Inatel
Instituto Nacional de Telecomunicações

Decentralized Deep Reinforcement Learning Approach for Channel Access Optimization

Sheila Cássia da Silva Janota Cruz
sheila.cassia@mtel.inatel.br

Felipe A. P. de Figueiredo
felipe.figueiredo@inatel.br

Instituto Nacional de Telecomunicações (INATEL), Santa Rita do Sapucaí - MG

O problema

- O mecanismo de múltiplo acesso adotado pelo padrão IEEE 802.11 (CSMA/CA), utiliza o algoritmo de *backoff* exponencial binário (BEB) para **evitar colisões**.
- A cada **nova colisão**, ele **aumenta o tempo de espera** para uma nova transmissão.
- Assim, o BEB aumenta a latência da rede, reduzindo o *throughput*.
- Esse problema se torna pior em situações de alta carga (i.e., muitas estações), pois a **chance de colisões aumenta**.

O BEB não é ótimo, especialmente em redes densamente povoadas.

Portanto, precisa-se de uma solução que aprimore a prevenção de colisões.

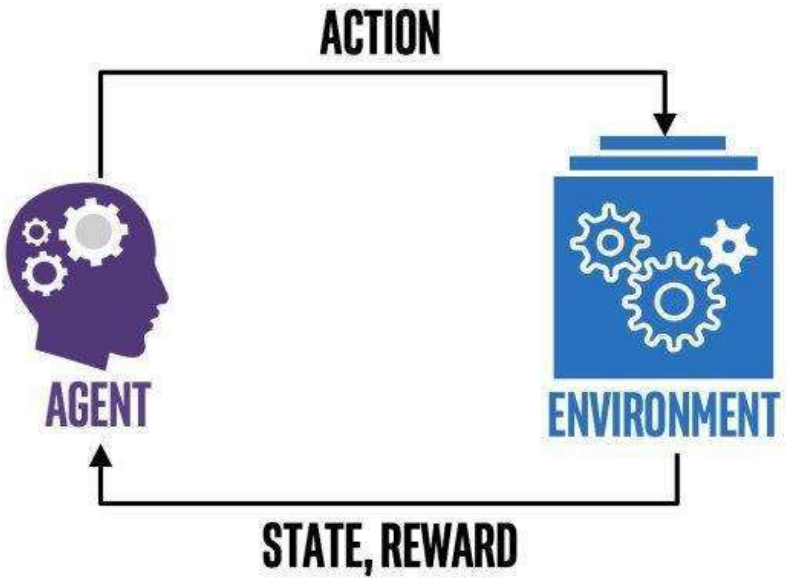
A hipótese

Uma abordagem **descentralizada** utilizando algoritmos de ***deep reinforcement learning*** pode otimizar o desempenho da rede, **reduzindo colisões e aumentando seu *throughput*** total.

Por que uma solução descentralizada?

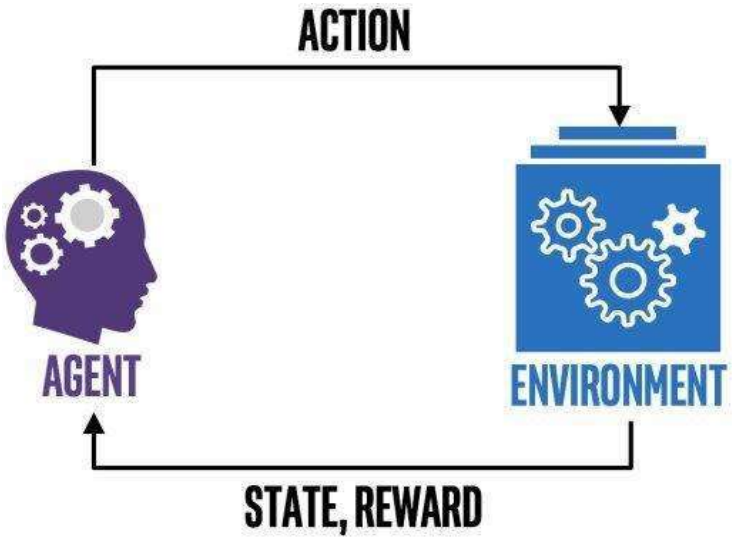
- Permite que as **janelas de contenção** (CW) sejam **alteradas localmente**, **reduzindo a latência** causada pela comunicação com o *access point* (AP).
- Permite que **cada estação** (STA) **ajuste seu CW** de forma **independente com base nas condições locais** (e.g., taxa de colisão, tamanho da fila de Tx, etc.).
  - Isso torna a **rede mais adaptável** às mudanças, como variações no **número de estações** ou na **demanda de tráfego** de **cada** STA.

O que é reinforcement learning?



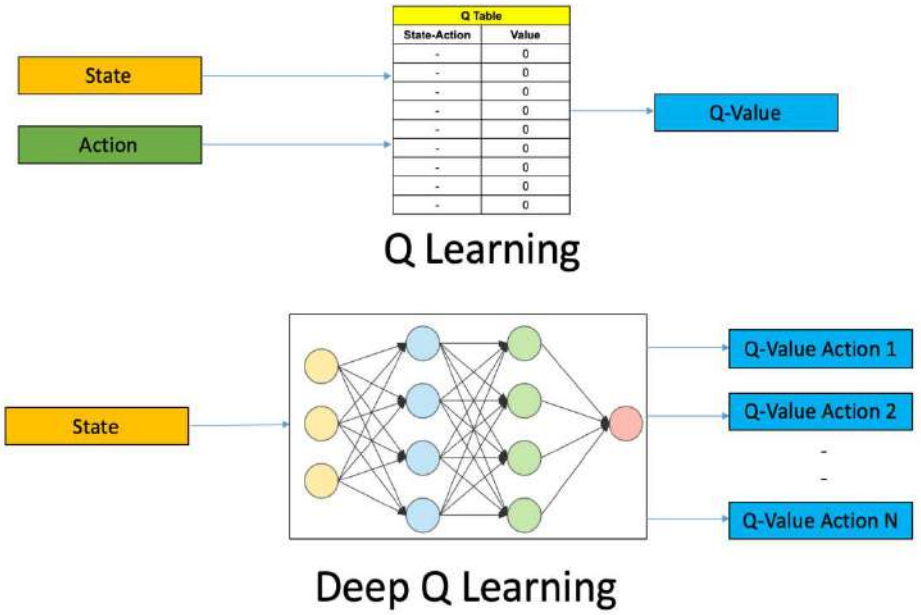
- Algoritmos de aprendizado por reforço (RL), chamados de **agentes**, aprendem como se **comportar em um ambiente** através de **interações do tipo tentativa e erro**.

O que é reinforcement learning?



- O agente **observa o estado** do ambiente, **seleciona e executa uma ação** e **recebe um reforço + ou -** em consequência da ação tomada.
- O **objetivo** do agente é aprender uma **estratégia**, chamada de **política**, que **maximize os reforços positivos** recebidos ao longo do tempo.

Deep Reinforcement Learning (DRL)

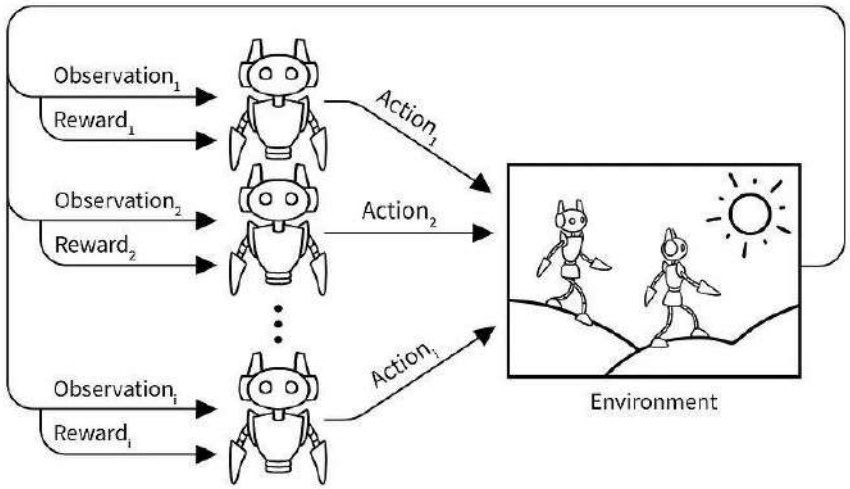


- **DRL** combina RL com redes neurais profundas para **lidar com estados e ações de alta dimensão**.

Os algoritmos de DRL mais conhecidos

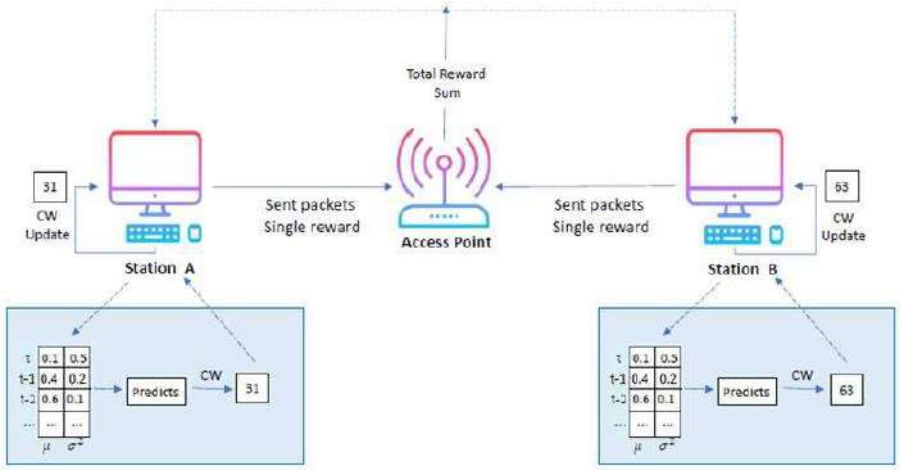
- DQN (*Deep Q-Network*) e DDPG (*Deep Deterministic Policy Gradient*) são dois algoritmos de DRL *off-policy* muito populares.
- Implementam *replay buffer* e *target networks* para maior estabilidade treinamento.
- As diferenças principais entre os dois são
  - DQN é adequado para problemas com espaços de ação discretos.
  - DDPG é adequado para problemas com espaços de ação contínuos.

Multi-Agent Reinforcement Learning (MARL)



- É uma subárea do RL.
- Estuda o comportamento de múltiplos **modelos** que **coexistem** em um **ambiente compartilhado**.
- Os **agentes agem** de forma **independente** com base em suas **observações locais**, mas **colaboram ou competem** para **maximizar** uma **recompensa global**.
- Assim, podemos implementar uma solução descentralizada.

Solução descentralizada



OBS.: Cada STA envia sua recompensa local para o AP.

- Propomos que o **CW de cada STA** seja **ajustado** de forma **independente** visando a **maximização do *throughput*** total da rede.
  - **Ambiente**: um rede 802.11ax com várias STAs e um AP.
  - **Agentes**: STAs com DQN ou DDPG.
  - **Ações**: CWs de cada STA: CW = $\lfloor 2^{(a+4)} \rfloor - 1$, onde a assume valor do intervalo [0, 6].
  - **Observações**: as médias e desvios padrão das taxas de colisão observadas por cada STA.
  - **Recompensa**: *throughput* total da rede normalizado.

Inatel
Instituto Nacional de Telecomunicações

www.inatel.br

Setup de simulação

Configuration Parameter	Value
Wi-Fi standard	IEEE 802.11ax
Number of APs	1
Number of static stations	5,15, 30 or 50
Number of dynamic stations	increases steadily from 5 to 50
Frame aggregation	disabled
Packet size	1500 [[bytes]
Max Queue Size	100 [packets]
Frequency	5 [GHZ]
Channel BW	20 [MHz]
Traffic	constant bit-rate UDP of 150 [Mbps]
MCS	HeMcs (1024-QAM with a 5/6 coding rate)
Guard Interval	800 [ns]
Propagation delay model	ConstantSpeedPropagationDelayModel
Propagation loss model	MatrixPropagationLossModel

Arquiteturas DQN e DDPG

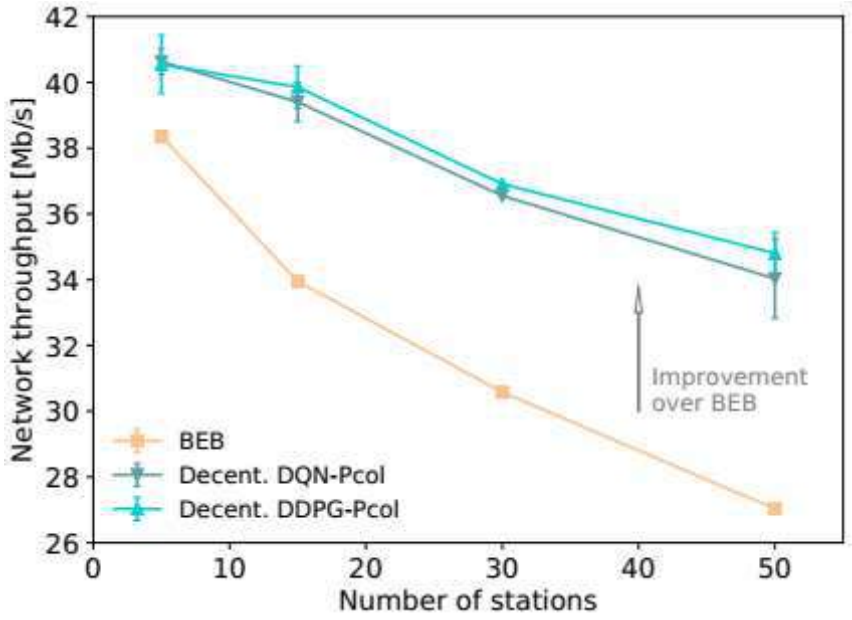
Layer	Units
LSTM	8 (relu)
DNN	128 (relu)
DNN	64 (relu)
DNN	7 (softmax)

Layer	Units
LSTM	2 (relu)
DNN	32 (relu)
DNN	1 (linear)

- Usamos o simulador NS-3 e o *framework* NS3-gym para criar o ambiente de treinamento dos agentes.
- **Cenários**: estático e dinâmico.
- STAs dispostas em um círculo de 1 [m] ao redor do AP.
- DQN e DDPG usam camadas LSTM e DNN com otimizador Adam.

Resultados com o cenário estático

Throughput versus número de STAs

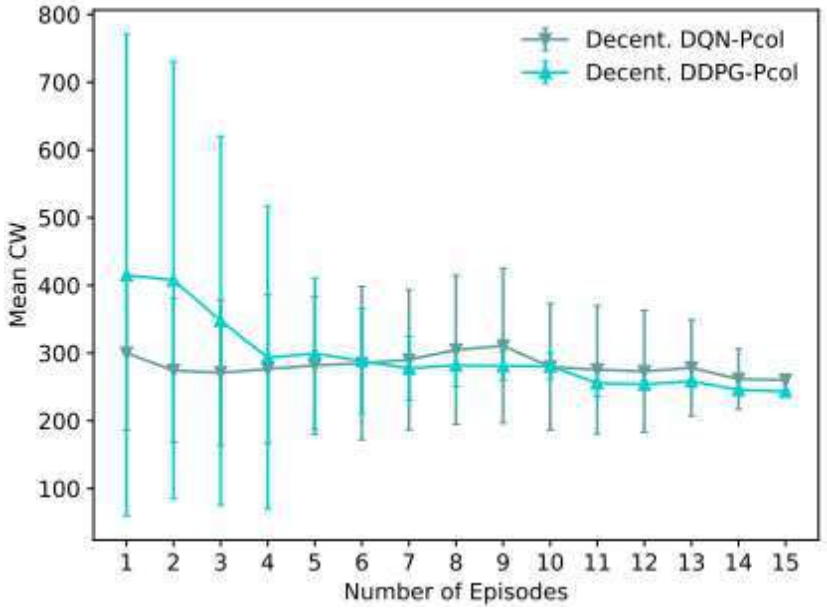


- Toda a simulação é feita com um número estático de STAs (5, 10, 30 e 50).
- O DDPG melhora o desempenho sobre o BEB entre 5,19% e 27,78% para 5 e 50 estações.
- O DQN melhora entre 5,19% e 27,10% no mesmo intervalo.
- O *throughput* de ambos reduz devido ao aumento da competição entre STAs (# de colisões).

Inatel
Instituto Nacional de Telecomunicações

www.inatel.br

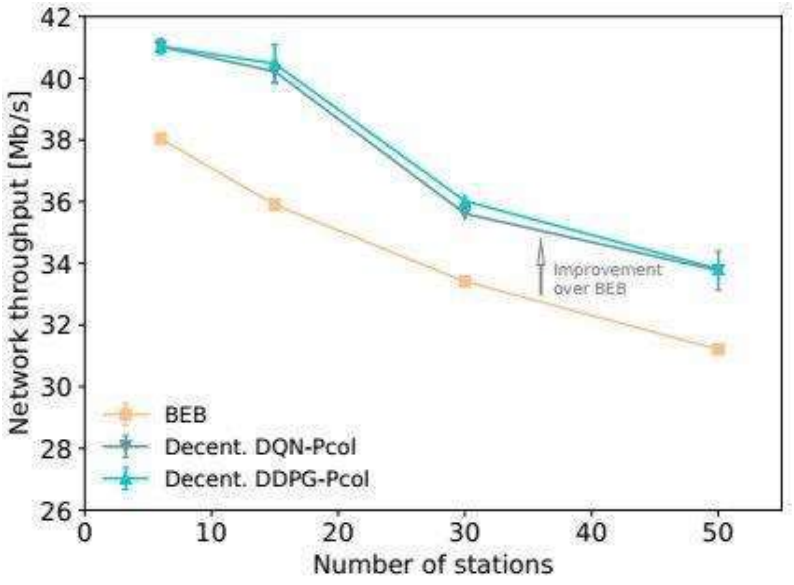
Valor médio de CW para 30 STAs versus episódios



- A figura apresenta a média e a variância do CW da trigésima STA.
- A média e a variância do CW diminuem conforme o agente aprende.
- Isso ocorre pois o número de ações aleatórias cai, indicando que o **agente aprendeu a escolher o melhor CW**.

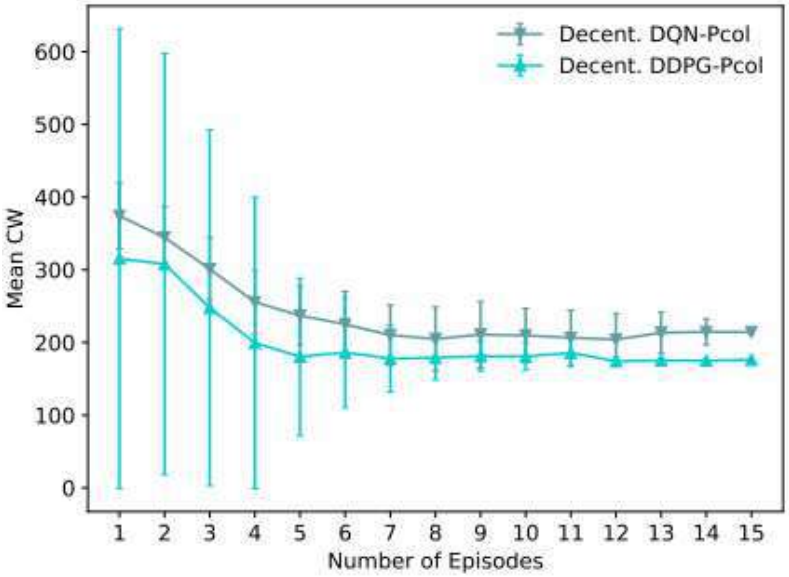
Resultados com o cenário dinâmico

Throughput versus número de STAs



- O número de STAs aumenta em 1 STA a cada 1.2 [s].
- DQN e DDPG apresentam melhorias similares em comparação com o BEB.
- Elas são de 7,89% e 8,43% sobre o BEB para 5 e 50 estações.
- Os agentes têm mais difculdade em aprender por ser um cenário mais complexo.
- O *throughput* reduz devido à maior competição (# de colisões).

Valor médio de CW para 30 STAs versus episódios



- A figura apresenta a média e a variância do CW da trigésima STA.
- A a média e a variância do CW diminuem conforme o agente aprende.
- Isso ocorre pois o **número de ações aleatórias diminui**, indicando que o **agente aprendeu a escolher o melhor CW**.

Conclusões

- Propusemos uma abordagem descentralizada usando DRL para a seleção independente dos CWs com o objetivo de reduzir colisões.
- Tal abordagem aumenta a vazão total da rede em relação ao BEB em cenários estático e dinámico.
- DDPG é melhor devido a sua maior granularidade.
- Pesquisas futuras envolvem o uso de outros algoritmos (e.g., *on-policy*), adoção de outras métricas de observação (e.g., níveis das filas de Tx) e uso de cooperação entre STAs.

Inatel
Instituto Nacional de Telecomunicações

Perguntas?

Inatel
Instituto Nacional de Telecomunicações

Obrigado!

Inatel
Instituto Nacional de Telecomunicações

BACKUP

Initialize $Q(s,a)$ arbitrarily
Repeat (for each episode):
  Initialize s
  Choose a from s using policy derived from Q (e.g., ε-greedy)
  Repeat (for each step of episode):
    Take action a, observe r, s'
    Choose a' from s' using policy derived from Q (e.g., ε-greedy)
    $Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma Q(s',a') - Q(s,a)\right]$
    $s \leftarrow s'$; $a \leftarrow a'$;
  until s is terminal

Figure 6.9: Sarsa: An on-policy TD control algorithm.

On-policy: a política de comportamento (que o agente segue para explorar o ambiente) e a política-alvo (que ele está aprendendo a otimizar) são as mesmas.

Initialize $Q(s,a)$ arbitrarily
Repeat (for each episode):
  Initialize s
  Repeat (for each step of episode):
    Choose a from s using policy derived from Q (e.g., ε-greedy)
    Take action a, observe r, s'
    $Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right]$
    $s \leftarrow s'$;
  until s is terminal

Figure 6.12:Q-learning: An off-policy TD control algorithm.

Off-policy: a política de comportamento (que o agente segue para explorar o ambiente) e a política-alvo (que ele está aprendendo a otimizar) não são as mesmas.

https://stackoverflow.com/questions/32846262/are-q-learning-and-sarsa-with-greedy-selection-equivalent

TABLE II

NS3-GYM AGENT CONFIGURATION PARAMETERS.

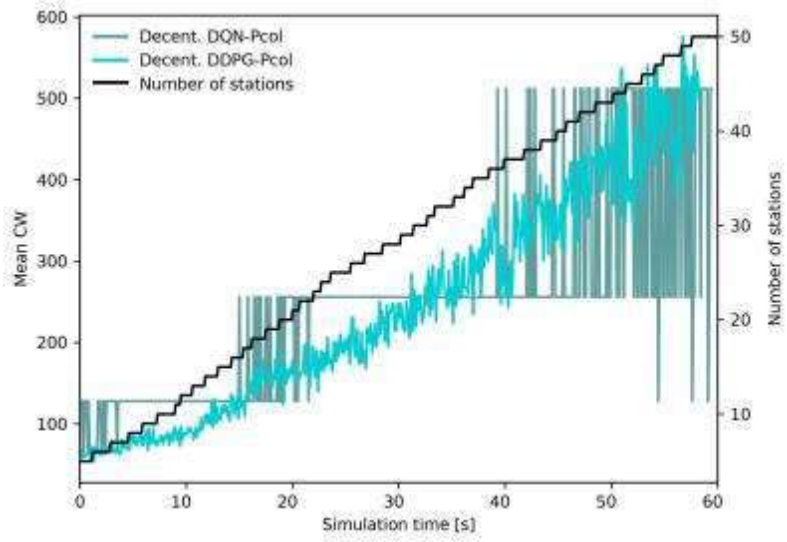
Configuration Parameter	Value
DQN's learning rate	4×10^{-4}
DDPG's actor learning rate	4×10^{-4}
DDPG's critic learning rate	4×10^{-3}
Reward discount rate	0.7
Batch size	32 samples
Replay memory size	18000 samples
Size of observation history memory	300 samples
$trainingPeriod$	840 [s]
$envStepTime$ (i.e., interaction interval)	10 [ms]

https://stackoverflow.com/questions/32846262/are-q-learning-and-sarsa-with-greedy-selection-equivalent

Inatel
Instituto Nacional de Telecomunicações

www.inatel.br

Valor médio de CW versus tempo e número de STAs



- À medida que o número de STAs aumenta, os valores de CW são ajustados adequadamente.
- Ambos aumentam o valor de CW conforme o número de STAs aumenta.
- Por ser discreto, o DQN oscila entre valores vizinhos.
- A abordagem contínua do DDPG acompanha melhor as mudanças da rede e alcança um CW final mais baixo.
  - A consequência é um maior throughput.