靜 宜 大 學

資訊工程學系

畢 業 專 題 成 果 報 告 書

分數多代理深度強化學習結合能量虛擬佇列於行動邊緣運算之 AoI 與能量最佳化

學生：

資工四 A 411147291 李語桐

資工四 A 411147534 張皓閔

資工四 A 411147699 陳鈺憲

指導教授：

劉建興 教授

西 元 二 O 二五 年 十二 月

分數多代理深度強化學習結合能量虛擬佇列於行動邊緣運算之 AoI 與能量最佳化

學生：　李語桐　　指導教授：劉建興
張皓閔
陳鈺憲

靜宜大學資訊工程學系

摘　　要

本研究聚焦於解決行動邊緣運算 (MEC) 環境中，資源受限條件下的資訊年齡 (AoI) 最小化問題。傳統方法難以在多代理分散式決策中，同時兼顧即時應用所需的資訊時效性與資源利用的長期穩定性（如能耗與傳輸成本）。為此，本研究提出一種混合優化框架，它透過創新性融合兩種領先的學術方法。我們以“Asynchronous Fractional Multi-Agent Deep Reinforcement Learning for Age-Minimal Mobile Edge Computing”為核心基礎，並將受“Minimizing Version Age of Information in Resource-Constrained Multi-Source Systems via Lyapunov and Learning Based Scheduling” 啟發的虛擬佇列約束機制整合至其中。我們將長期資源約束（如能耗）轉化為可即時追蹤的隊列動態（ΔZ 與 Z），並將其整合至 MARL 的瞬時獎勵函數中，實現節能調度。此混合優化公式巧妙地平衡了 AoI 最小化與資源約束滿足，既保留了 DPP 的穩定性直覺，又實現了在多代理環境中的可擴展、事件驅動訓練。我們的貢獻在於將異構學習模型（D3QN 處理卸載空間決策，PPO 處理更新時機連續決策）與虛擬佇列機制整合，建構出一個統一且高效的系統。透過模擬驗證，本方法能共同最小化資訊年齡，同時在實際運作條件下有效保證資源效率，為複雜 MEC 系統的時效性與能效優化提供了新的解決方案。

誌　　謝

本專題能夠順利完成，對我們來說是一段充滿挑戰、學習，也充滿成就感的旅程。在這一年多的時間裡，我們從不熟悉強化學習與 MEC 系統，到能自己閱讀論文、建立模型、撰寫程式，再到完整跑出實驗結果，我們要非常感謝指導教授**劉建興老師**。老師不只在技術面上給我們方向，也在研究方法、問題拆解、甚至是如何閱讀學術論文上提供很多關鍵性的指導。每次討論時，老師總是能快速抓住我們卡住的點，並用清楚的方式幫我們整理思路。從模型架構、環境設定，到 AoI、Lyapunov 虛擬佇列等比較不直覺的概念，老師都耐心帶著我們一步步理解。如果沒有老師的協助，我們不可能把這個跨多領域的專題整合起來。

本專題是我們在大學生涯中最完整的一次研究經驗，也是讓我們真正理解什麼是跨領域整合、什麼是科研精神的一次挑戰。能夠完成這份成果，是所有人共同努力的結果，我們在此再次表達最深的感謝。

目　　錄

摘 要 i
誌 謝 ii
目 錄 iii
圖目錄 iv
符號說明 iv
一、緒論 1
1.1 研究背景與動機 1
1.2 研究挑戰 2
1.3 研究目的與貢獻 3
二、專題內容與進行方法 4
2.1 系統機制與功能架構 4
2.1.1 系統機制與策略 4
2.1.2 主要功能模組 4
2.2 系統模型 5
2.2.1 資訊年齡(Age of Information,AoI) 5
2.2.2 Lyapunov 函數與漂移定義 Lyapunov Drift-Plus-Penalty(DPP) 5
2.3 能量收穫與消耗 6
2.4 實驗設計 7
2.4.1 虛擬佇列 7
2.4.2 獎賞函數(Reward Function) 7
三、專題流程與架構 8
3.1 系統架構圖 8
3.1.1 行動邊緣運算(MEC)系統 8
3.1.2 多行動裝置於 MEC 環境中決策之流程 9
四、專題成果介紹 10
4.1 數據比較 10
4.2 系統畫面 11
4.2.1 平均能量 11
4.2.2 平均 AoI 12
4.2.3 能量虛擬佇列 12
五、專題學習歷程介紹 13
5.1 專題相關軟體學習介紹 13
5.2 專題製作過程遭遇的問題與解決方法 15
六、結論與未來展望 16
6.1 結論 16
6.2 未來發展方向 16
七、參考文獻 18

圖目錄

圖 1 MEC 系統架構……8
圖 2 整體系統的運作概念……9
圖 3 AoI 和總能量使用在不同權重下後一百回合訓練的平均值……10
圖 4 各 Episode 的平均能量(包括本地運算、傳輸、收穫與淨變化)……11
圖 5 平均 AoI 隨 Episode 的變化 ……12
圖 6 能量虛擬佇列隨 Episode 的變化……12

符號說明

符號	定義
$\mathcal{M}$	行動裝置的集合
$Y_{m,k}$	完成k任務的總時間
$Z_{m,k}$	生成任務k之前更新間隔
$\mathbb{E}$	根據政策所做決策的期望
Z	虛擬佇列（Virtual Queue）
ΔZ	虛擬佇列的瞬時漂移量（Drift）
E_{harv}	能量收穫
E_{loc}	本地運算能耗
E_{tx}	由本地傳輸至邊緣伺服器能耗
E_{use}	能量總消耗
k	晶片常數
f	CPU 頻率(Hz)
C	計算所需週期
e_{bit}	bit 傳輸能量(J/bit)
B	任務大小（bit）
U	均勻分布（Uniform distribution）
λ	權重值
R	獎勵函數(reward)
a_t	代理於時間 (t) 的行動（Action）
s_t	代理於時間 (t) 的狀態（State）

一、緒論

1.1 研究背景與動機

隨著智慧城市、工業物聯網（IoT）、穿戴式設備以及無人機（UAV）任務的快速發展，大量即時資料的蒐集與運算已成為行動系統的基本需求。**行動邊緣運算**（Mobile Edge Computing, MEC）因能提供低延遲、高效能且貼近使用者的運算能力，逐漸成為支撐上述應用的重要基礎架構。

在諸多應用中，「資訊是否足夠即時」比單純的通訊延遲更為關鍵。例如無人機在長航程任務中（如災害搜救、態勢感知、邊境巡防），必須在有限的能源環境下完成飛行、視覺處理、資料上傳等多項任務。若資料更新不夠即時，將直接降低環境理解能力甚至影響任務安全。同樣情形也出現在智慧城市感測節點與穿戴式醫療裝置，其能量來源有限（如太陽能、小型電池或能量收穫機制），但又必須持續且及時地上傳資料。

傳統 MEC 研究多聚焦於延遲（Latency）或吞吐量，然而這些指標無法完整描述資料的新鮮程度。因此，**「資訊年齡」（Age of Information, AoI）**逐漸成為衡量即時性的核心指標，用以描述資料自產生至被接收到之間經過的時間。AoI 能更直接反映系統是否維持足夠即時的資訊，以滿足 UAV 感測、安全監測、行動控制與醫療應用中對高時效性的需求。

然而，要在能源有限、任務動態變化與通訊不穩定的條件下維持低 AoI，是一項極具挑戰性的問題。特別是在能量收穫波動（如光照不穩）、計算負載變動明顯的情況下，如何決定何時更新資料、是否將任務卸載至邊緣伺服器，以及如何保障長期能量穩定，是 MEC 系統面臨的核心難題。因此，本研究以「**分數階深度強化學習**（Fractional DRL）」與「**虛擬佇列**（Virtual Queue）」為基礎，建構一套能於非同步多代理環境下運作之 AoI 最佳化決策框架，以提升能源受限裝置的即時資訊傳輸與決策品質。

1.2 研究挑戰

在以 AoI 最小化為目標的 MEC 系統中，本研究需面對以下四項主要挑戰：

（1）能量隨機性與 AoI 最佳化的基本衝突

能量收穫具有高度波動（如太陽光照變化或使用者行為差異），使裝置無法穩定地以高頻率更新資料。而 AoI 的最佳化通常需要增加更新頻率，兩者在目標上形成天然矛盾，需透過額外機制取得平衡。

（2）分數型 AoI 目標難以以傳統強化學習處理

AoI 涉及更新間隔與任務延遲的比值，屬於分數型目標，無法直接套用傳統 RL 之累積獎勵架構。需要引入能處理比值目標的「分數階強化學習（Fractional RL）」。

（3）多代理互動帶來策略相依性與協作問題

多個 IoT 或 UAV 裝置會競爭「邊緣伺服器的計算資源」，因此每個代理的更新與卸載策略都會影響其他代理的 AoI 與能量消耗，使得整體問題呈現高度耦合，需設計可收斂至穩定解的**多代理強化學習**（MARL）架構。

（4）非同步任務完成造成的時序混亂

不同代理的：

- 任務生成時間
- 計算時延
- 任務完成時間

皆不同，使得決策時點不一致，形成 Semi-Markov Game（SMG）。常見的 MARL 演算法假設代理同步，因此無法直接應對此特性，需設計新的非同步資料蒐集與學習機制。

1.3 研究目的與貢獻

為解決上述挑戰，本研究提出一套整合 **分數階深度強化學習、能量虛擬佇列控制與非同步多代理架構** 的 AoI 最佳化 MEC 系統。主要目的與貢獻如下

（1）在原作者的基礎上建立具能量虛擬佇列約束的分數階多代理強化學習架構

本研究引入能量虛擬佇列 Z(t) 來約束能量使用以應對能量不足之場景，並以瞬時漂移 ΔZ 量化能量壓力，形成「漂移加懲罰（Drift-Plus-Penalty, DPP）」優化機制。此方式能兼具：

- 長期能量穩定性
- AoI 最小化
- 任務卸載與更新策略的自適應調整

實現 MEC 中能量與資訊即時性的雙重最佳化。

（2）實驗證明在能量受限下可顯著降低 AoI

實驗顯示本研究提出之方法可達到：

- 能量虛擬佇列保持穩定，避免能量耗盡
- 較同步 MARL 更具穩健性與可擴展性

上述結果證明本研究在能量受限的非同步 MEC 環境中具備實際應用價值。

二、專題內容與進行方法

2.1 系統機制與功能架構

2.1.1 系統機制與策略

本系統旨在於行動邊緣運算 (MEC) 環境中，模擬並訓練多個行動裝置（IoT agents）。核心目標是透過多代理強化學習 (MARL) 實現分散式策略的自我學習，以應對資源約束下的動態決策。

代理學習的關鍵策略 (Action Space)：

- 資源卸載決策： 決定任務是在本地端處理，還是卸載至可用的邊緣伺服器進行計算。
- 資訊更新決策： 確定生成新任務和上傳數據的時間間隔，以最佳化資訊年齡 (AoI)。
- 能量控制與隊列穩定化： 實時追蹤能量消耗與收穫，並依據能量虛擬佇列 Z 的狀態，調整傳輸與更新策略，確保資源利用的長期可持續性。

透過 MARL 框架，系統能逐步收斂至一種聯合優化策略，在保證資訊新鮮度的同時，實現能量使用的穩定性。

2.1.2 主要功能模組

1. 環境模擬模組

   建立一個高保真度的行動邊緣運算環境，涵蓋以下系統動態：

   - 資源動態建模： 模擬具備太陽能收穫、傳輸延遲、排隊延遲和能量消耗的複雜 MEC 資源模型。
   - 狀態與漂移動態： 建立包含排隊狀態、等待時間、能量收穫與虛擬佇列 Z 漂移的完整環境狀態空間。

2. 強化學習決策模組

   本模組採用異構學習方式，結合兩種獨立但協同的深度強化學習模型，形成雙決策路徑：

   R-D3QN (Double Dueling Deep Q-Network)：

   負責決定任務的計算位置（即本地執行或選擇性地卸載到特定邊緣伺服器）。D3QN 透過序列記憶體（如 GRU/RNN）捕捉非同步環境中的歷史依賴性。

R-PPO (Proximal Policy Optimization)：

負責調整任務產生或傳輸的間隔時間（即等待時間 wait）。PPO 根據當前環境狀態與虛擬佇列 Z 的反饋，動態調整任務更新的頻率策略。

兩種模型協同作用，D3QN 處理執行地點的空間選擇，而 PPO 處理執行時機的時間選擇，共同形成完整的資源分配與 AoI 最小化策略。

2.2 系統模型

2.2.1 資訊年齡(Age of Information, AoI)

資訊年齡 $\Delta_m(t)$是衡量單個代理 m 資訊新鮮度的關鍵指標

對於行動裝置 m，全域時鐘 T 的 AoI 定義為

$$\Delta_m(t) = t - T_m(t), \quad \forall m \in \mathcal{M},\ t \geq 0$$

其中 $T_m(t)$ 表示最近完成任務的時間。

完成任務 k 的總時長表示為 $Y_{m,k} \triangleq t\prime_{m,k} - t_{m,k}$

因此與時間間隔 $[t_{m,k}, t_{m,k+1}]$ 相關的梯形面積定義為

$$A(Y_{m,k}, Z_{m,k+1}, Y_{m,k+1}) \triangleq \frac{1}{2}(Y_{m,k} + Z_{m,k+1} + Y_{m,k+1})^2 - \frac{1}{2}(Y_{m,k+1})^2$$

而每個行動裝置 m 的目標就是最小化每個裝置 $m \in M$ 的時間平均 AoI。

$$\Delta_m^{(\text{ave})} \triangleq \liminf_{K\to\infty} \frac{\sum_{k=0}^{K} \mathbb{E}[A(Y_{m,k}, Z_{m,k+1}, Y_{m,k+1})]}{\sum_{k=0}^{K} \mathbb{E}[Y_{m,k} + Z_{m,k+1}]}$$

2.2.2 Lyapunov 函數與漂移定義 Lyapunov Drift-Plus-Penalty (DPP)

為了量化系統的不穩定程度和資源約束的累積違規，我們引入 Lyapunov 函數 $L(Q(t))$

$$L(Q(t)) = \frac{1}{2}Q(t)^2$$

其中 $Q(t)$ 是虛擬佇列（或資源約束追蹤隊列）在 t 時刻的長度。此函數用於衡量系統的不穩定程度（即隊列累積）。$L(Q(t))$ 越大，表示系統偏離穩定狀態越遠。

定義瞬時漂移量 $\Delta L(t)$ 為 Lyapunov 函數在一個時隙內的變化：

$$\Delta L(t) = L(Q(t+1)) - L(Q(t))$$

$\Delta L(t)$ 表示系統不穩定度在當前時隙的瞬時變化，我們旨在最小化 $\Delta L(t)$ 以確保長期穩定性。

2.3 能量收穫與消耗

本專題將能量收穫和消耗分別設定成行動裝置本地 CPU 運算的消耗 (E_{loc})、傳輸到邊緣伺服器的傳輸消耗(E_{tx})和太陽能板的能量收穫 (E_{harv})。

我們將行動裝置在傳輸和本地運算過程中的總能耗 E_{use} 定義為兩者的總和：

$$E_{use} = E_{loc} + E_{tx}$$

1. 本地運算能耗(E_{loc})：

   E_{use} 代表行動裝置在本地 CPU 運算時的能耗。E_{use} 可由下式計算：

   $$E_{\text{local}} = kf^2C$$

   其中，k為晶片常數、f為 CPU 頻率(Hz)、C為計算所需週期

2. 傳輸能耗(E_{tx})：

   E_{tx} 代表將本地資訊包傳輸到邊緣伺服器的能耗。我們將其簡化為與傳輸的資訊量 B 相關：

   $$E_{\text{tx}} = e_{\text{bit}} \times B$$

   其中e_{bit}為每 bit 傳輸能量(J/bit)、B為任務大小（bit）

3. 太陽能板的能量收穫(E_{harv})：

   $$E_{\text{harv}} \sim U(h_{\min}, h_{\max})$$

   其服從$U(h_{min}, h_{max})$均勻分布(J)

2.4 實驗設計

2.4.1 虛擬佇列

我們定義能量使用的總消耗 $E_{use} = E_{loc} + E_{tx}$

將虛擬佇列設計為

$$Z(t+1) = \max(Z(t) + E_{\text{use}} - E_{\text{harv}}, 0)$$

而漂移量定義成

$$\Delta Z = Z(t+1) - Z(t)$$

使用能量消耗減去能量收穫的形式紀錄能量虧損，進一步控制能量使用

2.4.2 獎賞函數(Reward Function)

我們將瞬時獎勵 R 旨在聯合最小化資訊年齡 (AoI) 與長期資源約束。此設計由 Lyapunov 漂移懲罰 (DPP) 原理啟發，將優化目標轉化為瞬時最大化獎勵：

$$R = -(AoI + \lambda_1 \Delta Z + \lambda_2 Z)$$

獎勵函數使用負號來將 AoI 和虛擬佇列 (Z ,ΔZ) 這些最小化目標，轉換為符合強化學習最大化獎勵的機制。

ΔZ 代表虛擬佇列的瞬時漂移量。它作為一個即時的反饋信號，追蹤資源約束的履行情況：

> 當 ΔZ 為負時（虛擬佇列下降），表示該時隙的資源消耗低於長期平均預算。此時，$-(\lambda_1 \Delta Z)$ 為正值，提升總獎勵 R，鼓勵行動裝置採取節省資源的行動或進行傳輸。
>
> 當 ΔZ 為正時，表示資源消耗超標，$-(\lambda_1 \Delta Z)$ 為負值，對總獎勵進行懲罰，以維持約束的穩定。

Z 項的作用： 引入虛擬佇列 Z 本身 $-(\lambda_2 Z)$ 是為了提供一個額外的抑制力，防止虛擬佇列累積過高，從而確保資源約束能夠更快速地被追蹤與滿足。

權重 λ_1 和 λ_2 用於調節 AoI 作為主要性能指標與資源約束追蹤之間的權衡。

三、專題流程與架構

3.1 系統架構圖

3.1.1 行動邊緣運算(MEC)系統

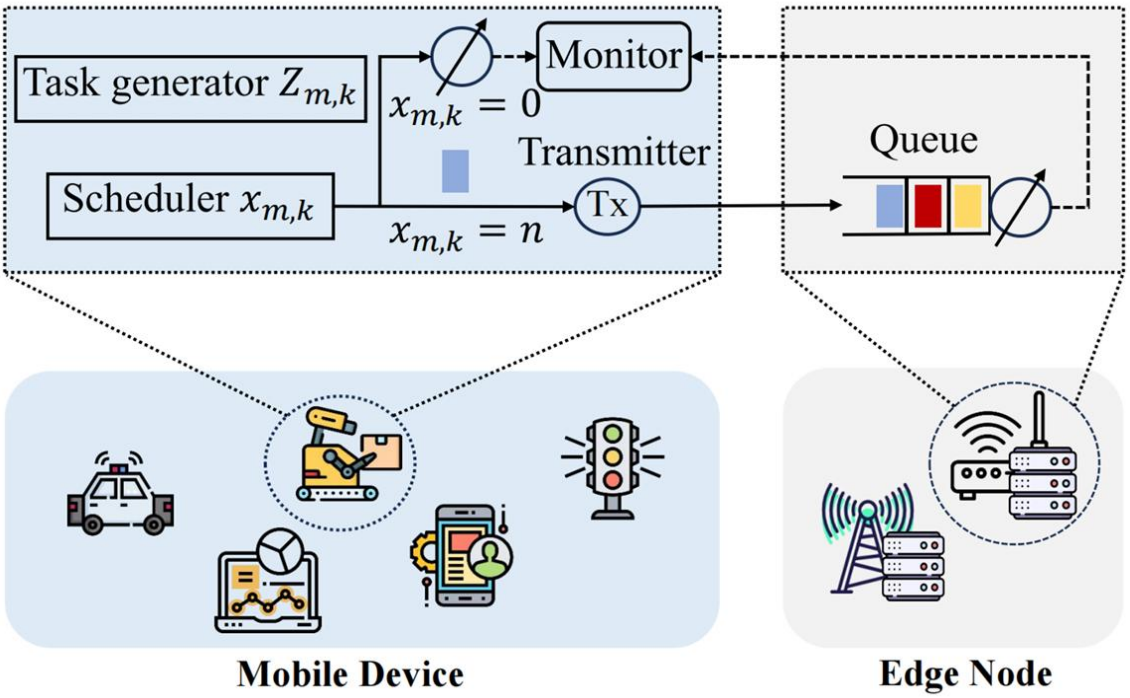


圖 1: MEC 系統架構

首先，在行動裝置 (Mobile Device) 有一個任務產生器 (Task generator)，這個模組會不斷產生新的計算任務，接著這些任務會交給排程器(Scheduler)決定要怎麼處理。

當排程結果是 0 的時候，表示這個任務會在本地裝置上執行，經過監測器 (Monitor) 監控，而當排程結果是 n 的時候，表示這個任務會透過發送器 (Transmitter) 傳送到邊緣節點 n 去執行。

任務會被接收並排入佇列 (Queue)，圖中不同顏色的方塊代表來自不同裝置卸載的任務，邊緣節點 (Edge Node) 會處理這些任務，減輕行動裝置的運算壓力，並提供更快的回應時間。

3.1.2 多行動裝置於 MEC 環境中決策之流程

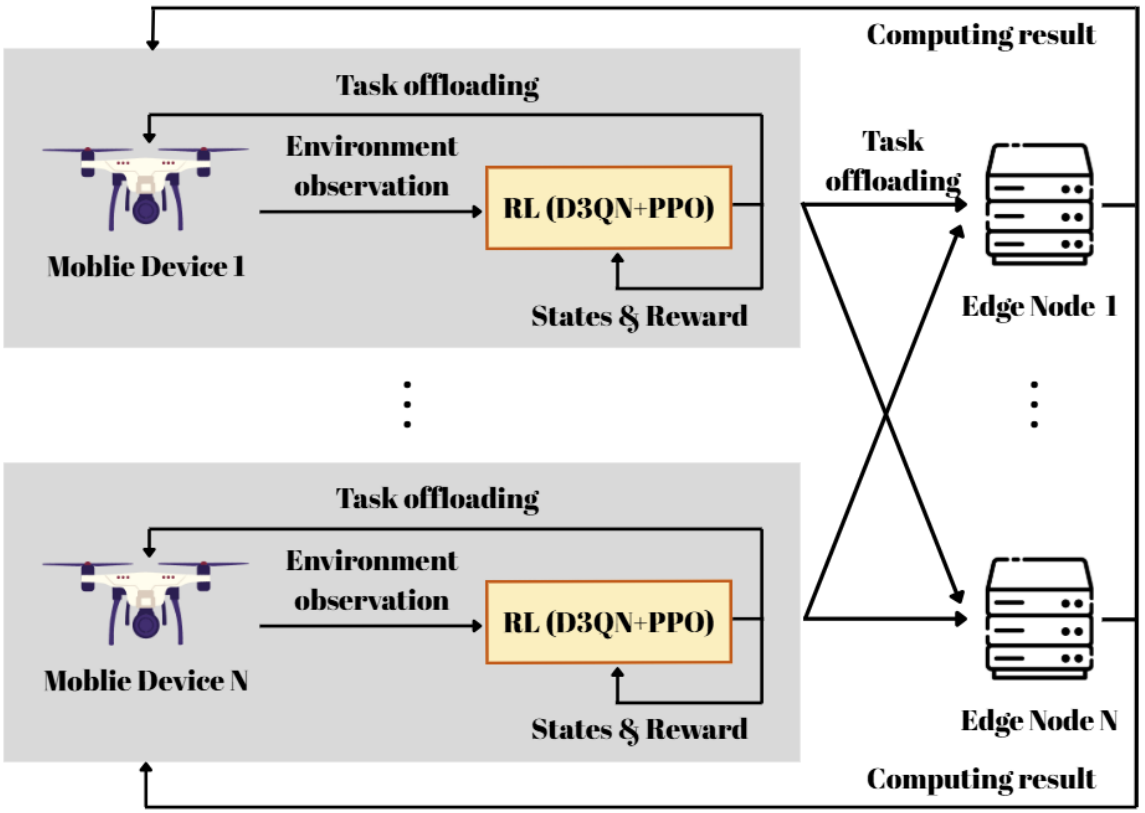


圖 2:整體系統的運作概念

系統由多個行動裝置(例如:無人機)與多個邊緣節點組成，每個裝置需要執行計算任務，由強化學習模型 (D3QN+PPO) 做決策，決定是否將任務在本地運算或卸載 (offload) 到哪一個邊緣節點。

每個行動裝置都是一個 RL agent，因此這是一個多代理系統 (Multi-Agent Reinforcement Learning)，不同行動裝置之間會競爭邊緣資源，因此系統需要考慮彼此的行為，避免所有裝置都集中將任務送往同一個邊緣節點，造成延遲與排隊時間暴增，在過程中藉由觀察環境狀態、回饋 (reward)，會逐步學習如何做出最優任務分配。

四、專題成果介紹

4.1 數據比較

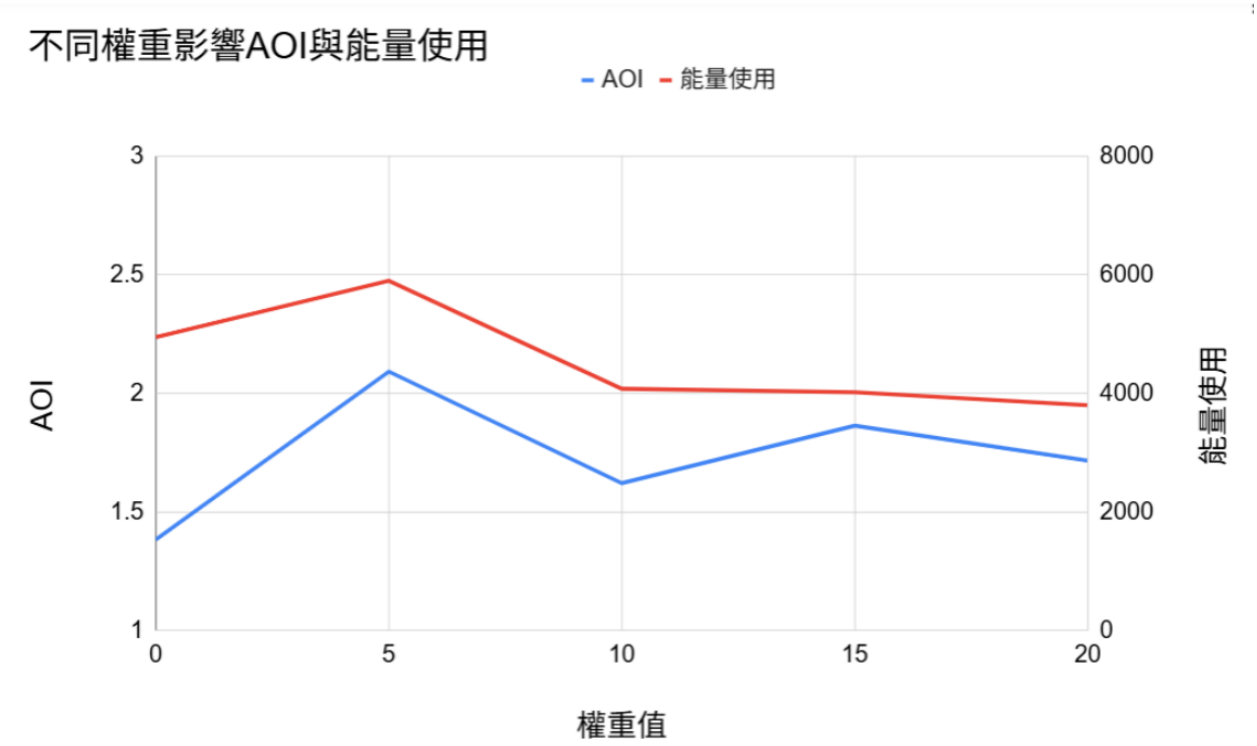


圖 3: AoI 和總能量使用在不同權重下後一百回合訓練的平均值

系統在權重 10～20 之間能達到能量與 AoI 的平衡，能量使用逐步下降，但 AoI 仍維持良好表現。

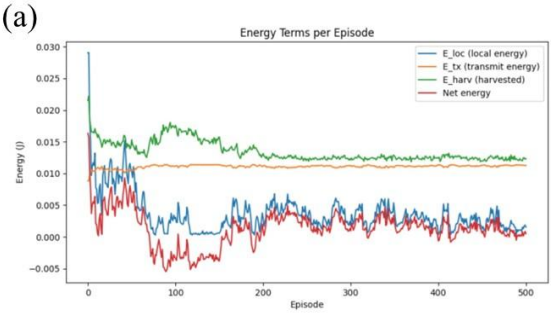
能量使用量從 4946 焦耳(能量虛擬佇列的權重值為 0)降到 3796 焦耳(能量虛擬佇列的權重值為 20)，降低約 **23%**。

所以適度調整權重值可以讓系統在資訊新鮮度(AoI)與能量消耗之間找到最佳平衡點，使系統能長期穩定運作。

4.2 系統畫面

4.2.1 平均能量

(a)



(b)

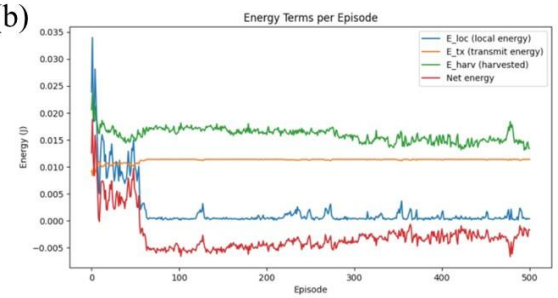


圖 4:各 Episode 的平均能量(包括本地運算、傳輸、收穫與淨變化)
(a)無 (b)有(權重值為 20) 啟用能量虛擬佇列之懲罰

圖 4 中的藍線為本地運算之能量、橘線為傳輸之能量、綠線為能量收穫、紅線為能量之淨變化(能量虧損:能量使用減收穫)，也就是藍線加橘線減綠線。

經過實驗發現啟用能量虛擬佇列之懲罰，圖 4(b)中的紅線會穩定小於 0，代表現在能量收穫逐漸大於能量使用，系統正在慢慢平衡能量的使用與收穫。

統計可見在能量收穫量不及使用量的情況下，系統學會節省能量使用並將使用量降低至收穫量。

4.2.2 平均 AoI

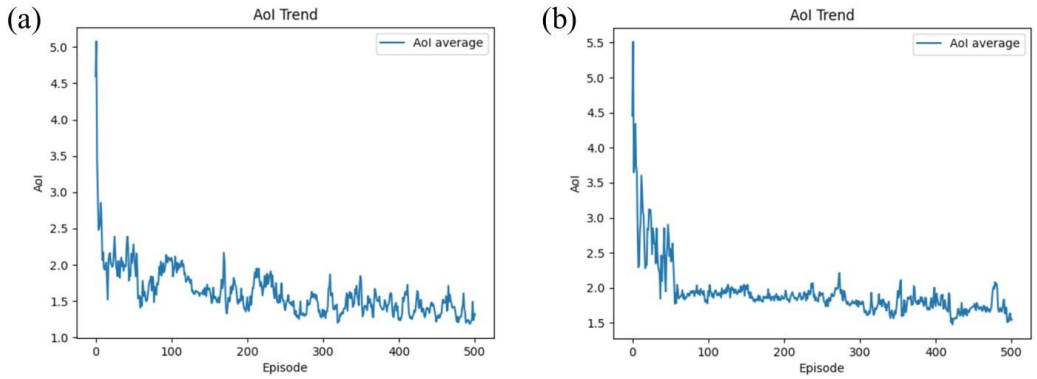


圖 5:平均 AoI 隨 Episode 的變化
(a)無 (b)有(權重值為 20) 啟用能量虛擬佇列之懲罰

圖 5(a)的平均 AoI 為 1.38 比(b)的 1.71 略低，原因是因為沒有啟用能量虛擬佇列之懲罰時，Agent 的策略會傾向更頻繁地計算或傳輸以直接最小化 AoI，因此平均 AoI 較低；但這是以違反能量約束(能量虛擬佇列 Z 膨脹、不穩定)為代價。

啟用能量虛擬佇列之懲罰後，為了保持 Z 穩定(滿足能量約束)，Agent 變得較保守，減少計算或傳輸頻率，從而導致平均 AoI 輕微上升。

4.2.3 能量虛擬佇列

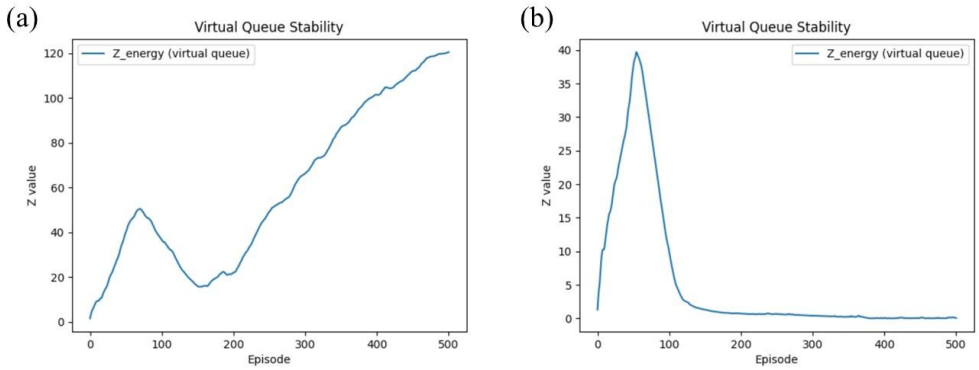


圖 6:能量虛擬佇列隨 Episode 的變化
(a)無 (b)有(權重值為 20) 啟用能量虛擬佇列之懲罰

能量虛擬佇列模擬一個違反約束的積累量，把「長期約束」變成「佇列穩定性」問題，因此要越小才能趨於穩定。

圖 6(b)可以發現，啟用能量虛擬佇列之懲罰的情況下，能量的虛擬佇列從原本的沒有收斂變成有所收斂。

五、專題學習歷程介紹

5.1 專題相關軟體學習介紹

本章在說明我們在本專題的開發過程中，為了建構「基於虛擬佇列約束的分數階多代理強化學習 MEC 系統」所進行的軟體工具學習、模型調適及系統實作歷程。我們需掌握深度強化學習、多代理系統與環境模擬等多項技術。同時，針對研究過程中所遭遇的技術與系統層面的挑戰進行回顧，並說明具體的解決方法，以呈現專題從理論理解到系統落地的完整軌跡。

（1）Python 深度學習生態系統與 PyTorch 建模能力

- 本研究以 PyTorch 作為主要深度學習框架，我們深入學習張量運算，以支援 D3QN 與 PPO 演算法。
- 在分數階與虛擬佇列等具時間依賴性的狀態下，我們須掌握 RNN、GRU 等時間序列模型，並於 PyTorch 中實作隱藏狀態傳遞以及避免梯度爆炸的技巧。
- 透過實作，我們熟悉各類 optimizer（Adam）、loss function（policy loss）、learning rate scheduler 等工具，用於穩定多代理訓練。

（2）強化學習演算法深度實作：D3QN 與 PPO 雙架構

- 透過主程式 main_MA_DQN_PPO.py，我們實作 D3QN 所需的核心機制，包括：
    a. Double Q-learning 避免 Q-value 高估
    b. Dueling Network 分離 Value 與 Advantage
    c. Experience Replay 結構化記憶管理
- 在 PPO 政策網路部分，我們學習：
    a. Actor–Critic 協同訓練機制
    b. Entropy bonus 增強策略探索能力
- 透過推導與實作，我們能明確理解混合動作空間（卸載／更新）如何以 D3QN 與 PPO 協作完成。

（3）非同步多代理訓練機制（Asynchronous MARL）

- 在 mec_env_test.py 中，我們建立一套事件驅動（Event-driven）模擬系統，使代理在任務完成、能量更新或通訊回傳時計算下一步動作。
- 我們需理解 Semi-Markov Game（SMG）理論，並重新定義動態決策時序，使不同代理可在不同時間步獨立進行策略更新與資料蒐集。
- 此外，我們建立多代理 Replay Buffer 管理策略，確保 RNN 的時間序列資料不被破壞，並確保 PPO 與 D3QN 使用的記憶分離且穩定。

（4）Lyapunov 虛擬佇列與漂移加懲罰（DPP）機制實作

- 團隊學習虛擬佇列 Z(t) 的設計方式，包含能量模型、消耗模型、能量收穫模型，並於環境中以參數化形式實作。
- 理解 Lyapunov Drift 與 Drift-Plus-Penalty 理論，並將瞬時漂移 ΔZ 與能量懲罰項整合至 reward，改寫 RL 目標，使其同時考量：
    a. 資訊新鮮度（AoI）
    b. 能量穩定性
- 團隊反覆調整 DPP 權重，使整體學習過程兼具穩定性與最優性。

（5）資料視覺化與系統調參能力

- 我們使用 Matplotlib 與 Pandas 分析：
    a. AoI 變化
    b. 能量虛擬佇列趨勢
    c. 更新頻率分布
    d. 卸載決策比例
- 這些圖表協助我們進行模型診斷、調整 learning rate、discount factor、fractional coefficient 等關鍵參數。
- 版本控制採用 Git 並搭配 Spyder，使團隊能有效協作並維護多代理架構。

5.2 專題製作過程遭遇的問題與解決方法

本專題整合能量虛擬佇列、分數階 RL、非同步 MARL 與混合動作空間，使開發過程面臨多項技術挑戰。

（1）分數型 AoI 目標無法直接以 RL 學習，導致 reward 震盪與策略發散

問題說明：

AoI 屬於「延遲／更新頻率」的分數型目標，並非 RL 能穩定處理的累積獎勵，因此初期直接使用 AoI 會導致 reward 噪音極大，策略反覆震盪。

解決方法：

- 採用 Dinkelbach's Reformulation 將分數目標轉換為可學習的差值形式。
- 加入 Lyapunov Drift 與能量懲罰，使 reward 從「不穩定」變為「具穩定性與物理意義」。

（2）能量虛擬佇列 Z 易爆增，造成 reward 不收斂

問題說明：

初期能量收穫模型波動過大，導致 Z 長期累積並失控，破壞強化訓練穩定性。

解決方法：

重新設計調整合理化各項權重，同時確保能量懲罰不壓過 AoI 目標。

六、結論與未來展望

6.1 結論

隨著行動邊緣運算（Mobile Edge Computing, MEC）在智慧城市、車聯網、工業物聯網與感測網等領域的應用日益普及，如何在能量受限且多代理非同步的環境中維持資訊的新鮮度（Age of Information, AoI）已成為關鍵的研究問題。本研究針對此挑戰，提出一套結合 **分數階多代理深度強化學習（Fractional Multi-Agent DRL）**、**虛擬佇列（Virtual Queue）** 與 **Lyapunov 漂移加懲罰（Drift-Plus-Penalty, DPP）** 的綜合性決策框架。

本研究的核心貢獻在於成功整合 **事件驅動的分數階學習機制** 與 **Lyapunov 能量控制理論**，使每個代理能根據能量狀態、佇列動態與網路負載做即時調整，從而在非同步的 MEC 系統中仍保持穩定收斂與高效更新。透過 D3QN 與 RNN-PPO 所構成的雙層決策架構，系統可同時處理卸載動作（offloading）與更新間隔（update scheduling），在空間與時間兩個面向達成最佳化。

實驗結果顯示，本研究方法能有效降低平均 AoI、節省能量消耗，並使虛擬佇列維持穩定，證實所提出架構在多代理環境中的可行性與優越性。綜合而言，本專題驗證了一個具備理論基礎、可擴展且具實務潛力的 MEC 控制策略，為未來在即時資訊系統中導入分數階深度強化學習與虛擬佇列控制提供重要借鏡。

6.2 未來發展方向

本研究所建立的分數階 MARL 加上 虛擬佇列架構具高度延展性，未來可應用至更多元、複雜且受限的運算環境。以下提出三項實際需求的發展方向：

（1）太陽能無人機（Solar-Powered UAV）自主通訊與任務協作

應用需求：

救災、偵察與邊境巡防等任務通常需同時蒐集影像、紅外線偵測、生命跡象分析等大量資訊，而無人機受限於太陽能供電，使能量與更新頻率之間的平衡變得格外重要。過時資訊可能導致任務判斷延誤，降低環境感知的可靠度。

本研究系統的貢獻：

利用虛擬佇列能量控制結合分數階策略更新，能為每台無人機自動規劃：

- 是否採用本地處理或卸載至邊緣站台
- 動態調整資料感測與上傳頻率
- 在能量有限下維持最高資訊即時性

此架構有助於建立高韌性、高省電且具協作能力的 UAV 群體通訊系統。

（2）智慧城市大規模感測網（Smart City Sensing Network）

應用需求：

智慧交通燈號、空氣品質監測、道路狀態偵測等感測器若更新延遲，將直接影響交通控制與城市治理品質。感測設備常依靠能量收穫（如太陽能微型板）維持運作，須避免頻繁更新導致能量枯竭。

本研究系統的貢獻：

透過多代理分數階 RL，可使每個感測節點依據：

- 電量
- 網路排程
- 當前佇列負載

  自動調整資料更新頻率並決定是否傳輸至邊緣伺服器。此方法能在不降低資訊新鮮度的前提下節省城市感測網的總體能耗。

（3）緊急醫療與穿戴式健康監控（Wearable and e-Health Monitoring）

應用需求：

穿戴式生醫裝置包含心率、血氧、呼吸監控等生命訊號，若資料傳輸延遲，即便只有十數秒，都可能錯過診斷或急救的黃金時機。然而此類設備多依賴小型電池或微型能量收穫模組，必須兼顧長期續航與即時傳輸。

本研究系統的貢獻：

系統可根據生理狀態與能量條件智慧調整更新策略：

- 正常生理狀態：降低傳輸頻率以延長續航
- 出現異常時：立即提升更新頻率與動態切換高優先通訊模式

在節能與即時性之間取得最佳平衡，提升穿戴式醫療設備的可靠性。

七、參考文獻

[1] Asynchronous fractional multi-agent deep reinforcement learning for age-minimal mobile edge computing, May 2025.

[2] Minimizing Version Age of Information in Resource-Constrained Multi-Source Systems via Lyapunov and Learning Based Scheduling.

靜宜大學
資訊工程學系

分數多代理深度強化學習結合能量虛擬佇列於行動邊緣運算之AOI與能量最佳化

西元二〇二四年十二月