ACTIVATION-FLEXIBLE ANN-TO-SNN CONVERSION WITH FINITE-STATE MARKOV NEURONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Most ANN-to-SNN conversion methods rely on a specific correspondence between the source activation and the spiking neuron dynamics. We propose a finite-state continuous-time Markov chain (CTMC) neuron framework whose stationary spike flux can approximate every continuous nonnegative monotone activation function on a compact interval. For a generalized CTMC family with affine input-dependent transitions, we prove uniform approximation to arbitrary accuracy over this function class and derive an explicit approximation error bound. In practice, two- and three-state CTMCs fit ReLU, sigmoid, softplus, and clipped ReLU on the evaluated input ranges, and we evaluate corresponding MLP conversions for each activation with layerwise rate scaling. Moderate clipping improves the conversion cost-accuracy tradeoff on the MNIST MLP and reduces SynOps by 27% on VGG-11/MNIST at matched ANN-SNN accuracy gap criteria, whereas the trend reverses on VGG-11/CIFAR-10. Mean-field and layerwise diagnostics indicate that finite-window sampling and terminal-layer mismatch are the main residual errors. Overall, our results establish finite-state CTMC neurons as a theoretically grounded framework for activation-flexible ANN-to-SNN conversion beyond fixed activation-neuron correspondences.

1 INTRODUCTION

ANN-to-SNN conversion is now a standard route to deep spiking inference, but most current methods are restricted to ReLU-like activation functions and optimized for low-latency rate or time-based coding (Ding et al., 2021; Wang et al., 2022; Stanojevic et al., 2023). This motivates a broader question: can spiking neurons approximate arbitrary monotonic activation functions while retaining simple internal dynamics? We address this question with neurons that store only a discrete state and whose transition rates depend affinely on the input.

Inspired by earlier Markov neuron models and network reductions (Cai et al., 2021; Wu et al., 2023; Chang et al., 2025; Wang et al., 2026), we introduce finite-state continuous-time Markov chain (CTMC) neurons for ANN-to-SNN conversion. Their stationary spike flux approximates a target activation after layerwise rate scaling. The transition structure shapes the input–output curve, while refractory transitions control saturation. Our low-state neurons accurately fit several monotone activations on their operating ranges; a generalized CTMC family uniformly approximates any continuous, nonnegative monotone activation on a compact interval. We then examine how activation shape and spike budget affect conversion efficiency.

On an MNIST MLP, sigmoid, softplus, and ReLU convert within a 1% accuracy gap, and moderate clipping reduces synaptic-event cost. On VGG-11/MNIST, ReLU6 improves conversion efficiency, with residual error mainly associated with finite-spike sampling. The CIFAR-10 results highlight the additional importance of layerwise distribution alignment. Together, these findings identify practical targets for improving conversion accuracy and efficiency.

We make four contributions: 1. We present a practical ANN-to-SNN conversion framework based on finite-state Markov neurons, fitted stationary firing curves, and layerwise rate scaling. 2. We show that boundedness can materially improve rate-coded conversion efficiency, most clearly through a clipped-ReLU ablation and a deep-MNIST VGG comparison. 3. We use mean-field and layerwise diagnostics to separate fitting bias from finite-spike variance and to localize conversion mismatch in deep networks. 4. In the appendix we prove that a generalized finite-state CTMC neuron with affine

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

input-dependent transition rates can uniformly approximate any monotone nonnegative continuous activation on a compact interval.

2 Related Work

ANN-to-SNN conversion. Threshold balancing and rate normalization established deep conversion (Cao et al., 2015; Diehl et al., 2015; Rueckauer et al., 2017). Later methods use residual potentials, quantized source activations, calibration, optimized activation-to-rate maps, or temporal codes to reduce error and latency (Han et al., 2020; Deng & Gu, 2021; Li et al., 2021; Ding et al., 2021; Bu et al., 2022; Wang et al., 2022; Hao et al., 2023; Jiang et al., 2023; Stanojevic et al., 2023); recent post-training work also emphasizes inference-scale conversion across broader vision tasks (Bu et al., 2025). Recent work targets non-ReLU activations and few-spike approximations (Oh & Lee, 2024; Jeong et al., 2024). We instead study activation flexibility under a compact finite-state stochastic representation.

Bounded activations and event-driven hardware. Explicit caps such as ReLU6 are used in efficient deep networks (Sandler et al., 2018). In conversion, however, a cap is useful only if its reduction in high-rate events outweighs any loss of source-model information. We test this tradeoff directly. Finite-state memory is compatible with event-driven implementation, but SynOps alone omits random-number generation, transition scheduling, memory access, and communication costs; therefore we do not claim measured hardware efficiency (Davies et al., 2018).

3 Finite-State Markov Conversion

3.1 Rate Curves as Effective Activations

For a rate-coded unit, the expected output rate acts as its effective activation. Ideal IF dynamics yield a positive linear current–rate map, whereas leak introduces a rheobase and curvature. More generally, the stationary spike-rate curve of a stochastic neuron determines which ANN nonlinearities it can represent. The following result shows that a generalized finite-state CTMC family is sufficiently expressive to approximate a broad class of activation functions to arbitrary accuracy.

Theorem 1 (Uniform approximation)**.** *Let $\phi : [I_{\min}, I_{\max}] \to [0, \infty)$ be continuous and nondecreasing. For every $\varepsilon > 0$, there exists a finite-state CTMC with affine input-dependent transition rates whose stationary spike flux F satisfies*

$$\sup_{I \in [I_{\min}, I_{\max}]} |F(I) - \phi(I)| < \varepsilon.$$

The proof is given in Appendix B. In practice, we use compact two- and three-state parameterizations for the activation functions considered here; we next describe the practical three-state construction.

3.2 Three-state neuron and stationary spike flux

Let $X(t) \in \{B, G, R\}$ and write $a = a_\theta(H) \geq 0$ and $c = c_\theta(H) \geq 0$ for fitted input-dependent transition rates. The generator is

$$Q_\theta(H) = \begin{bmatrix} -a & a & 0 \\ b & -(b+c) & c \\ d & 0 & -d \end{bmatrix}, \qquad b, d > 0. \tag{1}$$

A spike is emitted on $G \to R$. Solving $\boldsymbol{\pi} Q = 0$ and $\boldsymbol{\pi} \mathbf{1} = 1$ gives

$$\nu_\theta(H) = \pi_G(H) c_\theta(H) = \frac{d\, a_\theta(H) c_\theta(H)}{d\,[a_\theta(H) + b + c_\theta(H)] + a_\theta(H) c_\theta(H)}. \tag{2}$$

The practical model fits θ by least squares on a finite operating domain $\mathcal{D}$ sampled from ANN preactivations,

$$\min_\theta \; \frac{1}{M} \sum_{m=1}^{M} \left[\nu_\theta(H_m) - \phi(H_m)\right]^2, \qquad H_m \in \mathcal{D}, \tag{3}$$

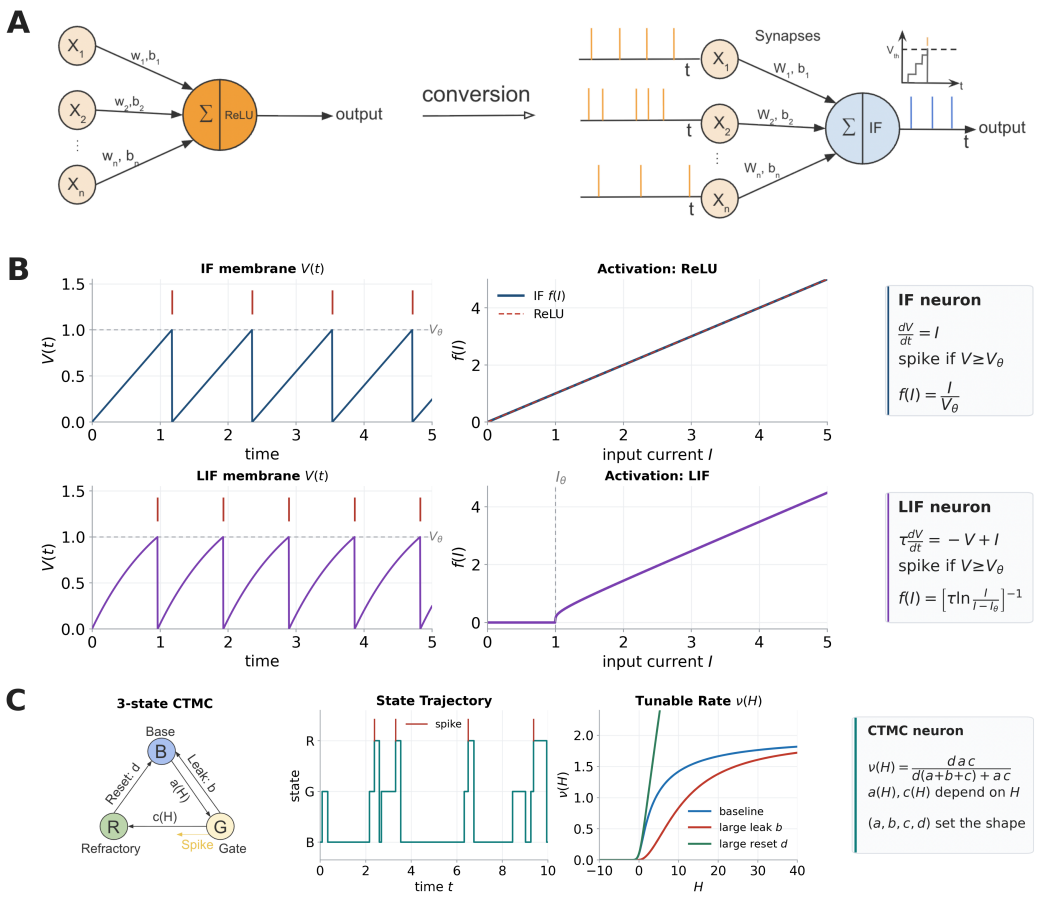

Figure 1: **From conventional neurons to a finite-state CTMC neuron. A:** Rate-coded ANN-to-SNN conversion at the level of one unit. **B:** IF and LIF membrane trajectories and their current-to-rate curves; IF aligns with the positive linear ReLU branch, whereas leak introduces a rheobase and curvature. **C:** The practical three-state neuron has base (B), gate (G), and refractory (R) states. Transitions $B \to G$, $G \to B$, $G \to R$ (spike), and $R \to B$ occur at rates $a(H)$, b, $c(H)$, and d, respectively. The displayed trajectories and parameter sweeps illustrate how leak and reset reshape the stationary firing curve $\nu(H)$.

subject to nonnegative rates. All claims for ReLU concern compact empirical ranges; a finite-state stationary model is not claimed to match an unbounded function globally. Appendix A derives equation (2).

3.3 Layerwise Scaling and Finite-Window Decoding

For ANN layer ℓ,

$$z^\ell = W^\ell a^{\ell-1} + b^\ell, \qquad a^\ell = \phi_\ell(z^\ell), \tag{4}$$

we select a positive scale α_ℓ and target the rate $r^{\ell,*} = \alpha_\ell a^\ell$. To preserve the next affine map in expectation, incoming weights are rescaled as $\widetilde{W}^\ell = W^\ell / \alpha_{\ell-1}$. A generic filtered spike input is

$$H_j^\ell(t) = b_j^\ell + \sum_i \widetilde{W}_{ji}^\ell \big(k_{\tau_H} * dN_i^{\ell-1}\big)(t), \quad k_{\tau_H}(t) = \tau_H^{-1} e^{-t/\tau_H} \mathbf{1}\{t \geq 0\}. \tag{5}$$

Because the kernel integrates to one, a stationary presynaptic rate $r^{\ell-1}$ yields the desired mean input. Over an inference window T, the decoded rate is $\widehat{r}_i^\ell = N_i^\ell(T)/T$ and $\widehat{a}_i^\ell = \widehat{r}_i^\ell / \alpha_\ell$. This construction separates the stationary mapping error from finite-T sampling error.

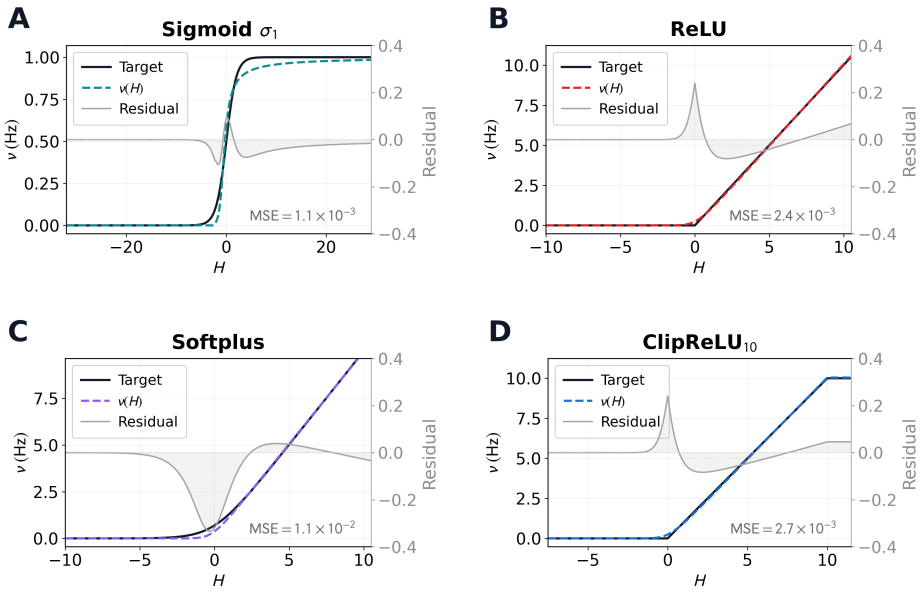

Figure 2: **Stationary-rate fits on the displayed input domains.** Solid curves are target activations, dashed curves are fitted CTMC rates $\nu(H)$, and grey curves are residuals (right axis, shared range). The reported mean-squared errors are 1.1×10^{-3} for sigmoid, 2.4×10^{-3} for ReLU, 1.1×10^{-2} for softplus, and 2.7×10^{-3} for $\mathrm{ClipReLU}_{10}$. The sigmoid uses the three-state family; the piecewise-linear/softplus examples use the low-state variant used in the implementation. These fits demonstrate finite-range flexibility, not exact global representation.

3.4 Evaluation quantities

The accuracy gap is $\Delta_{\mathrm{acc}} = A_{\mathrm{ANN}} - A_{\mathrm{SNN}}$ for the corresponding source ANN, with each accuracy expressed in percentage points. Thus Δ_{acc} is an absolute percentage-point difference, not a normalized relative error. For a sample x, the raw synaptic-event count is

$$\mathrm{SynOps}(x) = \sum_{\ell=1}^{L-1} \sum_{i} N_i^{\ell}(T; x)\ \mathrm{fanout}_i^{\ell+1}\,. \tag{6}$$

We also report emitted spikes per neuron. For deep models, a deterministic mean-field network replaces sampled spike counts by $\nu_\theta(H)$. Its accuracy A_{MF} gives the exact algebraic decomposition

$$A_{\mathrm{ANN}} - A_{\mathrm{SNN}} = (A_{\mathrm{ANN}} - A_{\mathrm{MF}}) + (A_{\mathrm{MF}} - A_{\mathrm{SNN}}), \tag{7}$$

which we call *mean-field mismatch* and *finite-window sampling gap*, and is an operational accuracy decomposition.

4 Experimental Design

We evaluate MNIST (LeCun et al., 1998) with an MLP of width 784–256–128–10 and a strided-convolution VGG-11, and CIFAR-10 (Krizhevsky, 2009) with VGG-11 (Simonyan & Zisserman, 2015). The MLP experiments compare sigmoid, ReLU, softplus, a right-shifted sigmoid, and clipped ReLU. We define $\mathrm{ClipReLU}_K(x) = \min\{\max(x, 0), K\}$. In the clipping ablation, one trained ReLU MLP is reused and each forward pass is clamped to $[0, K]$; consequently each K has its own ANN baseline. Deep experiments compare ReLU with $\mathrm{ClipReLU}_6$ on MNIST and with a threshold derived from the empirical 95th percentile of the ReLU preactivation distribution on CIFAR-10.

Pareto frontiers vary the available simulation and rate parameters. MLP error bars are standard deviations over four seeds. VGG-11/MNIST reports means over four seeds and eight stochastic trials per seed; VGG-11/CIFAR-10 uses four seeds and four trials per seed. The plotted SEM across

Table 1: Highlighted operating points. SynOps are raw spike transmissions per sample. "Criterion" is the ANN–SNN accuracy gap in percentage points.

Setting	Activation	Criterion	SynOps/sample	Spikes/neuron
MNIST MLP	sigmoid	$< 1\%$	$14.4 \pm 0.47\,\mathrm{k}$	0.90
MNIST MLP	softplus	$< 1\%$	$18.1 \pm 2.61\,\mathrm{k}$	1.02
MNIST MLP	ReLU	$< 1\%$	$13.0 \pm 0.58\,\mathrm{k}$	0.87
MNIST MLP	$\mathrm{ClipReLU}_4$	$< 1\%$	$9.1\,\mathrm{k}$	0.55
VGG-11/MNIST	ReLU	$< 0.5\%$	$2.64\,\mathrm{G}$	12.97
VGG-11/MNIST	$\mathrm{ClipReLU}_6$	$< 0.5\%$	$1.93\,\mathrm{G}$	9.51
VGG-11/CIFAR-10	ReLU	$< 2\%$	$2.06\,\mathrm{G}$	13.5
VGG-11/CIFAR-10	$\mathrm{ClipReLU}_{q95}$	$< 2\%$	$2.66\,\mathrm{G}$	15.7

runs describes run-to-run variability, while trials sharing an ANN are nested within a training seed. Table 1 records the operating points highlighted by the supplied figures. Since each accuracy gap is referenced to its corresponding source ANN, clipped and unclipped rows compare conversion fidelity rather than absolute end-to-end accuracy.

5 Results

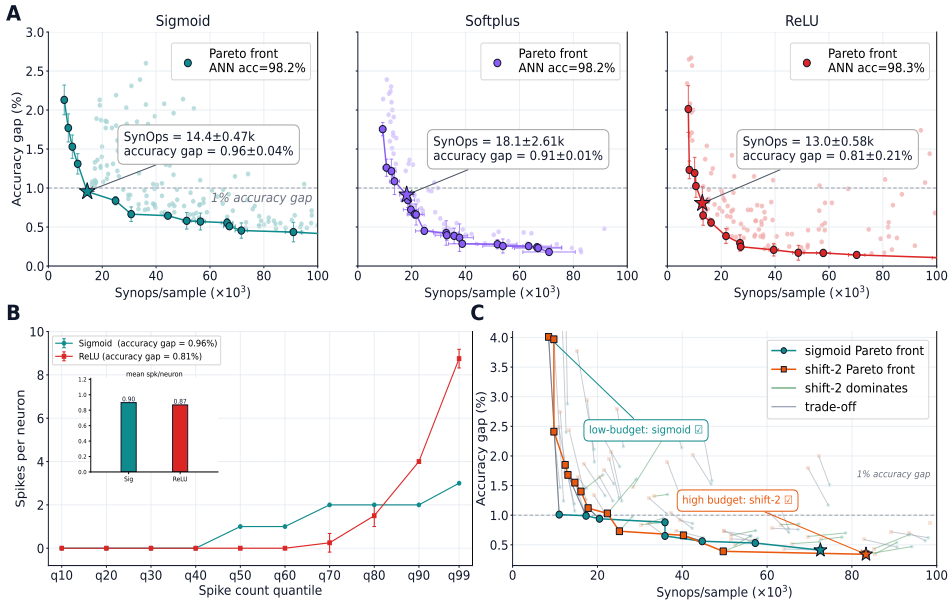

Figure 3: **MNIST MLP activation study** (784–256–128–10; error bars: four-seed standard deviations). **A:** Accuracy gap versus SynOps for sigmoid (source ANN 98.2%), softplus (98.2%) and ReLU (98.3%); stars mark the minimum-cost configurations below a 1% gap: 14.4 ± 0.47k SynOps at $0.96 \pm 0.04\%$ for sigmoid, 18.1 ± 2.61k SynOps at $0.91 \pm 0.01\%$ for softplus and 13.0 ± 0.58k at $0.81 \pm 0.21\%$ for ReLU. **B:** Per-neuron spike-count quantiles show a compressed sigmoid distribution and a sparse, heavy-tailed ReLU distribution; mean counts are 0.90 and 0.87. **C:** Pareto frontiers for sigmoid and a two-unit right shift. Matched-configuration arrows show that suppressing the negative-input tail helps at high budget but can trade away accuracy at low budget.

5.1 Activation shape changes where the spike budget is spent

Figure 3A shows that all sigmoid, softplus and ReLU reach a sub-1% MLP gap, with ReLU doing so at the lowest SynOps. As for sigmoid and ReLU, the aggregate spike counts are similar (0.90 versus 0.87 spikes/neuron), but their distributions differ: sigmoid activity is denser and compressed,

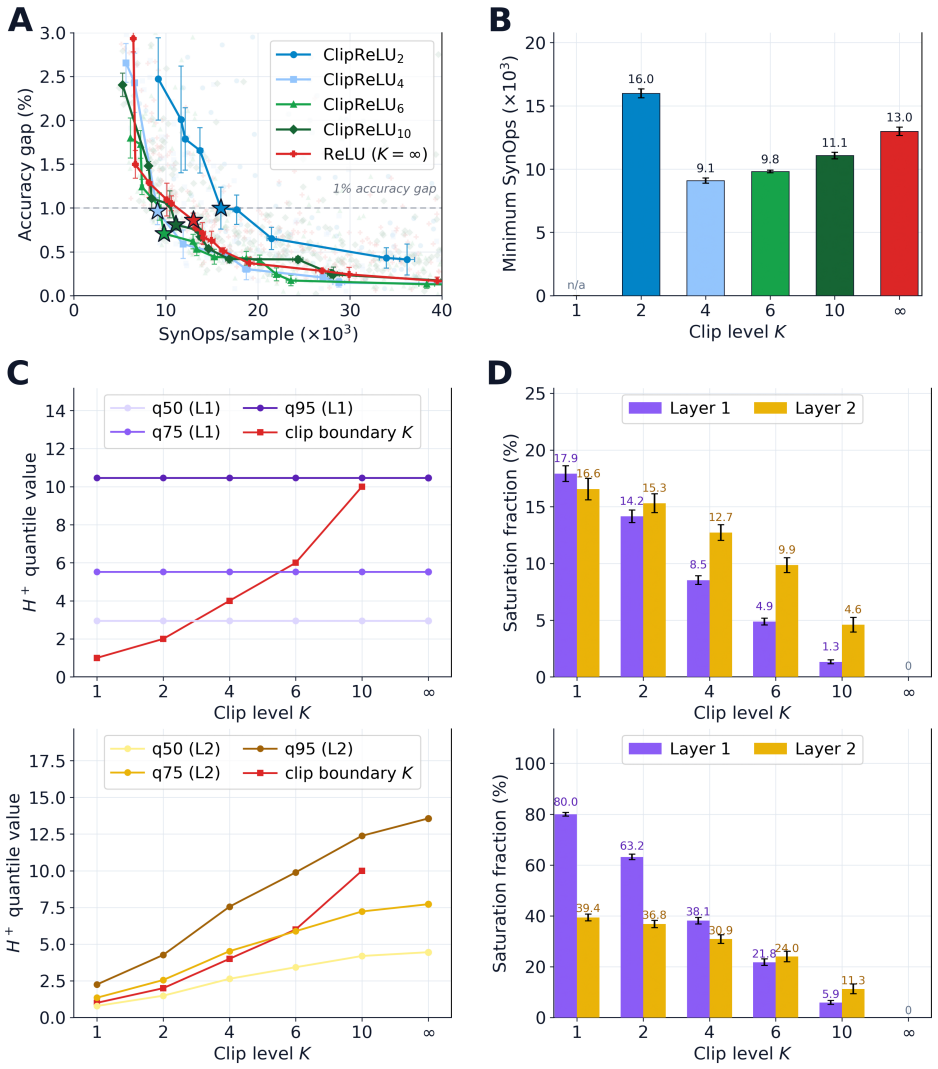

Figure 4: **$\mathbf{ClipReLU_K}$ ablation on the MNIST MLP, using shared ReLU weights and clip-on-forward evaluation. A:** Accuracy-gap/SynOps frontiers. **B:** Minimum SynOps for an accuracy gap below one percentage point; $K = 4$ is best, and $K = 1$ has no configuration satisfying the criterion. **C:** Layerwise positive-preactivation quantiles with the clip boundary. **D:** Fraction exceeding the cap, computed over all units (top) and active units (bottom). Error bars denote standard deviations across four training seeds. Because clipping changes the ANN forward map, each K is evaluated against its own source accuracy.

whereas ReLU is silent through much of the population and develops a heavy upper tail. A plausible source of the sigmoid cost is its nonzero output for negative preactivations. The shift-by-two control in Figure 3C is consistent with this interpretation but is budget dependent: the shift is worse at low budget and better only at sufficiently high budget, so it does not establish a universal benefit.

5.2 Moderate clipping improves the shallow frontier; aggressive clipping fails

The clip-on-forward ablation in Figure 4 isolates a non-monotone tradeoff. The minimum SynOps needed for an accuracy gap below one percentage point are 16.0k, 9.1k, 9.8k, 11.1k, and 13.0k for $K = 2, 4, 6, 10, \infty$, respectively; $K = 1$ never meets the criterion. Thus $K = 4$ reduces the event count by 30% relative to ReLU, while tighter clipping is harmful. Layerwise positive-preactivation quantiles and saturation fractions explain the transition: moderate caps primarily remove the high-rate

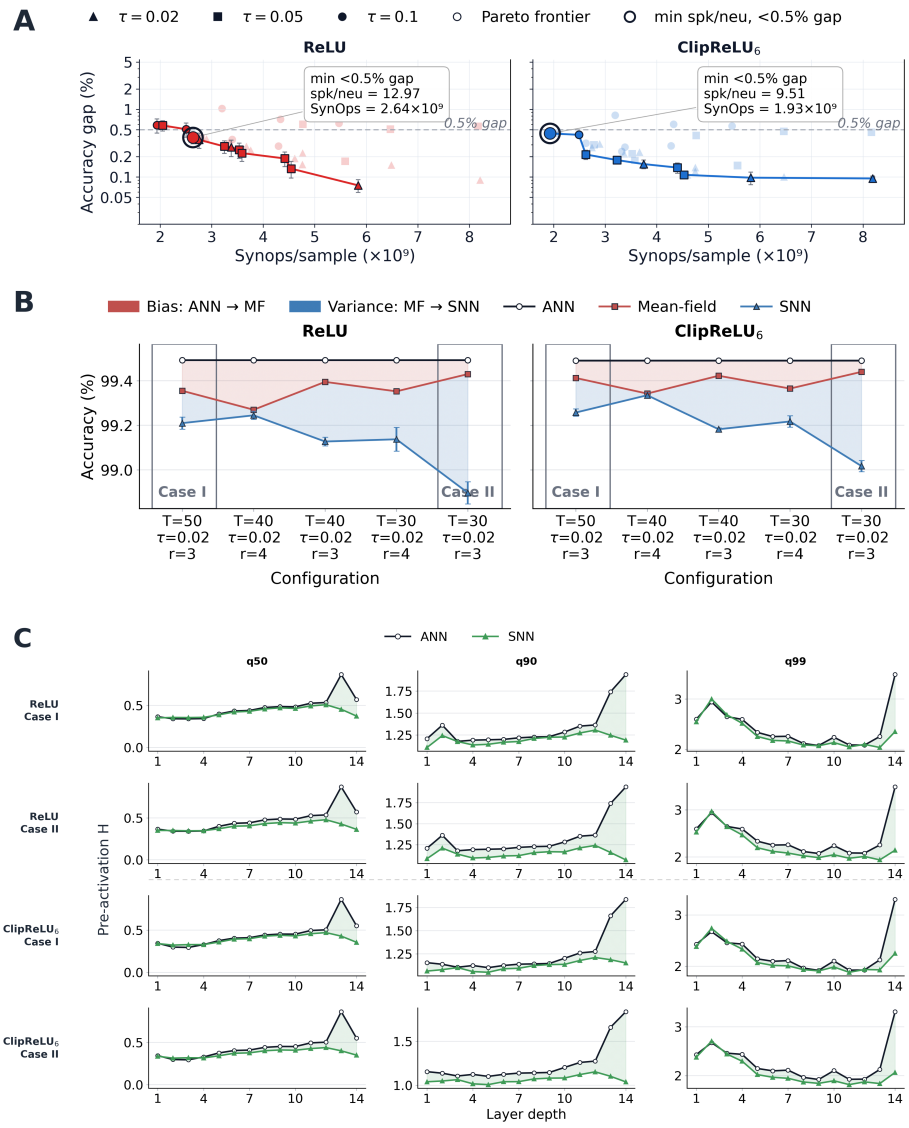

Figure 5: **VGG-11/MNIST: ReLU versus** $\mathrm{ClipReLU}_6$**.** Means over four seeds and eight stochastic trials per seed; error bars are the supplied SEM across runs. **A:** Pareto frontiers; marker shape denotes τ, and open circles mark the minimum-spike point below a 0.5% gap. **B:** ANN, mean-field, and SNN accuracies; Cases I/II use $(T, \tau, r) = (50, 0.02, 3)$ and $(30, 0.02, 3)$. **C:** Preactivation quantiles across 14 layers.

tail, whereas small K intersects the distribution body. At $K = 1$, approximately 80% of active layer-1 units and 39% of active layer-2 units exceed the cap. This experiment supports a conditional mechanism—tail truncation can help—rather than a general advantage of bounded activations.

5.3 On VGG-11/MNIST, clipping lowers cost and sampling dominates the residual gap

In Figure 5A, $\mathrm{ClipReLU}_6$ reaches a sub-0.5% gap with 1.93×10^9 SynOps and 9.51 spikes/neuron, compared with 2.64×10^9 and 12.97 for ReLU—approximately 27% reductions in both quantities.

Figure 5B shows that $A_{\mathrm{ANN}} - A_{\mathrm{MF}}$ remains small across representative configurations, while $A_{\mathrm{MF}} - A_{\mathrm{SNN}}$ is larger and varies with the finite simulation budget. Therefore the dominant residual in this experiment is finite-window stochastic sampling, not stationary-rate fit alone. Layerwise

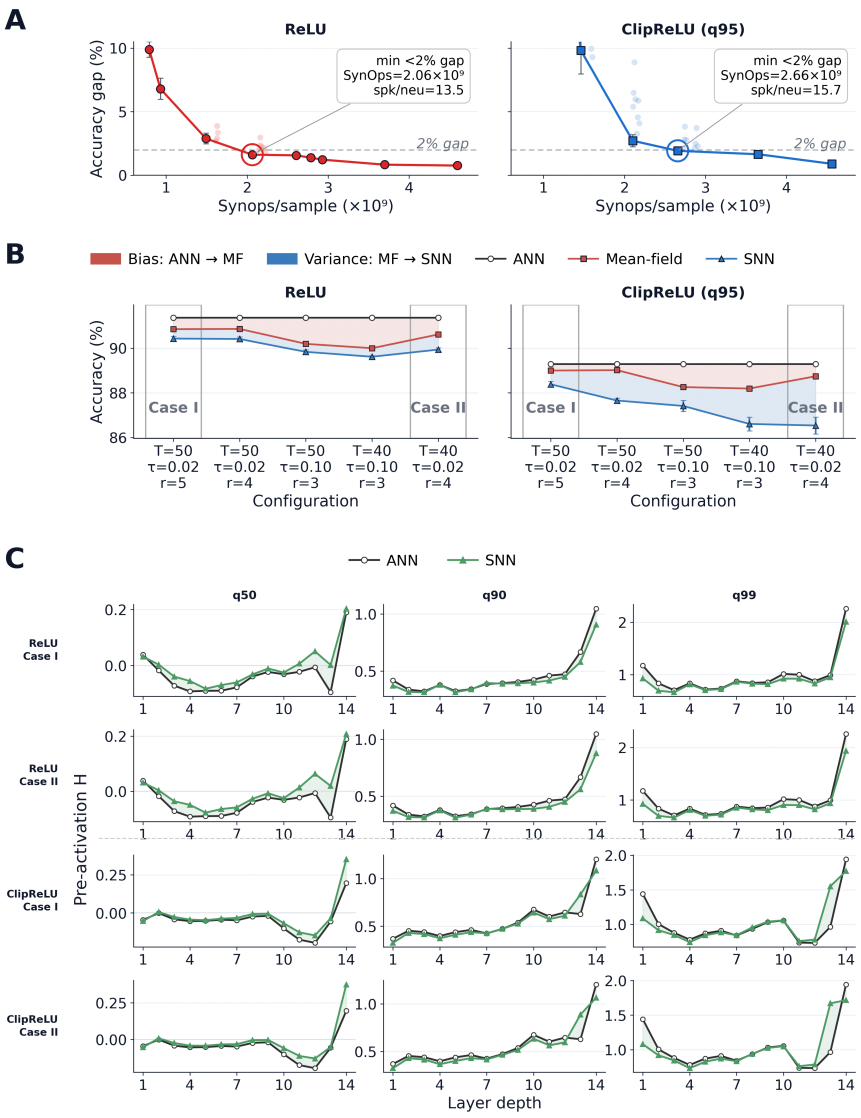

Figure 6: **VGG-11/CIFAR-10: ReLU versus a 95th-percentile clipped variant.** Means over four seeds and four stochastic trials per seed; error bars are the supplied SEM. **A:** Pareto frontiers and points below a 2% gap. **B:** ANN, mean-field, and SNN accuracies. **C:** Preactivation quantiles across 14 layers for Cases I/II, $(T, \tau, r) = (50, 0.02, 5)$ and $(40, 0.02, 4)$; the largest discrepancy is terminal and high-quantile.

$q50/q90/q99$ traces in Figure 5C agree through early and middle layers and diverge mainly in the last two layers and upper quantiles; the bounded variant reduces this terminal mismatch.

5.4 On CIFAR-10, the clipped variant is less efficient

The VGG-11/CIFAR-10 result reverses the MNIST trend (Figure 6). ReLU reaches the sub-2% band at 2.06×10^9 SynOps and 13.5 spikes/neuron, whereas $\mathrm{ClipReLU}_{q95}$ requires 2.66×10^9 and 15.7. The clipped operating point therefore uses 29% more SynOps. The same qualitative trend holds when cost is measured by CTMC state transitions on ReLU and $\mathrm{ClipReLU}_6$ (Appendix D, Fig. 11). For reference, conventional ReLU IF conversion shows a comparable accuracy–cost tradeoff on VGG-11/CIFAR-10; full results are provided in Appendix C.

The operational decomposition again assigns the larger component to finite-window sampling, especially at lower budget. Layerwise quantiles track well until the final layers, where upper-quantile underestimation is strongest for the clipped network. These observations are compatible with clipping removing useful upper-tail information, but they do not by themselves prove that causal explanation; a matched retraining and calibration study is needed.

6 Discussion

Finite-state Markov neurons provide a common conversion framework for ANNs with different monotone activations. Combining stationary firing-rate approximation with layerwise scaling connects activation design to spike cost. The clipped-ReLU and VGG-11/MNIST experiments demonstrate that moderate clipping can reduce synaptic-event counts. The CIFAR-10 results complement this finding by identifying layerwise mismatch as a target for further improvement.

Mean-field and layerwise diagnostics offer a basis for jointly optimizing clipping thresholds, rate scaling, and spike allocation across layers. Future work can investigate how finite-window sampling errors propagate through deep networks and use this analysis to improve calibration. Evaluating event-driven hardware implementations would test whether the observed reductions in synaptic events also reduce inference latency and energy consumption.

7 Conclusion

Finite-state Markov neurons extend ANN-to-SNN conversion beyond ReLU-like activations through simple state dynamics and affine transition rates. The framework combines a general approximation guarantee with accurate low-state fits and demonstrated reductions in synaptic-event cost. These results provide a foundation for jointly designing activation functions, neuron dynamics, and layerwise calibration to support efficient spiking inference.

Reproducibility Statement

Equations (1)–(7) specify the model, scaling, decoding, and reported cost proxy; Appendices A and B give the analytical derivations, including the stationary firing rate and the universal approximation result. Appendices C and D provide additional baseline and diagnostic results. The supplementary material includes an anonymous code archive containing selected implementation code supporting the experiments and analyses reported in the paper.

Ethics Statement

The reported experiments use standard image-classification benchmarks and do not involve human participants or sensitive personal data. The principal risk is over-interpreting SynOps as physical energy; throughout the paper it is described only as a raw event-transmission proxy.

AI Usage Statement

A large language model was used to assist with editorial restructuring, prose refinement, LaTeX organization, reference-format checking, reviewer-style critique, and limited code writing, debugging, and refactoring for the experimental implementation. All AI-assisted code was reviewed and tested by the authors. The authors independently verified all equations, citations, empirical values, interpretations, and experimental results and take full responsibility for the final manuscript and code.

References

Tong Bu, Wei Fang, Jianhao Ding, PengLin Dai, Zhaofei Yu, and Tiejun Huang. Optimal ANN–SNN conversion for high-accuracy and ultra-low-latency spiking neural networks. In *International Conference on Learning Representations*, 2022.

Tong Bu, Maohua Li, and Zhaofei Yu. Inference-scale complexity in ANN–SNN conversion for high-performance and low-power applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24387–24397, 2025.

Yuhang Cai, Tianyi Wu, Louis Tao, and Zhuo-Cheng Xiao. Model reduction captures stochastic gamma oscillations on low-dimensional manifolds. *Frontiers in Computational Neuroscience*, 15: 678688, 2021. doi: 10.3389/fncom.2021.678688.

Yongqiang Cao, Yang Chen, and Deepak Khosla. Spiking deep convolutional neural networks for energy-efficient object recognition. *International Journal of Computer Vision*, 113(1):54–66, 2015. doi: 10.1007/s11263-014-0788-3.

Jie Chang, Zhuoran Li, Zhongyi Wang, Louis Tao, and Zhuo-Cheng Xiao. Minimizing information loss reduces spiking neuronal networks to differential equations. *Journal of Computational Physics*, 537:114117, 2025. doi: 10.1016/j.jcp.2025.114117.

Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, et al. Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro*, 38(1):82–99, 2018. doi: 10.1109/MM.2018.112130359.

Shikuang Deng and Shi Gu. Optimal conversion of conventional artificial neural networks to spiking neural networks. In *International Conference on Learning Representations*, 2021.

Peter U. Diehl, Daniel Neil, Jonathan Binas, Matthew Cook, Shih-Chii Liu, and Michael Pfeiffer. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. In *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2015. doi: 10.1109/IJCNN.2015.7280696.

Jianhao Ding, Zhaofei Yu, Yonghong Tian, and Tiejun Huang. Optimal ANN–SNN conversion for fast and accurate inference in deep spiking neural networks. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2328–2336, 2021. doi: 10.24963/ijcai.2021/321.

Bing Han, Gopalakrishnan Srinivasan, and Kaushik Roy. RMP-SNN: Residual membrane potential neuron for enabling deeper high-accuracy and low-latency spiking neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13558–13567, 2020.

Zecheng Hao, Jianhao Ding, Tong Bu, Tiejun Huang, and Zhaofei Yu. Bridging the gap between ANNs and SNNs by calibrating offset spikes. In *International Conference on Learning Representations*, 2023.

Dayena Jeong, Jaewoo Park, Jeonghee Jo, Jongkil Park, Jaewook Kim, Hyun Jae Jang, Suyoun Lee, and Seongsik Park. A more accurate approximation of activation function with few spikes neurons. *arXiv preprint arXiv:2409.00044*, 2024.

Haiyan Jiang, Srinivas Anumasa, Giulia De Masi, Huan Xiong, and Bin Gu. A unified optimization framework of ANN–SNN conversion: Towards optimal mapping from activation values to firing rates. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14945–14974, 2023.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Yuhang Li, Shikuang Deng, Xin Dong, Ruihao Gong, and Shi Gu. A free lunch from ANN: Towards efficient, accurate spiking neural networks calibration. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6316–6325, 2021.

George G. Lorentz. *Bernstein Polynomials*. Chelsea Publishing Company, 2 edition, 1986.

J. R. Norris. *Markov Chains*. Cambridge University Press, 1998.

Hyunseok Oh and Youngki Lee. Sign gradient descent-based neuronal dynamics: ANN-to-SNN conversion beyond ReLU network. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 38562–38598, 2024.

Bodo Rueckauer, Iulia-Alexandra Lungu, Yuhuang Hu, Michael Pfeiffer, and Shih-Chii Liu. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in Neuroscience*, 11:682, 2017. doi: 10.3389/fnins.2017.00682.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Ana Stanojevic, Stanislaw Wozniak, Guillaume Bellec, Giovanni Cherubini, Angeliki Pantazi, and Wulfram Gerstner. An exact mapping from ReLU networks to spiking neural networks. *Neural Networks*, 168:74–88, 2023. doi: 10.1016/j.neunet.2023.09.011.

Yuchen Wang, Malu Zhang, Hong Qu, and Yi Chen. Towards simple, accurate and high-efficient ANN–SNN conversion. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 2501–2508, 2022. doi: 10.24963/ijcai.2022/347.

Zhongyi Wang, Louis Tao, and Zhuo-Cheng Xiao. Finite-state Markovian surrogates for long-time neuronal state distributions and firing rates, 2026. URL `https://arxiv.org/abs/2505.08254v3`. arXiv:2505.08254, version 3.

Tianyi Wu, Yuhang Cai, Ruilin Zhang, Zhongyi Wang, Louis Tao, and Zhuo-Cheng Xiao. Multiband oscillations emerge from a simple spiking network. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 33(4):043121, 2023. doi: 10.1063/5.0106884.

A Stationary Rate of the Three-State CTMC

Let $\boldsymbol{\pi} = (\pi_B, \pi_G, \pi_R)$ denote the stationary distribution of equation (1). The gate and refractory balance equations are

$$\pi_B a = \pi_G(b + c), \qquad \pi_G c = \pi_R d. \tag{8}$$

Hence $\pi_B = \pi_G(b + c)/a$ and $\pi_R = \pi_G c/d$. Normalization yields

$$\pi_G = \left(\frac{b + c}{a} + 1 + \frac{c}{d} \right)^{-1}. \tag{9}$$

Multiplication by the spike-transition rate c gives

$$\nu(H) = \pi_G c = \frac{acd}{d(a + b + c) + ac}, \tag{10}$$

which is equation (2). This calculation assumes an irreducible chain at the evaluated input; boundary inputs with zero rates are obtained by continuity or by restricting the fitting domain to positive-rate parameters (Norris, 1998).

B Universal Approximation by a General Finite-State CTMC

The low-state neurons used in our experiments are deliberately compact. To prove Theorem 1, we consider a more general finite-state CTMC construction that addresses expressivity in principle while keeping every input-dependent transition rate affine in the input.

Let $\phi : [I_{\min}, I_{\max}] \to [0, \infty)$ be continuous and nondecreasing. Normalize the input as

$$u = \frac{I - I_{\min}}{I_{\max} - I_{\min}} \in [0, 1], \qquad \psi(u) = \phi\big(I_{\min} + (I_{\max} - I_{\min})u\big). \tag{11}$$

Consider active states $A_0, \ldots, A_N$ with birth–death transitions

$$A_i \to A_{i+1} \qquad \text{at rate } \gamma(N - i)u, \tag{12}$$

$$A_i \to A_{i-1} \qquad \text{at rate } \gamma i(1 - u), \tag{13}$$

with absent boundary transitions omitted. Attach a refractory state R_i to each active state:

$$A_i \xrightarrow{\kappa_i,\, \text{spike}} R_i, \qquad R_i \xrightarrow{\rho} A_i, \qquad \kappa_i \geq 0,\ \rho > 0. \tag{14}$$

Since u is affine in I, all input-dependent transition rates are also affine functions of I.

The active birth–death chain has binomial stationary weights

$$b_i(u) = \binom{N}{i} u^i (1 - u)^{N-i}. \tag{15}$$

Writing

$$M_N(u) = \sum_{i=0}^{N} \kappa_i b_i(u), \tag{16}$$

direct balance and normalization give the stationary spike flux

$$F_N(u) = \frac{\rho M_N(u)}{\rho + M_N(u)}. \tag{17}$$

Proof of Theorem 1. Choose $\rho > \max_{u \in [0,1]} \psi(u)$ and define

$$g(z) = \frac{\rho z}{\rho + z}. \tag{18}$$

On the target range, g is invertible, with

$$h(u) = g^{-1}(\psi(u)) = \frac{\rho \psi(u)}{\rho - \psi(u)}. \tag{19}$$

Because ψ is continuous, nonnegative, and nondecreasing, so is h. Set $\kappa_i = h(i/N)$. Then M_N in equation (17) is the Bernstein polynomial of h:

$$M_N(u) = B_N h(u) = \sum_{i=0}^{N} h(i/N) \binom{N}{i} u^i (1 - u)^{N-i}. \tag{20}$$

By the Bernstein approximation theorem,

$$\|B_N h - h\|_\infty \to 0 \tag{21}$$

as $N \to \infty$ (Lorentz, 1986). Moreover,

$$g'(z) = \frac{\rho^2}{(\rho + z)^2} \leq 1, \tag{22}$$

so g is 1-Lipschitz on $[0, \infty)$. Therefore,

$$|F_N(u) - \psi(u)| = |g(B_N h(u)) - g(h(u))| \tag{23}$$

$$\leq |B_N h(u) - h(u)|. \tag{24}$$

The right-hand side converges uniformly to zero. Hence, for every $\varepsilon > 0$, there exists N such that

$$\sup_{u \in [0,1]} |F_N(u) - \psi(u)| < \varepsilon. \tag{25}$$

Applying the affine change of variables between u and I gives

$$\sup_{I \in [I_{\min}, I_{\max}]} \left| F_N\left(\frac{I - I_{\min}}{I_{\max} - I_{\min}}\right) - \phi(I) \right| < \varepsilon, \tag{26}$$

which proves the result. □

Corollary 1 (One quantitative bound)**.** *If h is L_h-Lipschitz, then*

$$\|F_N - \psi\|_\infty \le \frac{L_h}{2\sqrt{N}}. \tag{27}$$

Consequently, $N \ge (L_h/(2\varepsilon))^2$ is sufficient for error at most ε.

Proof. For $K \sim \mathrm{Binomial}(N, u)$, $B_N h(u) = \mathbb{E}[h(K/N)]$. Hence

$$|B_N h(u) - h(u)| \le L_h \mathbb{E}|K/N - u| \tag{28}$$

$$\le L_h \sqrt{\mathrm{Var}(K/N)} = L_h \sqrt{u(1-u)/N} \le \frac{L_h}{2\sqrt{N}}. \tag{29}$$

Apply the 1-Lipschitz property of g used above. □

Remark 1. *The theorem is an expressivity statement, not a practical complexity guarantee. The main experiments use two- or three-state parameterizations, whereas the construction uses $2(N+1)$ active-plus-refractory states. Establishing tight state, spike, and finite-time complexity for useful activation classes remains open.*

Fully connected three-state example. The universal approximation result concerns a generalized finite-state CTMC family, whereas the main experiments use deliberately restricted low-state parameterizations. As an intermediate example, we also fit a fully connected three-state CTMC in which transitions are allowed between all pairs of states. Figure 7 shows that this additional transition flexibility substantially improves the softplus fit.

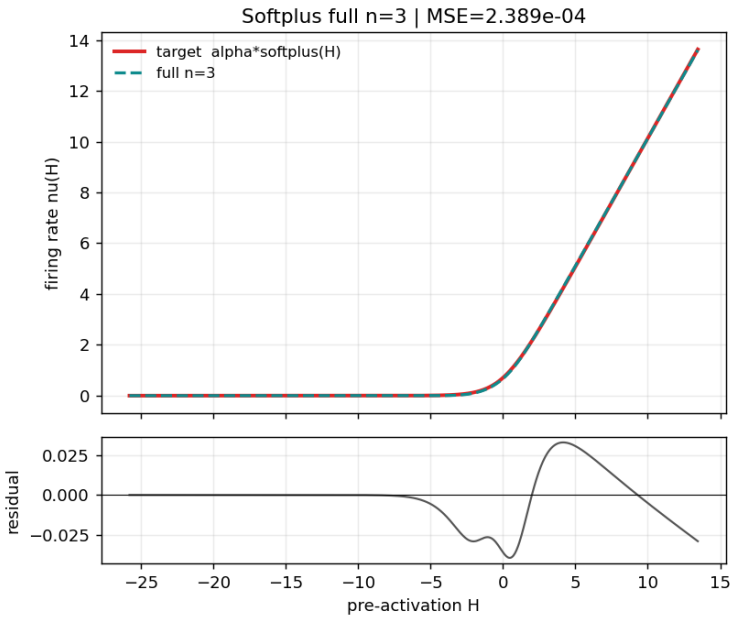

Figure 7: Fully connected three-state CTMC fit to the softplus transfer curve. Unlike the restricted low-state CTMC used in the main experiments, this model permits transitions between all pairs of states, with a spike counted on transitions into the designated "base" state. *Top:* target rate $\alpha\,\mathrm{softplus}(H)$ and the stationary spike-rate prediction over $H \in [-26, 14]$. *Bottom:* residual between the fitted and target rates. The residual remains within ± 0.03 Hz over the evaluated range, with mean-squared error 2.39×10^{-4}.

C BASELINE

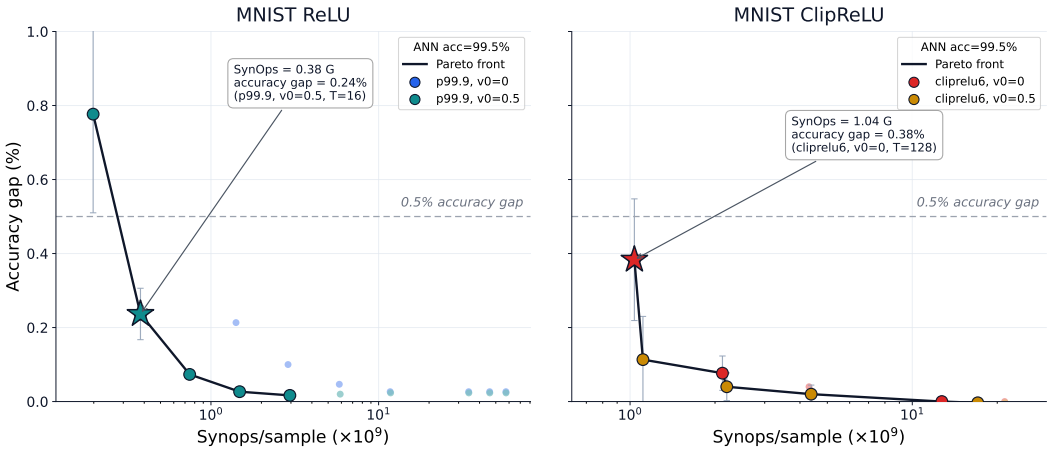

Figure 8: Accuracy–cost trade-off for conventional integrate-and-fire (IF) conversion on VGG-11/MNIST, comparing the ReLU (left) and ClipReLU6 (right) source networks. Each point represents one conversion configuration averaged over three seeds. Curves sweep the simulation length T under the indicated normalization and initial-membrane settings $v_0 \in \{0, 0.5\}$. The Pareto front shows the lowest attainable gap at each cost, with error bars denoting across-seed standard deviation. The dashed line marks the 0.5% gap criterion, and the star marks the lowest-SynOps configuration satisfying it. The selected ReLU configuration reaches a 0.24% gap at 0.38 G SynOps ($p99.9$, v_0=0.5, T=16), while ClipReLU6 reaches a 0.38% gap at 1.04 G SynOps (v_0=0, T=128).

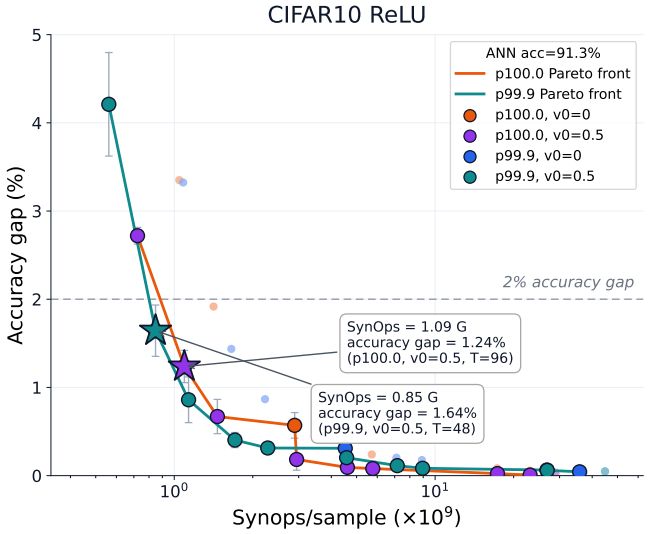

Figure 9: Accuracy–cost trade-off for conventional IF conversion on VGG-11/CIFAR-10 with a ReLU source network, averaged over three seeds. The source ANN accuracy is 91.3%. Results are shown for $p100.0$ and $p99.9$ normalization with $v_0 \in \{0, 0.5\}$ while sweeping the simulation length T. A separate Pareto front is shown for each normalization setting. The dashed line marks the 2% ANN–SNN accuracy-gap criterion, and each star marks the lowest-SynOps configuration satisfying it. Under this criterion, the best $p99.9$ configuration reaches a 1.64% gap at 0.85 G SynOps (v_0=0.5, T=48), while the best $p100.0$ configuration reaches a 1.24% gap at 1.09 G SynOps (v_0=0.5, T=96).

D ADDITIONAL REPORTING DETAILS

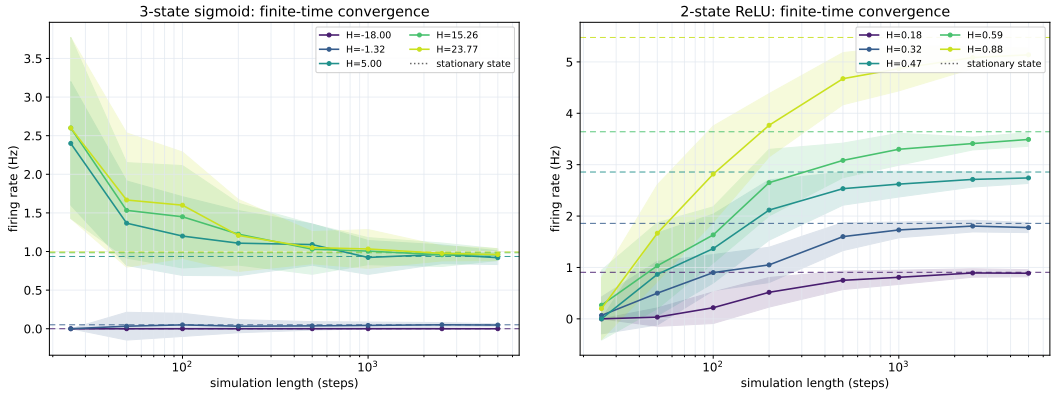

Figure 10: Finite-time convergence of empirical firing rates to stationary rates for representative low-state CTMC neurons. Left: a 3-state sigmoid neuron evaluated at several input values H. Right: a 2-state ReLU neuron evaluated at several input values H. Shaded regions indicate variability across stochastic trials, and dashed horizontal lines mark the corresponding stationary firing rates. The results illustrate how finite simulation length contributes to the gap between stationary-rate predictions and actual SNN behavior in different time steps.

Single-neuron finite-time convergence To illustrate the effect of finite simulation length, Figure 10 shows how empirical firing rates approach their stationary values for representative 2-state ReLU and 3-state sigmoid neurons.

Clipping interpretation. The MLP clipping ablation is "clip-on-forward": the same trained ReLU weights are evaluated after clamping every layer activation to $[0, K]$. This design controls the weights but changes both the ANN predictions and downstream preactivation distributions. Accordingly, the plotted accuracy gap answers whether the SNN tracks each altered forward map; it does not by itself show that the clipped system has higher absolute task accuracy at equal cost.

Deep operating points. The VGG-11/MNIST diagnosis labels Case I as $(T, \tau, r) = (50, 0.02, 3)$ and Case II as $(30, 0.02, 3)$. The CIFAR-10 diagnosis uses $(50, 0.02, 5)$ and $(40, 0.02, 4)$ respectively.

Transition-count cost. Figure 11 reports the corresponding accuracy–cost trade-off on VGG-11/CIFAR-10 when CTMC state transitions, rather than SynOps, are used as the cost measure. The qualitative trend is consistent with the SynOps analysis: the clipped variant does not reduce the minimum cost required to satisfy the 2% accuracy-gap criterion.

Uncertainty. For deep experiments, repeated stochastic trials share each trained ANN. Figures 5 and 6 report the SEM across the displayed stochastic runs and therefore primarily characterize run-to-run stochastic variability; trials sharing the same trained ANN are not independent training replicates.

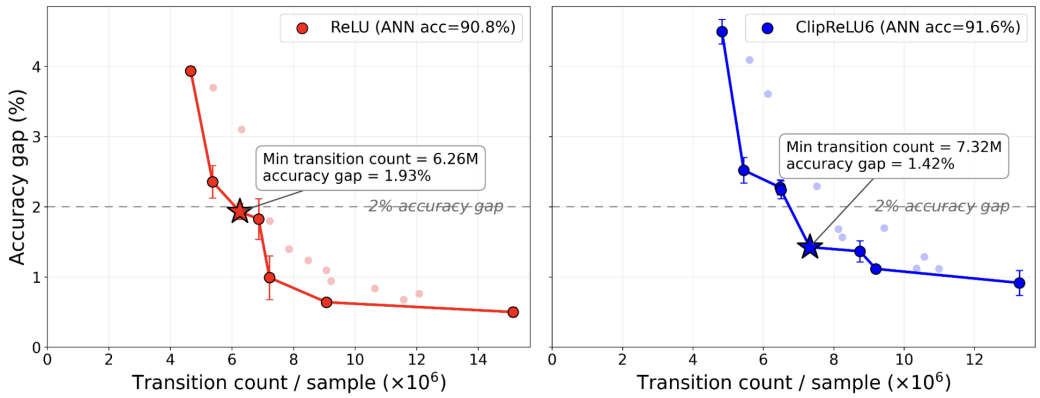

Figure 11: Transition-count trade-off for CTMC conversion on VGG-11/CIFAR-10, comparing ReLU (left) and ClipReLU6 (right). Solid lines show the Pareto frontiers, and the minimum-transition configurations satisfying the 2% accuracy-gap criterion: ReLU reaches a 1.93% gap at 6.26×10^6 transitions/sample, while ClipReLU6 reaches a 1.42% gap at 7.32×10^6 transitions/sample.