Proceedings of the 4th International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/2026.33244

Low Power Consumption Control Algorithm of the AI Chip in the Electronic Control Unit of the Embodied Robot

Zhexi Lu

College of Information Science and Technology & Artificial Intelligence, Nanjing Forestry University, Nanjing, China
luzhexi2004@outlook.com

Abstract. With the large-scale application of embodied robots in complex scenarios, the electronic control unit, as the core control, is faced with problems such as the difficulty in balancing the performance and power consumption of the AI chip and the lack of adaptability between the algorithm and the hardware architecture. These problems seriously restrict the endurance ability and real-time response efficiency of robots. To address the above issues, this paper conducts research on low-power control algorithms. First, it systematically sorts out the research status of electronic control units for robots, low-power technologies of AI chips, and collaborative design of software and hardware. Then, it analyzes the performance characteristics, power consumption performance, and application limitations of three mainstream hardware platforms. On this basis, a dynamic precision-adjustable control strategy is proposed, and a quantitative calculation model for task complexity is established, followed by the construction of a three-level collaborative optimization framework of algorithm-architecture-hardware. This framework can achieve precise adaptation between algorithms and hardware, intelligent task shunting, and refined hardware regulation, effectively realizing the balanced optimization of control performance, real-time response, and energy efficiency. In the future, it will further deepen the in-depth adaptation between algorithms and chips, promote the engineering implementation of related technologies, and ultimately realize the large-scale application of embodied robots in complex scenarios.

Keywords: Embodied robot, electronic control unit, AI chip, low-power control, hardware-software co-design

1. Introduction

The early research on humanoid robots only realized basic motion control. They had low requirements on computational power and did not pay attention to energy consumption. In order to realize human-computer interaction, complex terrain movement, and other functions, Honda ASIMO and other new-generation products have constantly improved their computing power requirements. The serial architecture of the central processing unit (CPU) makes it difficult to meet the parallel computing power requirements of the deep learning model, while the energy efficiency ratio of the graphics processing unit (GPU) is low. Therefore, Field-Programmable Gate Array (FPGA),

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.33244

Application-Specific Integrated Circuit (ASIC), and other special AI chips appeared, and finally promoted the migration of AI algorithms from cloud offline computing to robot real-time computing.

The intelligence level of the embodied robot relies on the computing power of the AI chip of the electronic control unit (ECU). Although the new generation robots can realize the functions of human-computer interaction, complex terrain movement, and so on, they generally suffer from high power consumption, limiting endurance. The low-power optimization of the hardware layer before was costly. After 2015, algorithm-level optimization, such as Dynamic Voltage and Frequency Scaling (DVFS), task diversion, and computing power scheduling, has become mainstream. These methods can adapt to the hardware characteristics of AI chips, and are the main methods to reduce high power consumption and achieve long endurance.

However, the existing low-power schemes often sacrifice accuracy or latency, ignore hardware constraints (tensor cores, DVFS), and lack hardware-software co-design, resulting in low resource utilization. Also, hardware design and algorithm optimization are separated from each other, and there is a lack of an integrated method integrating control algorithm.

Through Y. Wu's research, it can be found that if an AI chip and a low-power algorithm are co-optimized, it can reduce energy consumption by 39% and delay by 40% while maintaining accuracy, improve resource utilization to more than 80%, extend endurance, and ensure the stability of the electronic control unit [1].

This paper focuses on the low-power control algorithm of an AI chip in a robot electronic control unit, aiming to achieve the balance optimization of control performance, real-time response, and energy efficiency, so as to solve the above problems.

2. Technical status and limitations

2.1. Low-power technology of the robot ECU and the AI chip

The core of systems such as embodied intelligence and mobile robots lies in their electronic control unit (ECU), which is responsible for converting AI decision instructions into precise physical actions, such as precise grasping, dynamic gait control, and real-time obstacle avoidance. Its working mechanism is shown in Figure 1. At present, the research on gait control algorithms is mainly based on the methods of learning and optimizing traditional control parameters, combining model predictive control (MPC) with reinforcement learning (RL), and combining RL and whole body control (WBC) [2]. To handle complex control algorithms such as RL, MPC, and visual perception fusion, modern robots' ECUs are increasingly relying on NVIDIA Jetson, FPGA, and other AI chips to provide computing support. Meanwhile, in order to improve the computing efficiency and take into account the requirements of computing power and parallel processing, the current chip generally integrates multiple independent computing cores.

The massive computing requirements of these multi-core processors have led to a significant increase in the power consumption of AI chips, which poses a challenge to battery-powered robots. The limited endurance limits their working hours and scenarios. The power consumption model of the robot electronic control unit is shown in Formula 1.

$$P = P_{static} + \alpha \cdot C \cdot f \cdot V^2 + \eta \cdot A \tag{1}$$

P --total power consumption of the ECU (W), P_{static} --static power consumption of the chip (W). α --switching probability (activity factor) of the circuit within one clock cycle. C --

DOI: 10.54254/2755-2721/2026.33244

computation load of the AI algorithm (GOPS). f -- operating frequency of the chip (GHz). V -- operating voltage (V). η -- control accuracy coefficient. A -- complexity of the control task.

The power reduction methods at the algorithm level (such as model pruning and quantization) are often energy-efficient due to the lack of adaptation to the underlying hardware architecture. The traditional method of combining software and hardware is dynamic voltage frequency regulation (DVFS), which refers to the technology that CPU and GPU dynamically adjust the working frequency and voltage in real time according to the current load condition, to realize the optimization of power consumption and energy efficiency on the premise of ensuring performance. However, the traditional DVFS is a single threshold adjustment, and the frequency increases rapidly when the load is high, which easily leads to a surge in power consumption [3]. Therefore, the current research focus is to build a low-power control framework that cooperates with algorithm and hardware design, which can not only maintain the high control performance of the robot, but also reduce the power consumption of the AI chip.

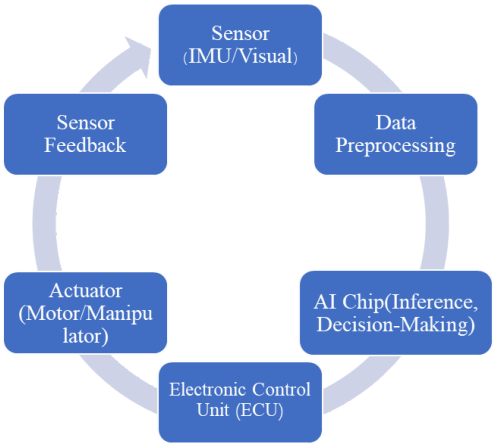


Figure 1. Working mechanism of the Electronic Control Unit (ECU) (picture credit: original)

2.2. Software and hardware co-design in robot field

In the field of robotics, a GPU is a common visual reasoning hardware, such as NVIDIA Jetson Orin, which can accelerate CNN deployment with TensorRT [4]. However, many robot algorithms are designed without considering the tensor core utilization of the GPU or DMA transmission overhead, resulting in the actual energy efficiency far below the theoretical peak. For example, when a mobile robot uses a custom lightweight detection network, the reasoning delay may exceed expectations due to frequent small-scale convolution misaligned tensor core computing mode.

Customized ASIC or SoC is mainly used in mass production platforms or scenes with high requirements for delay and energy efficiency. It realizes the optimal energy efficiency ratio through the on-chip integration of dedicated acceleration units, reducing the delay and additional power consumption caused by data transmission in the traditional multi-chip architecture. According to the research of S. S. Bharadwaj, compared with the multi-chip modular scheme, the optimized SOC achieves 14.7% delay reduction, 16.2% power consumption reduction, and 40.1% energy efficiency improvement [5].

The flexibility of an FPGA makes it suitable for algorithm hardware collaborative research. For example, Zhu C, Huang K, Yang S, et al verified the feasibility of energy efficiency improvement through structured sparse convolution acceleration [6]. However, in actual robot projects, the development cycle of an FPGA usually takes several months, and small changes in the algorithm

DOI: 10.54254/2755-2721/2026.33244

(such as convolution kernel size changes) often require re-synthesis and layout, which leads to iteration difficulties. For example, when the SLAM system accelerates feature extraction on an FPGA, if the front-end algorithm needs to add a new descriptor, the hardware logic must be rewritten, which will inevitably delay the progress of the project.

The above cases reflect a common problem. Hardware features, such as tensor core, DMA constraint, and quantization accuracy, are generally ignored in the algorithm design stage, resulting in serious deviation between actual energy efficiency and theoretical value. In addition, the contradiction between the long development cycle of customized hardware and the fast iteration of algorithms has become increasingly prominent. V. Sze, et al suggested introducing hardware constraints in the early stage of algorithm design, and emphasized the end-to-end power consumption monitoring and verification based on real robots. This idea is particularly critical for improving the engineering reliability of software and hardware cooperation schemes [7].

2.3. Robot ECU and AI chip architecture

The robot's ECU is the core hardware carrier for the interaction between motion control and peripherals, mainly including sensors, actuators, AI chips, etc. At present, AI chips can be divided into three categories: General chips, special chips, and reconfigurable chips. Mainstream general-purpose chips include CPU and GPU, and special AI chips mainly refer to customized processors designed based on ASIC and FPGA methods. These chips are designed specifically for computing tasks [8].

However, in the scene of a robot with high real-time requirements, a multi-chip heterogeneous architecture has significant drawbacks. First, the split architecture leads to high data exchange latency. In the traditional humanoid robot controller, the sensing unit and motion control unit are independent of each other, and the data exchange delay can reach 50-100ms, while the motion control task often requires millisecond or even microsecond response. For example, the gait imbalance of many humanoid robots occurred frequently in the early stage, one of the main reasons is the control delay [9].

Second, the power consumption and heat dissipation requirements of high-performance AI chips conflict with the compact structure of robots. Although the high-performance GPU has strong computing power, its high power consumption makes it difficult to be compatible with the limited battery capacity and cooling space of the robot. For example, the NVIDIA 4090 graphics card consumes about 250W, and the robot battery can be exhausted in dozens of minutes. Its huge volume and heat dissipation noise are more difficult to embed into the compact human body. Xingxing Wang, CEO of Yushu Technology, once said that "the mobility and space constraints of robots determine that they have more stringent requirements for chip computing power and power consumption control". This triangular dilemma of "computing power energy efficiency volume" has become one of the core obstacles to the implementation of specific intelligence [10].

2.4. Design of low-power control algorithm -- dynamic precision adjustable control

The core of dynamic precision adjustable control is to predict the current task complexity (such as environment, operation difficulty, terrain roughness, etc.), as shown in formula 2, and dynamically switch among multiple control models with different scales and precisions trained in advance according to the prediction results, to minimize the power consumption of AI reasoning on the premise of ensuring the success rate of the task. The working mechanism is shown in Figure 2. At the algorithmic level, the multi-precision control model base is constructed by model compression,

DOI: 10.54254/2755-2721/2026.33244

quantification, and knowledge distillation, and the lightweight strategy network is trained to realize the real-time evaluation of task complexity. At the hardware level, it mainly relies on the dynamic power consumption regulation and reasoning acceleration ability of the AI chip to achieve low delay and high energy efficiency of model switching.

$$c = \omega_1 \cdot \Delta v + \omega_2 \cdot \sigma_{env} + \omega_3 \cdot d_{obs} \tag{2}$$

c --Task complexity score (0-1). $\omega_1/\omega_2/\omega_3$ --Weight coefficient (the sum is 1). Δv --Robot speed change rate (normalized to 0-1, the more unstable the speed, the greater the value). σ_{env} -- Variance of environmental dynamics (0.1 for flat road surface and 0.8 for rough road surface). d_{obs} -- Normalized value of obstacle distance (the closer the distance, the greater the value, 0-1).

The purpose is to change the task complexity assessment from qualitative description to quantitative calculation, reflecting the realizability of the algorithm.

Although the dynamic precision adjustable control algorithm shows good potential in low-power design, it still faces some challenges in practical application. For example, the traditional single threshold DVFS method has the problem that the frequency rise and fall are not flexible enough, which may lead to a surge in power consumption during load fluctuations. Therefore, the single threshold DVFS strategy can be upgraded to a multi-threshold control strategy. By establishing multiple independent power consumption thresholds, according to the characteristics of the task (such as computational complexity, environmental adaptability, etc.), the fine adjustment of frequency division and then the coarse adjustment of reference frequency can effectively reduce energy consumption under low accuracy. Through the experiments of D. Chen and others, it is found that the dual threshold power adaptive (DTPA) DVFS scheduling algorithm can maintain the performance above 90%, and the maximum power optimization is above 35% [3].

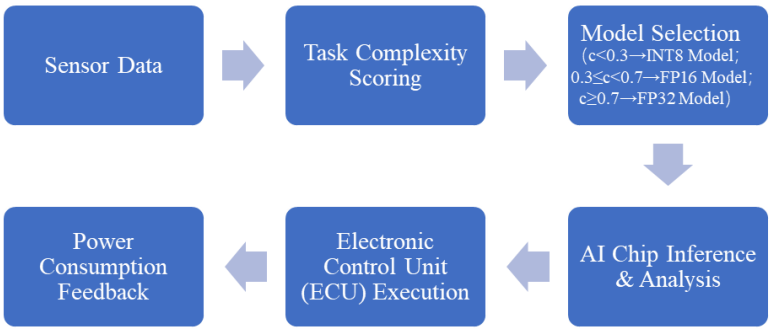


Figure 2. Flowchart of dynamic precision adjustable control strategy (picture credit: original)

3. Collaborative optimization of AI chip and hardware

In section 2, it is pointed out that the robot ECU has some problems, such as high delay, difficulty in FPGA development iteration, contradiction between GPU power consumption and compact body. However, most of the existing schemes are aimed at a single link and lack systematic coordination. To solve this problem, this section proposes a three-level collaborative optimization framework of algorithm layer, architecture layer, and hardware layer, which allows the low-power control algorithm to adapt to the hardware system of the robot electronic control unit, gives full play to the function of hardware resources, and allows the hardware resource scheduling to accurately match the control algorithm [11].

DOI: 10.54254/2755-2721/2026.33244

At the algorithmic level, it can build a multi-precision control model library through model compression, quantization, knowledge distillation, and other technologies, optimize each algorithm module, ensure the accuracy, calculation, and hardware ability match accurately, and use the quantization error compensation formula to realize the operator-level hardware adaptation, and remove the calculation logic incompatible with the hardware. The quantization error compensation formula is shown in Formula 3, in which the compensation coefficient λ and complexity attenuation coefficient K are dynamically adjusted according to the task complexity to ensure that the algorithm accuracy matches the hardware capability.

$$\hat{y} = y_q + \lambda \cdot (y_{fp32} - y_q) \cdot e^{-k \cdot C} \tag{3}$$

$\hat{y}$ --Compensated control output. y_q --Output of quantitative model (INT8/FP16). y_{fp32} -- Output of high precision model (FP32). λ -- Compensation coefficient (0.8-1.0). k -- Complexity attenuation coefficient (0.5-0.8). C : Current task complexity score (same as formula 2).

The function is to quantify the design idea of a low-precision model and error compensation, reflecting the synergy of the algorithm layer and hardware quantization characteristics.

The main goal for the architecture layer is to focus on the intelligent task allocation of AI chips and robot ECUs. CPU is responsible for serial logic control and data preprocessing, GPU/FPGA is responsible for high parallel AI reasoning and visual perception, FPGA/ASIC completes hard real-time tasks such as motor drive, and reduces cross-chip communication delay through PCIe/AXI high-speed bus, to alleviate the problem of data exchange delay mentioned earlier. As shown in Figure 3.

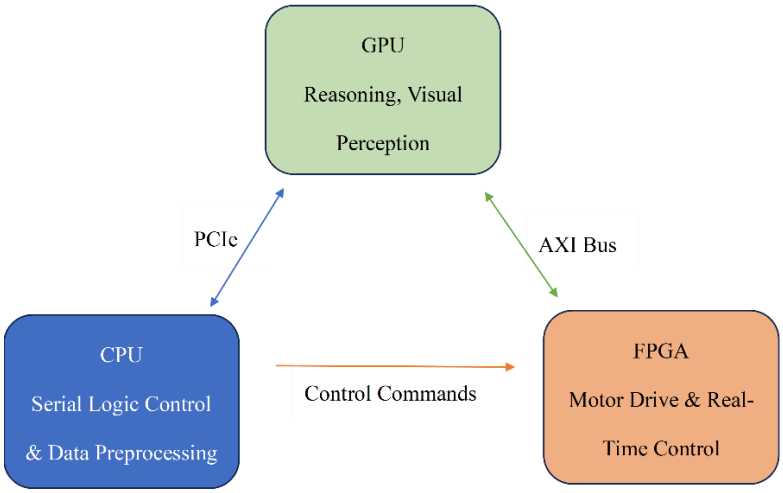


Figure 3. Framework diagram of task diversion at architecture level (picture credit: original)

The hardware layer implements fine regulation for different chip characteristics. GPUs such as Jetson series complete hardware optimization through the TensorRT acceleration framework, and adjust the computing power in real time, combined with multi-threshold DVFS. FPGA can achieve a high degree of adaptation between the algorithm and hardware through its own characteristics of repeatable modification and writing. ASIC completes special hardware scheduling through the integration of perception decision control. The above three-level framework is expected to reduce the end-to-end delay to less than 30ms and significantly improve the energy efficiency ratio.

DOI: 10.54254/2755-2721/2026.33244

4. Conclusion

Focusing on the electronic control unit and low-power control algorithm of the embodied robot, this paper examines the technical status of the robot's electronic control unit, the architecture characteristics of the AI chip, and the method of software and hardware co-design, and analyzes the limitations of the current technology in terms of performance power trade-off, algorithm, and hardware adaptation. Through consulting relevant research, it is found that dynamic precision adjustable control is the key algorithm to reduce power consumption, and the algorithm architecture hardware three-level collaborative framework can realize the unification of algorithm hardware matching, task intelligent shunt, and refined power consumption control.

Future research needs to further deepen the deep adaptation of control algorithms and AI chips, explore adaptive collaboration strategies for robot dynamic tasks, promote the engineering implementation of software and hardware collaboration technologies such as low-precision quantization and network tailoring, provide technical solutions with more practical value for the long-term and reliable operation of battery powered robots, and promote the large-scale application of embodied robots in complex scenes.

References

[1] Wu, Y., Yan, Z., Yin, X., He, L., & Zhuo, C. (2025). ANAS: Software-hardware co-design of approximate neural network accelerators via neural architecture search. Integration, 104, 102469.
[2] Ma, S. (2025). Research on motion control of biped robot based on reinforcement learning. Master's thesis/Doctoral dissertation, Harbin Institute of Technology.
[3] Chen, D., Wu, L., Luo, C., Fan, W., & He, Z. (2023). Research on power test and DVFS scheduling algorithm of high performance multi-core processor chip. Automation Technology and Applications, 42(8), 149–152, 160.
[4] Tian, J. (2023). Hardware-software co-design of FPGA-based convolutional neural network accelerator. Master's thesis/Doctoral dissertation, Harbin University of Science and Technology.
[5] Bharadwaj, S. S., & Ramkumar, P. (2025). Chiplet-based RISC-V SoC with modular AI acceleration. 2025 3rd International Conference on Computational Intelligence and Network Systems (CINS), Dubai, United Arab Emirates.
[6] Zhu, C., Huang, K., Yang, S. (2020). An efficient hardware accelerator for structured sparse convolutional neural networks on FPGA. IEEE Transactions on Very Large Scale Integration Systems, 28(9), 1953–1965.
[7] Sze, V., Chen, Y. H., Yang, T. J., & Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE, 105(12), 2295–2329.
[8] Gao, J., Fan, Z., Bao, M. (2024). A review of AI chip design technology for robots. Integrated Circuits and Embedded Systems, 24(11), 60–77.
[9] APUQI. (2025, April 21). Humanoid robots enter the "dual-brain collaboration" era: Breaking the technical bottleneck of core controllers. https: //www.apuqi.com/news/
[10] Wang, X. (2025, September 26). The computing power dilemma: Why humanoid robots cannot accommodate NVIDIA's high-performance chips. CSDN Blog. https: //blog.csdn.net/weixin_44421595/article/details/142552515
[11] Dai, Y. (2024). Hardware-software co-design of transformer deep learning accelerator. Master's thesis/Doctoral dissertation, Tiangong University.