binary neural networks using nanoelectromechanical memory switches," *IEEE Trans. Electron Devices*, vol. 71, no. 12, pp. 7955–7962, Dec. 2024.