Proactive Supply Management of Digital System by Reinforcement Learning

Jongmin Lee[1]*, Hangyeol Mun[2]*, Jaehoon Lee[3,4], Hojin Lee[4], Byung-Jun Kim[3], Heesung Roh[4], Hyunwoo Son[5], Kyu-Jin Choi[6], Jae-Yoon Sim[3]
[1]University of California San Diego, La Jolla, CA, USA, [2]Cornell Tech, New York, NY, USA, [3]POSTECH, Pohang, Korea,
[4]Samsung electronics, Hwaseong, Korea, [5]Gyeongsang National University, Jinju, Korea, [6]ETH Zurich, Zurich, Switzerland
E-mail: jol156@ucsd.edu, hm632@cornell.edu, jysim@postech.ac.kr, *Equal Credit Authors

***Abstract*— This paper presents a Reinforcement Learning (RL)-based digital low-dropout regulator (DLDO) for proactive supply management. A hardware–software co-designed framework is introduced, featuring a Power Cache-based Actor that replaces high-precision neural network computation with precomputed results for low-overhead on-chip inference, along with LDO-specific state and reward designs to improve control accuracy. The framework is trained using post-silicon measurement data to enable robust control under real conditions. The proposed DLDO is implemented in 28 nm CMOS and applied to a general digital signal processor, achieving an 88.47% reduction in mean square error and a 73.37% reduction in maximum voltage droop compared to conventional PI control.**

Keywords—Power management, Reinforcement Learning, Low dropout regulator, Actor-Critic algorithm

I. Introduction

An LDO providing a stable internal supply is essential for SoC power management [1–4]. Voltage droop is a critical issue in LDO design, as it can cause timing violations in load circuits. Conventional LDOs rely on purely reactive control, responding to voltage errors only after they occur through a feedback loop. This inherent control-loop delay fundamentally limits their ability to suppress voltage droop during sudden load current changes. As a result, they must operate with a voltage guard-band at the output, leading to reduced power efficiency [5–6] or constrained performance for a given power budget. Prior work has explored proactive supply voltage regulation [7], but it relies on explicit control signals and has simple load current profiles, limiting its applicability to general processor workloads. There have been ML-based approaches that predict supply current profiles using supervised learning on simulated data [8–12] (Fig. 1, top). However, this modeling-based approach fails to reflect real operating conditions due to several limitations. First, it relies on simulated data at a fixed supply voltage, whereas post-silicon behavior exhibits fluctuations under varying supply voltages and is sensitive to inaccuracies in the power distribution network (PDN) model. Second, since the control flow is composed of a sequence of independent models rather than a single end-to-end model, it cannot achieve globally optimal performance. Finally, as these approaches disable the clock or throttle the processor based on the prediction, performance loss is inevitable [8–10].

To overcome these limitations, we formulate DLDO control as a sequential decision-making problem and propose a proactive DLDO based on RL (Fig. 1, bottom). At each time

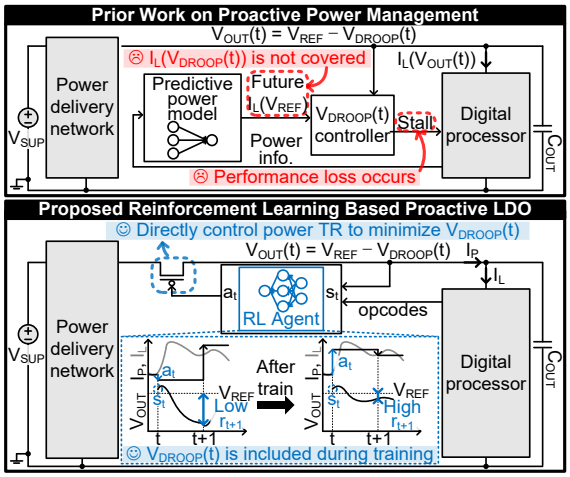


Fig. 1. Prior proactive power management work (top) and proposed RL based proactive LDO (bottom).

step, the agent observes a state that reflects the output voltage condition and instruction-level activity and selects an action that adjusts the on-resistance of the PMOS array. The reward is defined to penalize the deviation of the output voltage from the target, allowing the agent to minimize voltage droop over time. By following those definitions, RL can directly optimize long-term control performance in an end-to-end manner without requiring explicit label of the exact number of ON PMOS for each state. This eliminates the need for costly labeling and thereby enables more efficient learning. Unlike prior approaches based on simulated models, the proposed method is trained using measurement data, allowing the RL agent to capture real operating behavior such as dynamic voltage fluctuations. The proposed hardware–software co-design integrates RL-based control with efficient on-chip implementation and LDO-specific RL training framework, as summarized below:

1) RL-based proactive DLDO trained with post-silicon data, enabling the controller to capture dynamic voltage fluctuations during load circuit's real operation with various workload,

2) A cache-based inference computation that reduces computation overhead by separating the neural network into off-chip and on-chip components, and storing precomputed results in a cache at compile time,

3) LDO-specific state and reward design for improved voltage regulation, which accounts for hardware constraints such as computation delay and blind regions.

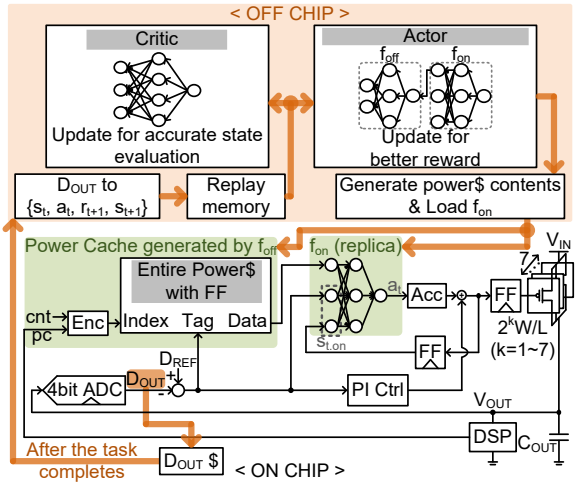


Fig. 2. Overall architecture of the proposed RL-based proactive DLDO framework.

II. Proposed Reinforcement Learning Based LDO

A. System Architecture

Fig. 2 shows the overall architecture and RL training flow. The on-chip components include a 4-bit ADC, Power Cache, a low-overhead computing unit, PI controller, PMOS switches, and a Digital Signal Processor (DSP) as the load. The off-chip side consists of an LDO-specific RL training framework with an Actor and Critic.

The proposed LDO targets a DSP that executes compute-intensive workloads and induces large supply fluctuations. The RL agent observes the state (s_t), which consists of the fetched opcode sequence, the measured voltage droop (V_{DROOP}), and the number of enabled power transistors. Based on s_t, the agent selects an action (a_t), defined as the number of additional power transistors to turn on. The reward (r_{t+1}) is defined as the negative squared error between the output voltage (V_{OUT}) and the target voltage (V_{REF}), and the model is trained to maximize cumulative reward.

During training, the ADC outputs (D_{OUT}) are buffered on chip while the DSP executes workloads composed of RISC instructions with multi-cycle floating-point operations. The recorded D_{OUT} sequence is then transferred off chip and reconstructed into $\{s_t, a_t, r_{t+1}, s_{t+1}\}$ tuples for RL training. The Actor and Critic are updated off chip, and the updated on-chip network weights (f_{on}) and Power Cache contents generated from off-chip network weights (f_{off}) are sent back to the chip for the next episode, which will be described in detail in the following section. This process is repeated until the training converges and the workload space is sufficiently covered. The proposed RL-based control is orthogonal to conventional LDO techniques and safety mechanisms, such as a backstop based on a simple threshold detector after the ADC to identify abnormal voltage droop events.

B. Power Cache based Actor Implementation

Fig. 3 shows the proposed RL-based LDO. We adopt an Actor–Critic algorithm, but deploy only the Actor on chip, as real-time control requires only policy inference. Naive on-chip neural network implementation incurs significant

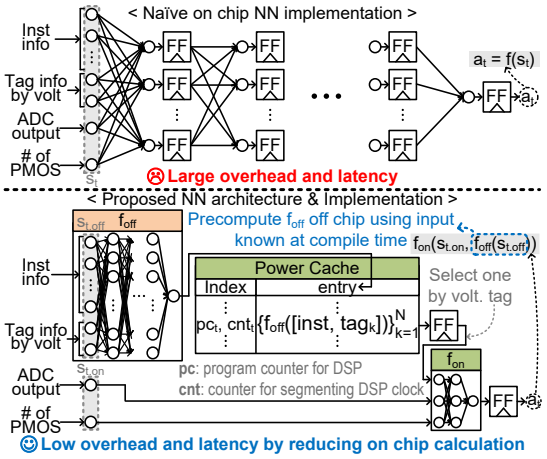


Fig. 3. Actor Neural Network using Power Cache.

computation overhead and latency due to high-precision operations and multi-stage flip-flop delays. To address this, the Actor is divided into a high-precision off-chip network, f_{off}, and a low-precision on-chip network, f_{on}. The f_{off} coarsely estimates voltage-dependent power using instruction-related information, while f_{on} processes real-time data during operation. Since the inputs to f_{off} are determined by the program, its outputs are precomputed at compile time and stored in an on-chip Power Cache. This allows the on-chip Actor to replace expensive computations with simple memory access, completing inference in two cycles: one for cache read and one for f_{on} computation.

To account for instruction's power variation based on supply voltage level, a tag is introduced, where the deviation of the output voltage from the target is nonlinearly quantized into four levels (2 bits) and combined with a 1-bit indicator of whether the voltage is increasing or decreasing. In addition, since the DLDO operates at a higher clock frequency than the DSP, a counter (cnt) is introduced to capture sub-instruction-level timing information. The program counter (PC) and cnt are jointly used as inputs to the Power Cache to account for both instruction context and fine-grained timing variations.

C. LDO-Specific State and Reward Design

Context-Aware State Augmentation: Fig. 4 illustrates the proposed droop-aware state and reward formulation. The power consumption of an instruction depends not only on the opcode but also on its execution context. For example, the current consumption of an add instruction varies depending on whether it is part of a loop-indexing operation or a multiply–accumulate operation. Since such behavior can be inferred from instruction history, we append the three most recently completed opcodes to the state of f_{off} without additional hardware overhead.

Delay-Aware Critic State Redesign: Due to the hardware-limited 2-cycle latency, the Actor determines the number of PMOS at time t using information up to t–2, resulting in inaccuracy. To address this problem, the Critic, implemented off chip, incorporates ADC outputs at both t and t+1 in its state s_t^{cri} along with context-aware state augmentation, enabling more accurate value estimation. Introducing delay-

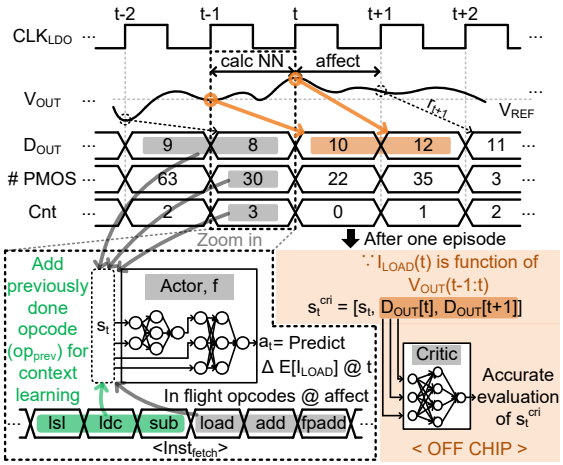


Fig. 4. Timing diagram and the Actor and Critic with context-aware state augmentation and delay-aware Critic state redesign.

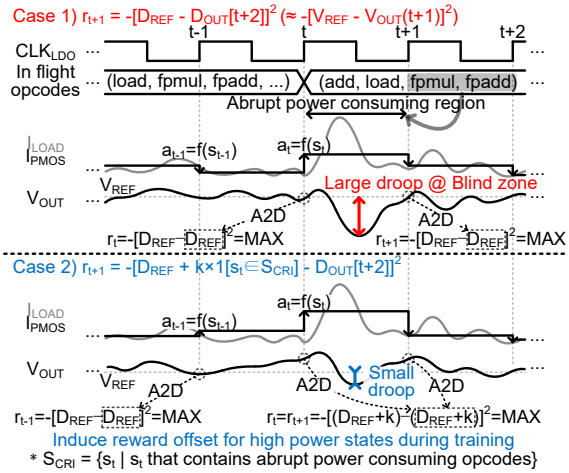


Fig. 5. Effect of the droop-aware reward offset (Case 1: without offset, Case 2: with offset).

aware Critic state redesign can compensate for imperfections of the on-chip Actor due to the delays and guides it so that V_{OUT} converges toward V_{REF} more.

Droop-Aware Reward Offset: Since the ADC samples D_{OUT} only at clock rising edges, defining the reward solely based on the error between D_{OUT} and D_{REF} can lead to unintended voltage droop between sampling points (blind zones), especially under abrupt load changes (Fig. 5, Case 1). To mitigate this, we introduce a droop-aware reward formulation with an intentional offset for high-power events. This encourages the RL agent to proactively increase current and reduce voltage droop (Fig. 5, Case 2).

III. Measurement Results

Fig. 6 (top left) shows the learning curves for four cases depending on whether context-aware state augmentation and delay-aware Critic state redesign are applied. The tasks used for training and testing the RL agent are loop-unrolled versions of FFT, matrix multiplication, image filtering, CORDIC, and FIR filtering. Even though our evaluation

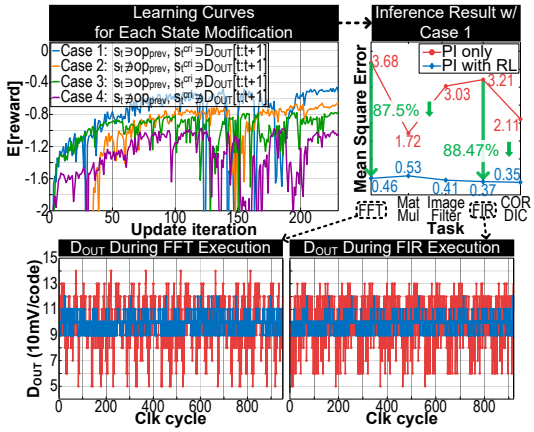


Fig. 6. Measured learning curves for each state modification and inference results.

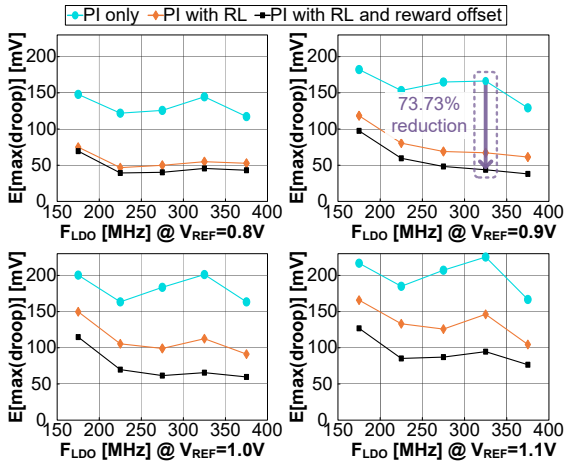


Fig. 7. Summary of Measured maximum voltage droop.

uses a representative subset of DSP workloads, off-policy RL allows additional workloads to be learned without catastrophic forgetting, making the approach scalable to the full DSP workload. Among the four cases, Case 1, which includes both previous opcodes and D_{OUT}[t: t+1] in state achieved the best performance. Using the weights obtained in Case 1, the mean square error of the D_{OUT} sequence during FFT task and FIR task execution were reduced by 87.5% and 88.47% respectively (Fig. 6, top right, bottom), compared to PI-only control, which shows a 358mV voltage droop under a 24mA / 5ns load current surge with 1nF C_{OUT} at V_{IN} of 1.2V, V_{OUT} of 1.1V, and clock frequency (F_{LDO}) of 300MHz in simulation.

Fig. 7 shows the measured voltage droop using the trained weights with different F_{LDO} and V_{REF}. The dropout voltage was fixed to 0.1V. The RL with the droop-aware reward offset on top of PI control shows the best performance, achieving a 73.37% reduction compared to PI control only at V_{REF} of 0.9 V and F_{LDO} of 325 MHz (Fig. 8). Table I compares performance with previous works on proactive control. This work is the first RL-based LDO that controls VDD of general digital processor trained by data obtained from real measurements. Fig. 9 and 10 shows measurement setup and a die photo.

Table 1. Performance comparison.

	IBM J'20 [8]	VLSI'20 [9]	ISSCC'24 [10]	VLSI'23 [11]	ISSCC'26 [12]	**This work**
Process	14nm	7nm	28nm	65nm	28nm	**28nm**
(Volt, Freq)	(N/A, 5.2GHz)	N/A	(0.9V, 500MHz)	(1V, 850MHz)	(1V, N/A)	**(0.99V, 320MHz)**
Power	N/A	N/A	2.2mW	3.5mW	0.9-15mW	**4.175mW ***
Train with I(V_{DROOP})?	N/A	No	No	No	No	**Yes**
Main Proactive System	Critical Path Monitor Linear Regression	Linear Regression PDN Modeling	Linear Regression PDN Modeling	Linear Regression Decision Tree	Linear Regression Neural Network	**RL-Neural Network Power Cache**
Learning Algorithm	N/A	Supervised Learning	Supervised Learning	Supervised Learning	Supervised Learning	**Reinforcement Learning**
Predicted Value	Power	Current, Voltage	Current, Voltage	Current, Voltage	Current, Voltage	**Required # PMOS in LDO**
Actuation Method	Throttling	Clock Gating	Clock Gating	Feedforward Control (DC-DC Converter)	Feedforward Control (DC-DC Converter)	**Change of # PMOS in LDO**
C_{OUT}	N/A	N/A	20nF	2.1nF	N/A	**146pF *****
Target Performance Improvement	25% Max Droop	10% F_{MAX}	32.0% [45.2mV] Max Droop	9.9% F_{MAX}	59% [79mV] Max Droop	**73.7% [122.66mV] Max Droop**
Processor's Performance Loss	> 0% (N/A)	2%	0.6%	0%	3.75-23x throttling reduction	**0%**

* 1.727mW for DLDO controller (mainly f_{on}) calculation, 2.448mW for Power Cache ** 132pF for DSP VDD parasitic cap, 14pF for oscilloscope input cap

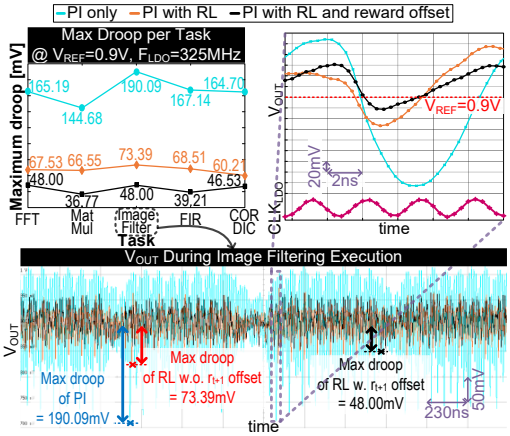


Fig. 8. Measured transient voltage response at V_{REF} = 0.9V and F_{LDO} = 325MHz.

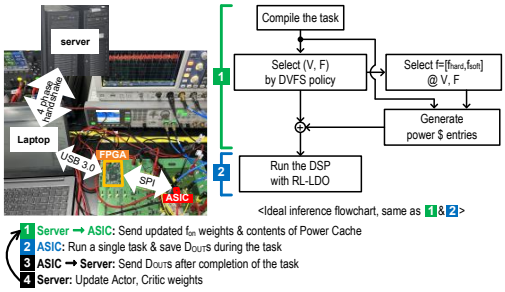


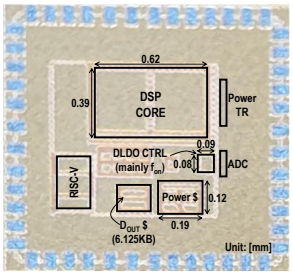
Fig. 9. Test setup.

IV. Conclusion

This work proposes an RL-based proactive DLDO framework with a hardware–software co-design, featuring a Power Cache-based Actor that replaces high-precision neural network computation with precomputed results, and LDO-specific state and reward designs for improved voltage regulation. The framework is trained using post-silicon measurement data to capture real operating behavior. The proposed RL-based proactive DLDO, implemented in 28 nm CMOS, reduces MSE by 88.47% and the maximum voltage droop by 73.37% compared to conventional PI control.

Acknowledgment

This work was supported by NRF under grant No. RS-2026-25467770 and Samsung Electronics.



	Specification
V_{IN} [V]	0.9-1.2 ($V_{dropout}$ = 0.1V)
Freq [MHz]	175-375
Area [mm^2]	0.007 (DLDO CTRL) 0.023 (Power $)
Power $ [kB]	7

Fig. 10. Chip micrograph and summary.

References

[1] S. -Y. Peng et al., "Instruction-Cycle-Based Dynamic Voltage Scaling Power Management for Low-Power Digital Signal Processor With 53% Power Savings," IEEE JSSC, vol. 48, no. 11, pp. 2649-2661, 2013.

[2] M. E. Perez et al., "Distributed Network of LDO Microregulators Providing Submicrosecond DVFS and IR Drop Compensation for a 24-Core Microprocessor in 14-nm SOI CMOS," IEEE JSSC, vol. 55, no. 3, pp. 731-743, 2020.

[3] Y. Han et al., "A DVS-Enabled Distributed Digital LDO Providing Rapid Uniform Power Grid and Ripple Reduction Achieving 20.1-ps FOM in 28 nm CMOS," IEEE TCAS-1, vol. 71, no. 11, pp. 5081-5090, 2024.

[4] J. Arenas et al., "A Dynamically Reconfigurable Digital-Integrated Voltage-Regulator Fabric for Energy-Efficient DVFS in Multi-Domain SoCs," ISSCC, pp. 1-3, 2025.

[5] M. Zelikson et al., "14.3 A Digital Low-Dropout (LDO) Linear Regulator with Adaptive Transfer Function Featuring 125A/mm2 Power Density and Autonomous Bypass Mode," ISSCC, pp. 230-232, 2023.

[6] T. Webel et al., "8.1 Dynamic Guard-Band Features of the IBM zNext System," ISSCC, pp. 1-3, 2025.

[7] J. Kim et al., "8.5 A Command-Aware Hybrid LDO for Advanced HBM Interfaces with 150µA Quiescent Current and 20pF On-Chip Capacitor Achieving Sub-10mV Voltage Droop in 400ps Settling Time," ISSCC, pp. 1-3, 2025

[8] T. Webel et al., "Proactive Power Management in IBM z15," IBM J. Res. Dev., vol. 64, no. 5/6, pp. 15:1-15:12, 2020.

[9] V. K. Kalyanam et al., "A Proactive Voltage-Droop-Mitigation System in a 7nm Hexagon™ Processor," IEEE Symp. VLSI Circuits, 2020.

[10] W. Shan et al., "Proactive Voltage Droop Mitigation Using Dual-Proportional-Derivative Control Based on Current and Voltage Prediction Applied to a Multicore Processor in 28nm CMOS," ISSCC, pp. 256-258, 2024

[11] X. Chen et al., "Proactive Power Regulation with Real-time Prediction and Fast Response Guardband for Fine-grained Dynamic Voltage Droop Mitigation on Digital SoCs," IEEE Symp. VLSI Circuits, 2023

[12] X. Chen et al., "Proactive Power Management-Based Supply Regulation with Online Learning for Variation-Tolerant Workload-Aware Droop Mitigation in 28nm CMOS," ISSCC, pp. 176-178, 2026