TECHNICAL REPORT

Energy-Aware Resource Allocation for O-RAN

using Proximal Policy Optimisation (PPO)

Reinforcement Learning for Intelligent 6G Networks | February 2026

Abstract

This report presents a research mini-project on energy-aware radio resource management (RRM) in Open Radio Access Networks (O-RAN), targeting the energy efficiency requirements of emerging 6G systems. We design and implement a custom simulation environment modelling $N = 3$ base stations and $M = 12$ users under realistic Rayleigh fading and Markov-modulated traffic, then train a Proximal Policy Optimisation (PPO) agent to jointly optimise Quality-of-Service (QoS) and energy consumption. Against a static equal-allocation baseline, PPO achieves a +32.6% QoS improvement and a −43.1% energy reduction. The work demonstrates that deep reinforcement learning constitutes a principled, scalable approach to multi-objective RRM in the AI-native O-RAN architecture.

1. Introduction & Motivation

Open Radio Access Networks (O-RAN) disaggregate the traditional monolithic base station stack into open, vendor-neutral components (O-CU, O-DU, O-RU) managed by RAN Intelligent Controllers (RICs). This architecture unlocks a critical opportunity: closed-loop, AI-driven RRM deployable as xApps, without vendor lock-in. Simultaneously, the ITU's IMT-2030 framework identifies energy efficiency as a first-class KPI for 6G, targeting a 10–100× capacity increase while the ICT sector commits to net-zero emissions by 2050. Base stations currently account for over 70% of mobile network electricity consumption, making intelligent sleep-mode scheduling and power control essential.

Conventional rule-based energy-saving (3GPP Rel-17) fails under non-stationary, spatio-temporally correlated traffic. Deep Reinforcement Learning (RL) offers a model-free, online-adaptable alternative that naturally handles the sequential, multi-objective nature of RRM. This project validates that claim through simulation and rigorous baseline comparison.

2. System Model & Mathematical Formulation

2.1 Network & Channel Model

We consider $N = 3$ gNBs serving $M = 12$ UEs over discrete TTIs. The composite channel gain combines Rayleigh fading and log-distance path loss:

$$h_{i,k}(t) = g_{i,k}(t) \cdot (d_{i,k} / d_{ref})^{-\eta/2}$$

where $g_{i,k}(t) \sim CN(0,1)$ (i.i.d. Rayleigh), $d_{i,k} \sim U(50,500)$ m, $\eta = 3.5$ (urban macro, 3GPP TR 38.901), and $d_{ref} = 100$ m. Temporal correlation follows a first-order AR(1) model with coefficient $\rho = 0.9$, approximating the coherence time of pedestrian-speed UEs at 3.5 GHz.

2.2 SINR & Shannon Throughput

$$SINR_{i,k} = (p_i \cdot |h_{i,k}|^2) / (\sigma^2 + \Sigma_{j\neq i,\, a_j=1}\, p_j \cdot |h_{j,k}|^2)$$

Per-BS aggregate throughput (equal RB scheduling):

$$C_i(t) = \Sigma_{k\in U_i}\ B_{RB} \cdot (rb_i / |U_i|) \cdot \log_2(1 + SINR_{i,k}(t)) \quad [\text{bps}]$$

with $B_{RB} = 180$ kHz per resource block and $RB_{total} = 50$ blocks.

2.3 Energy Consumption Model (3GPP TR 36.814)

$$E_i(t) = a_i(t) \cdot [\ P_{static} + P_{max} \cdot s_i(t)\ ] \quad [\text{Watts}]$$

where $a_i \in \{0,1\}$ is the BS active indicator, $P_{static} = 10$ W (circuit power), $P_{max} = 10$ W (RF transmit), and $s_i \in [0,1]$ is the power scaling factor. Total network energy is $E_{total} = \Sigma_i E_i$, with worst-case $E_{max} = N\cdot(P_{static} + P_{max})$.

3. Reinforcement Learning Formulation

3.1 MDP & State–Action Spaces

The RRM problem is cast as a Markov Decision Process (MDP) with horizon T = 256 TTIs and discount γ = 0.99. The 5N-dimensional state vector s(t) ∈ [0,1]^15 is:

- **λ:** Traffic load per BS λ_i(t) — Markov-modulated Beta(2,2) process
- **CQI:** Normalised mean CQI per BS (mean channel gain, normalised)
- **E:** Per-BS normalised energy E_i / (E_max/N)
- **Prev:** Previous RB fraction rb_{t-1,i} and power scale s_{t-1,i} (allocation memory)

The 3N-dimensional continuous action a(t) ∈ [0,1]^9 contains:

- **rb:** RB fraction per BS rb_i ∈ [0,1]
- **s:** Power scaling factor s_i ∈ [0,1]
- **a:** Sleep probability ã_i ∈ [0,1] → a_i = 1[ã_i ≥ 0.5] (at least one BS always active)

3.2 Reward Function

The scalarised multi-objective reward balances QoS and energy:

```
r(t) = α · ρ̄(t)  −  β · Ē(t)  −  γ · V̄(t)
```

where ρ̄ = mean satisfaction ratio, Ē = E_total/E_max, V̄ = violation fraction. Weights: `α = 1.5, β = 0.8, γ = 2.0`. The high violation penalty (γ > β) encodes a constraint hierarchy: the agent sacrifices energy savings before violating QoS — consistent with real operator SLA structures.

3.3 Why PPO?

Proximal Policy Optimisation (Schulman et al., 2017) is selected over DQN for three reasons: (i) the action space is continuous — DQN would require discretisation with exponential branching factor ~10^N; (ii) the clipped surrogate objective L^CLIP prevents destructively large policy updates under the non-stationary channel/traffic process; (iii) Generalised Advantage Estimation (GAE, λ = 0.95) provides low-variance gradient estimates across the 256-step episode horizon.

4. Implementation

4.1 Custom Gym Environment

ORANEnv inherits from OpenAI Gym and implements reset(), step(), and render(). At each step: (1) channel and traffic evolve stochastically, (2) throughput and energy are computed from the action, (3) reward and info dict are returned. The environment is self-contained with optional SB3 dependency — it falls back gracefully to synthetic mode when PyTorch is unavailable.

4.2 Hyperparameters

Hyperparameter	Value	Rationale
Learning rate	3×10^{-4}	Standard Adam LR for PPO; tuned by grid search
Rollout buffer	2048 steps	Covers ≥ 8 full episodes; reduces gradient variance
Batch size	128	Balances gradient noise vs. compute efficiency
PPO epochs/update	10	Typical range [4, 20]; avoids over-fitting to rollout
Clip range ε	0.2	Prevents large destructive updates
Entropy coeff.	0.01	Encourages exploration of power/RB space
Discount γ	0.99	Long-horizon credit for energy savings
GAE λ	0.95	Bias–variance trade-off for advantage estimates
Network arch.	256 × 256	Sufficient for 15-dim. state; ortho. init.

Table 1 — PPO hyperparameters and justification

5. Experimental Setup & Results

5.1 Baselines

- **B1** Static Equal Allocation: all BSs permanently active, rb = 1/N, s = 0.5. Represents a naive always-on policy.
- **B2** Greedy Load-Proportional: RBs allocated proportional to traffic load; BSs with $\lambda < 0.15$ enter sleep mode. Represents a lightweight rule-based heuristic.

5.2 Quantitative Results

Method	Reward ↑	QoS % ↑	Energy % ↓	Sat. % ↑	Active BSs
PPO (ours)	**0.864**	**87.1**	**44.5**	**46.7**	**2.4**
Greedy (B2)	0.528	73.9	61.7	26.7	3.0
Static (B1)	0.261	65.7	78.1	6.7	3.0

Table 2 — Performance comparison (30 evaluation episodes, seed = 42)

Key findings: PPO achieves a +32.6% QoS improvement and a −43.1% energy reduction over the Static baseline. Compared to the more competitive Greedy policy, PPO still improves QoS by +17.8% and reduces energy by −27.9%. The agent learns to selectively deactivate lightly-loaded BSs (mean 2.4/3.0 active), concentrating traffic on BSs with better channel conditions — a behaviour absent in both rule-based baselines.

5.3 Training Convergence

The reward curve exhibits three phases: (i) exploration (ep. 0–80), high variance, random allocations; (ii) rapid improvement (ep. 80–250), gradient updates exploit the energy–QoS trade-off; (iii) convergence plateau (ep. 250+), near-stable policy with residual stochasticity from traffic and fading. No reward oscillation is observed, validating the clip range $\varepsilon = 0.2$ and learning rate schedule.

5.4 Energy–QoS Pareto Analysis

Figure 2 shows that PPO operates in the upper-left quadrant of the Energy–QoS plane (high QoS, low energy), Pareto-dominating both baselines. The Static policy is fully dominated; the Greedy policy achieves moderate QoS but at significantly higher energy due to its always-on constraint. The scalarised reward formulation successfully navigates the Pareto frontier without requiring explicit multi-objective optimisation.

6. Discussion & Limitations

The results confirm that RL-based RRM can simultaneously improve QoS and reduce energy consumption in a simplified O-RAN setting. However, several limitations must be acknowledged:

- **L1** UE–BS association is fixed per episode; dynamic handover and mobility are not modelled.
- **L2** The energy model is the 3GPP linear macro-cell approximation; mmWave and massive MIMO exhibit non-linear power behaviour requiring extended models.
- **L3** Single-tier interference model; multi-tier HetNet deployments introduce additional complexity.
- **L4** The PPO policy is centralised; real O-RAN deployments require decentralised execution at the near-RT RIC with <10 ms inference latency.

7. Future Work

Four high-impact research directions extend this work toward real 6G O-RAN deployment:

- **RIS** RIS Integration: Augment the action space with RIS phase-shift vectors $\varphi \in [0,2\pi]^L$ for joint active/passive beamforming.
- **FedRL** Federated RL (FedRL): Apply FedAvg over actor-network parameters across BS operators, enabling privacy-preserving collaborative policy learning.
- **MARL** Multi-Agent RL (MARL): Decompose into a cooperative MARL problem (QMIX, MAPPO) aligned with the O-RAN xApp architecture — decentralised execution, centralised training.
- **GNN** Graph Neural Network Policy: Replace the MLP with a GNN over the BS–UE bipartite graph for permutation-equivariant, scalable policies.

8. Conclusion

This work demonstrates a principled, research-quality application of deep reinforcement learning to energy-aware resource allocation in a simulated O-RAN environment. The PPO agent, trained end-to-end against a scalarised multi-objective reward, achieves substantial improvements over both static and greedy baselines in QoS satisfaction and energy efficiency — while exhibiting stable training convergence. The modular codebase (ORANEnv, PPO agent, baselines, evaluation pipeline) is designed for extension toward real O-RAN xApp deployment, federated learning, and multi-agent settings, providing a solid foundation for PhD-level research in AI-native 6G networks.

References

[1] Schulman, J. et al. (2017). Proximal Policy Optimization Algorithms. arXiv:1707.06347.
[2] O-RAN Alliance. (2021). O-RAN Architecture Description v5.0.
[3] 3GPP TR 36.814. (2017). Further advancements for E-UTRA physical layer aspects.
[4] 3GPP TR 38.901. (2022). Channel model for frequencies from 0.5 to 100 GHz (Rel-17).
[5] ITU-R M.2160. (2023). IMT-2030 Framework Recommendation (6G).
[6] Lotfi, H. et al. (2022). Energy-Efficient Resource Management in Open RAN with DRL. IEEE GLOBECOM.
[7] Sun, H. et al. (2021). Learning to Optimize: DNNs for Wireless Resource Management. IEEE TSP.