TACT: A Transfer Actor-Critic Learning Framework for Energy Saving in Cellular Radio Access Networks

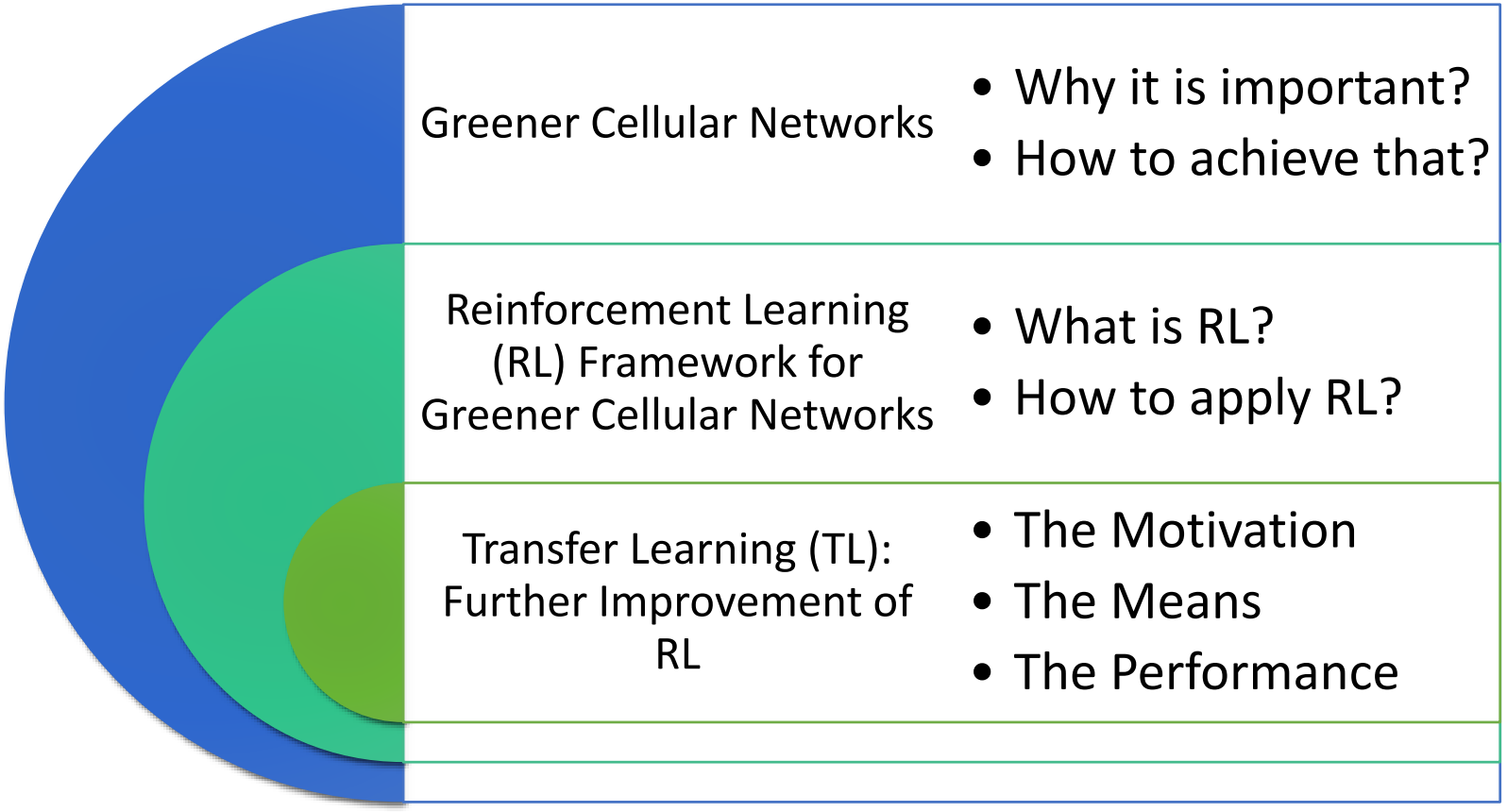
LI RONGPENG

ZHEJIANG UNIVERSITY

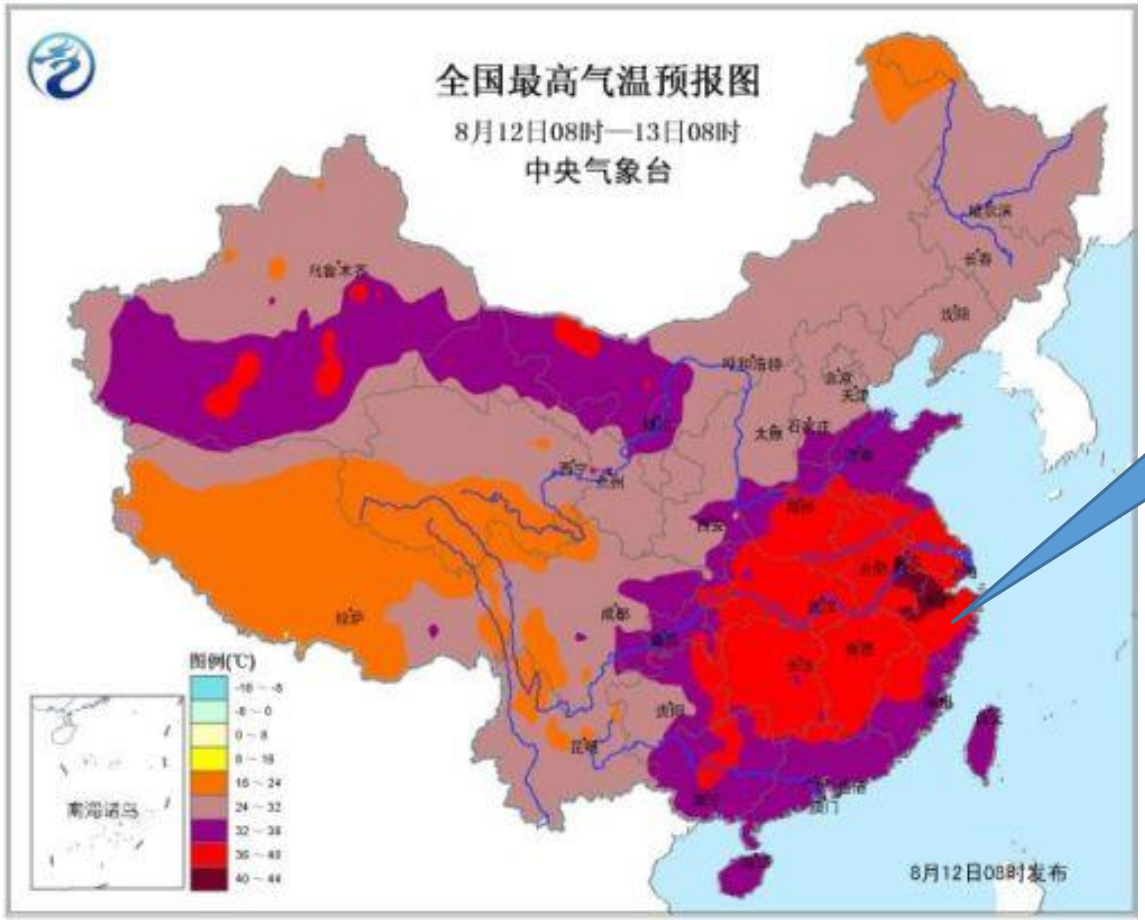
EMAIL: LIRONGPENG@ZJU.EDU.CN

WEB: HTTP://WWW.RONGPENG.INFO

Content

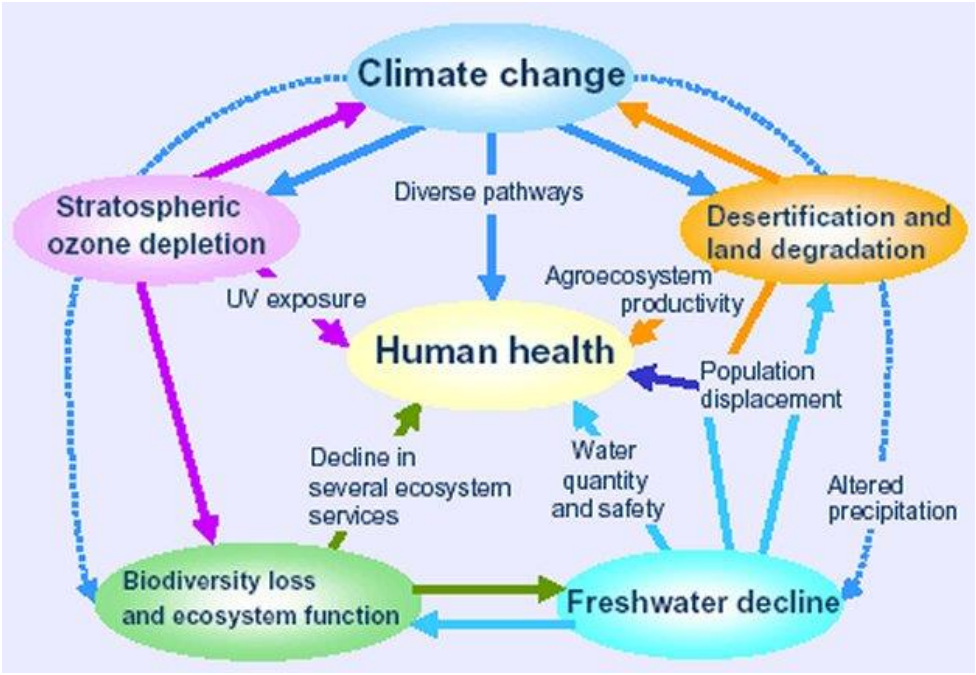

Greener Cellular Networks
• Why it is important?
• How to achieve that?
Reinforcement Learning (RL) Framework for Greener Cellular Networks
• What is RL?
• How to apply RL?
Transfer Learning (TL): Further Improvement of RL
• The Motivation
• The Means
• The Performance

Weather in Hangzhou, the So-called Paradise in China



Temperature of Hangzhou, Aug. 12
Highest: **41** ℃
Lowest: **28** ℃

Global Climate Change



Greener Cellular Networks: “Research for the Future”

So
He

Ultimate immersive experience & Data Explosion

Learning

Gaming

1000X Traffic Growth in 10 Years

Sharing

The Next-Generation Cellular networks

- Objectives:
  - **Green Wireless Network Requirement**
  - **Explosive Traffic Demands**
- Means:
  - More Power? Not Green!
  - More bandwidth? Limited Help
  - ✓ Advanced Physical Layer Technologies
    - ✓ Cooperative MIMO, Spatial Multiplexing, Interference Mitigation
  - ✓ Advanced Architecture
    - ✓ Cloud RAN, HetNet, Mass
  - ✓ "**Network Intelligence**" is th
    - ✓ Networks must grow and work where data is demanded.

Increase Power
Cooperative systems

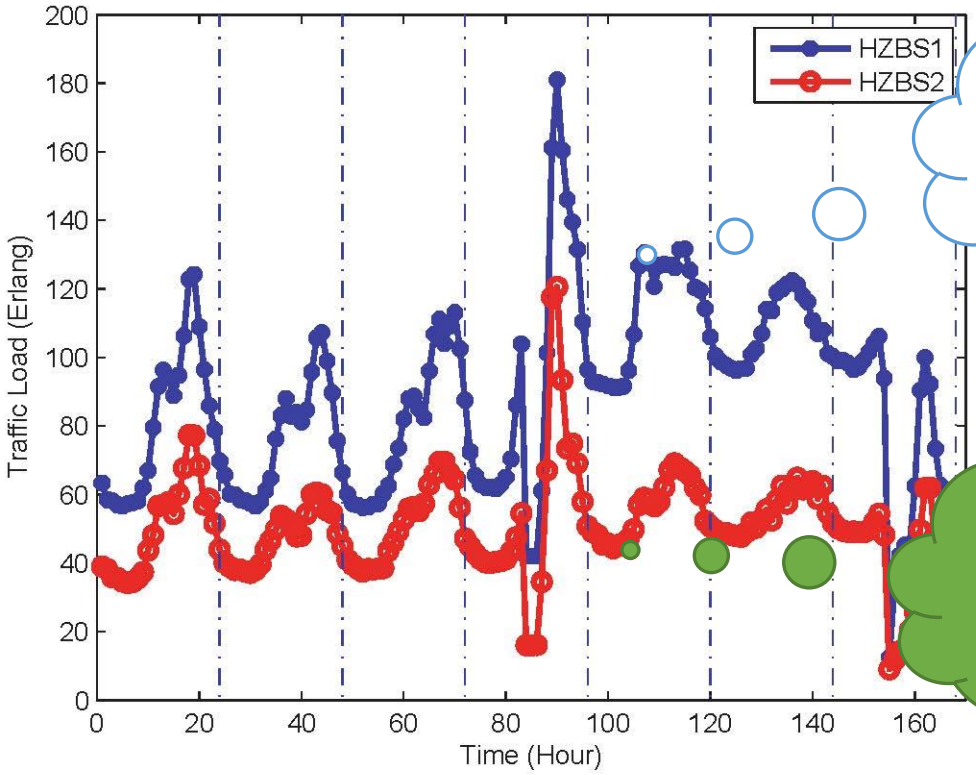
$$C = \sum_{Channels} B_i \log_2(1 + P_i / N)$$

More Channels
MIMO

Increase Bandwidth
Cognitive Radio

Courtesy to the public reports of Prof. Vincent Lau (Zhejiang University Seminar, March 2013) and Dr. Jeffrey G. Andrews (University of Notre Dame Seminar, May 2011).

Temporal Characteristics of Traffic Loads



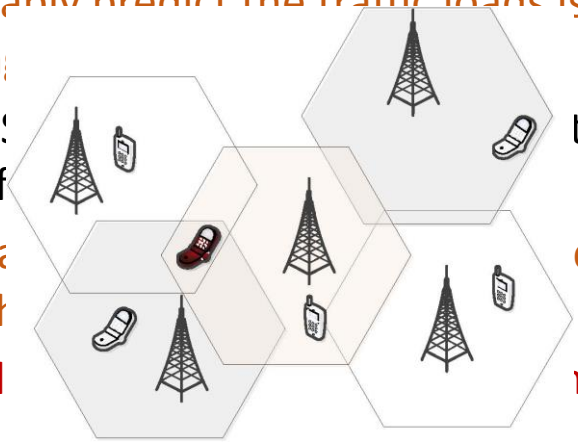
BSs are deployed on the basis of peak traffic loads.

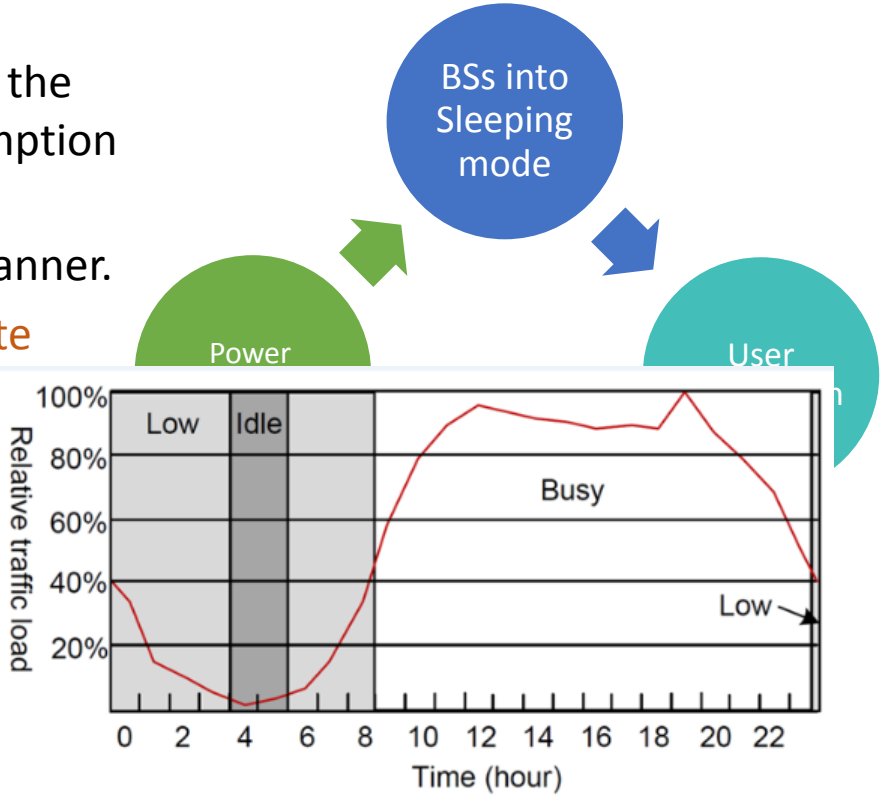
How to make it green at low traffic loads?

- Rongpeng Li, Zhifeng Zhao, Yan Wei, Xuan Zhou, and Honggang Zhang, "GM-PAB: A Grid-based Energy Saving Scheme with Predicted Traffic Load Guidance for Cellular Networks," in Proceedings of IEEE ICC 2012, Ottawa, Canada, June 2012.
- Rongpeng Li, Zhifeng Zhao, Xuan Zhou, and Honggang Zhang, "Energy Savings Scheme in Radio Access Network via Compressed Sensing Based Traffic Load Prediction," Transactions on Emerging Telecommunications Technologies (ETT), Nov. 2012.

Energy Saving Scheme through BS Switching Operation

- Towards traffic load-aware BSs adaptation.
  - Turn some BSs into sleeping mode on the basis of minimizing the power consumption when the traffic loads are low.
  - Zooming other BS in a coordinated manner.
- To reliably predict the traffic loads is still quite challen
- One BS tly relate the traf
- Compa chemes, foresigh heavily r
- Actual ient lear





Machine Learning

Supervised Learning

- Data
- Desired Signal/Teacher

Reinforcement Learning

- Data
- Rewards/Punishments

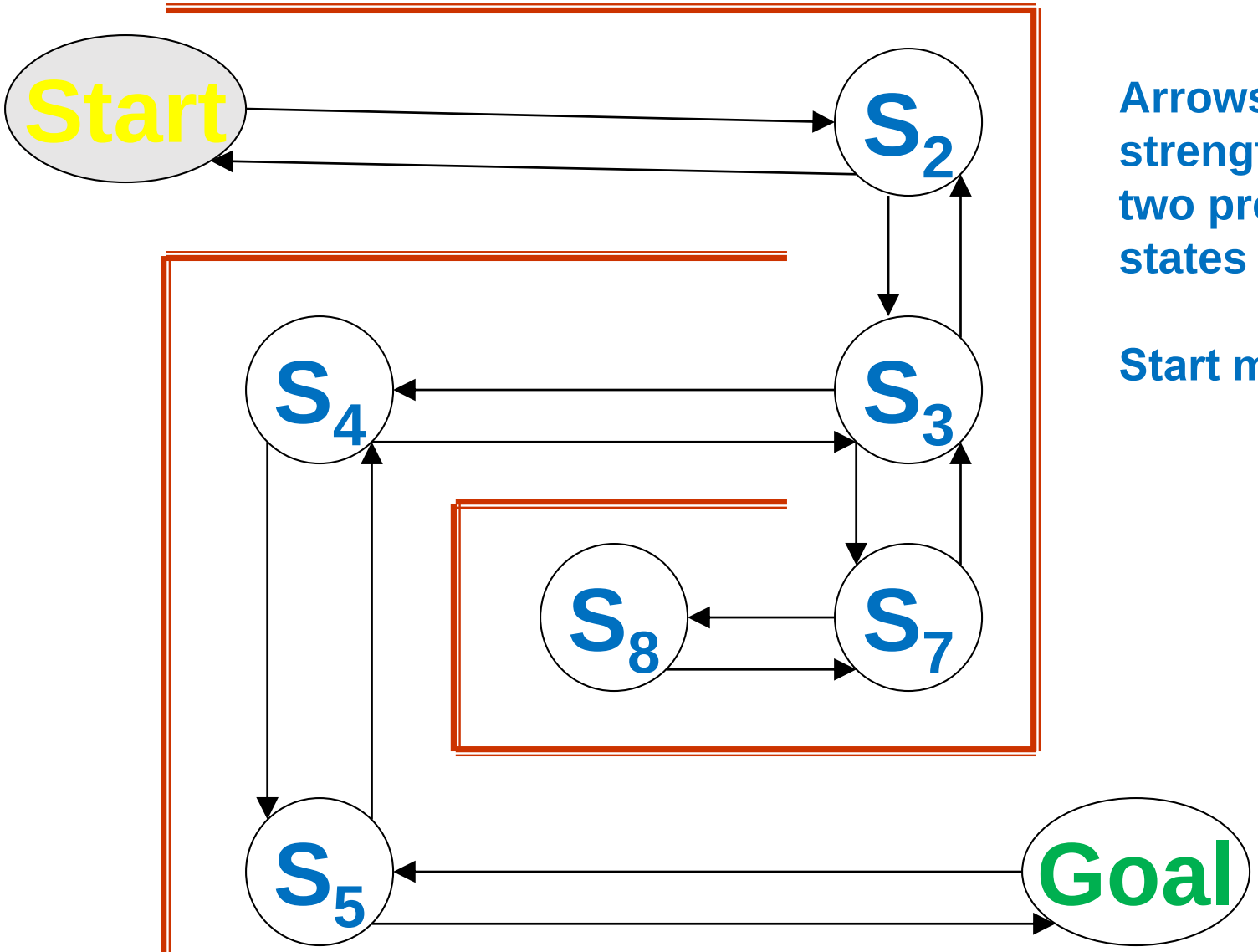
Unsupervised Learning

- Only the Data

Reinforcement Learning

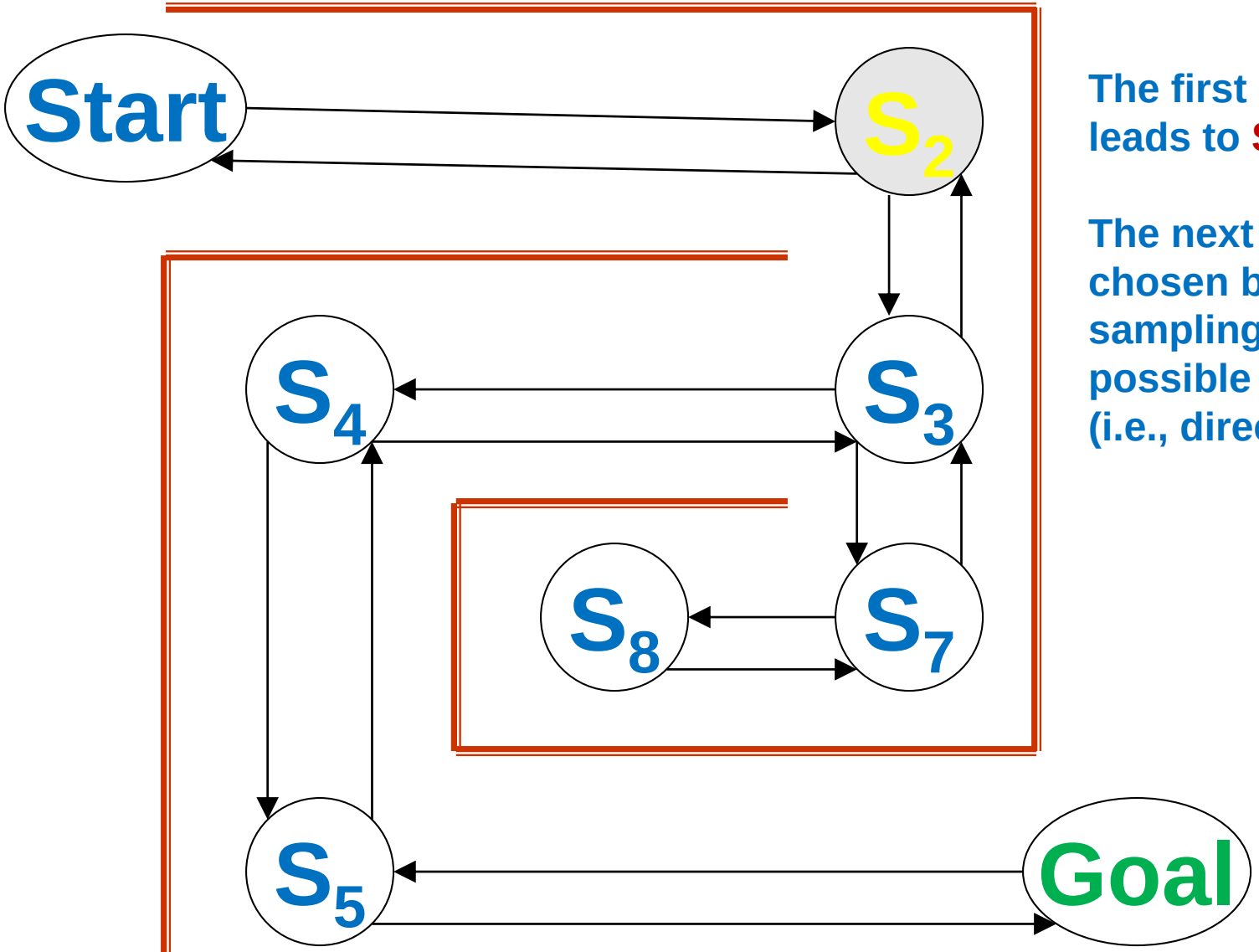
- Reinforcement learning (RL) is learning by interacting with an environment. An RL agent **learns from the consequences of its actions** to maximize the accumulated reward over time, **rather than from being explicitly taught** and it selects its actions on basis of its past experiences (**exploitation**) and also by new choices (**exploration**), which is essentially trial and error learning.

-- Scholarpedia



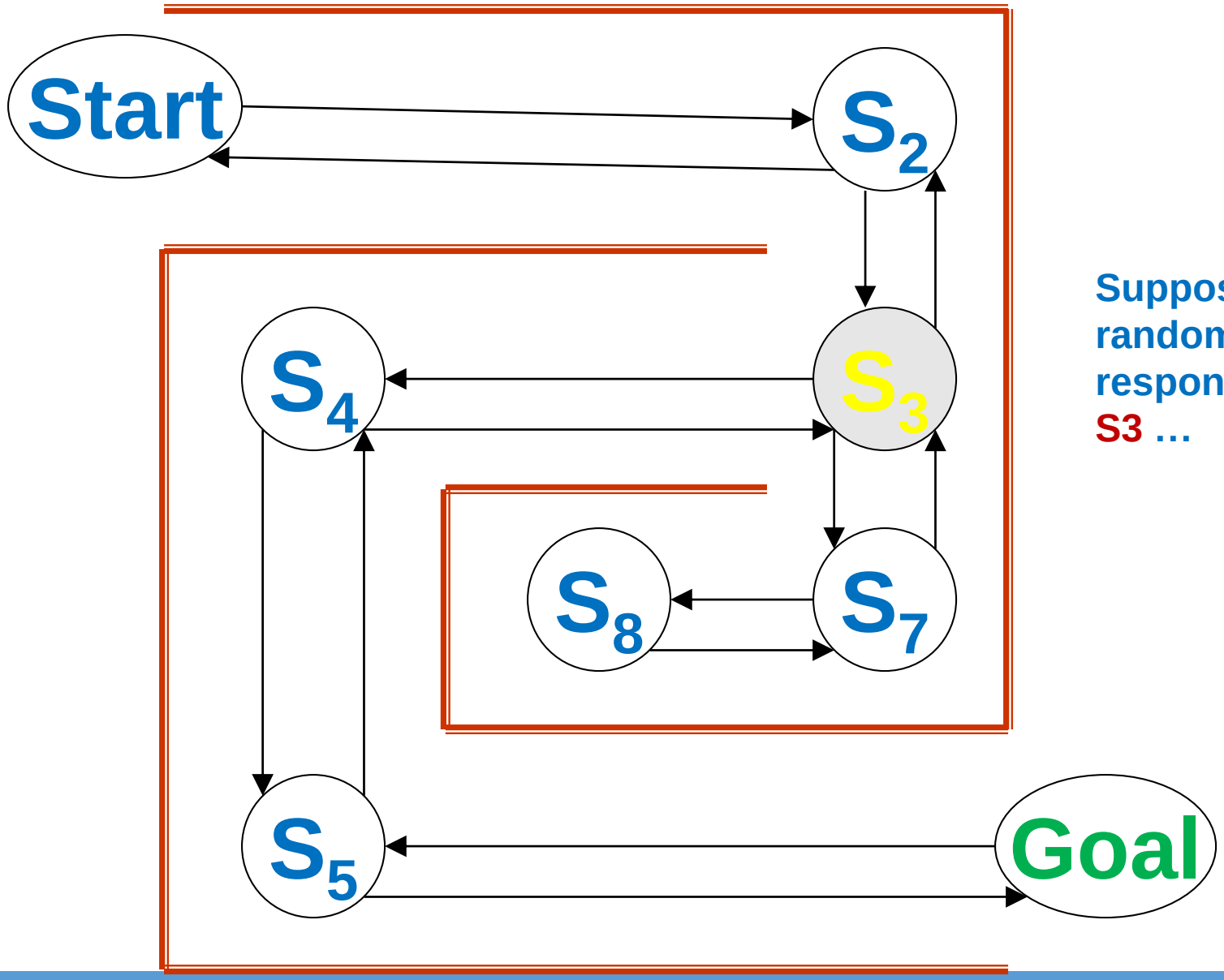
Arrows indicate strength between two problem states

Start maze …

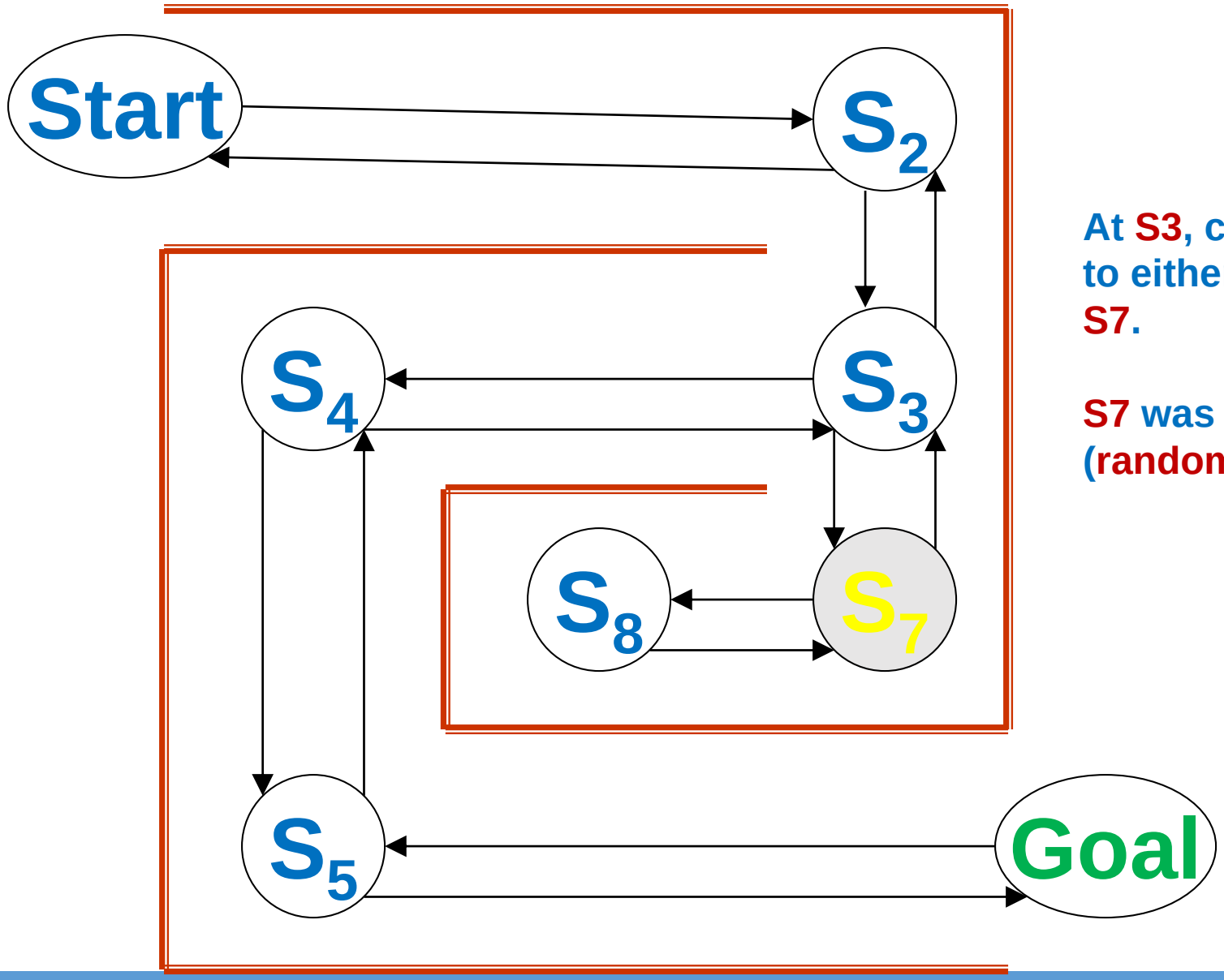


The first response leads to **S2** …

The next state is chosen by ***randomly*** sampling from the possible next states (i.e., directions).

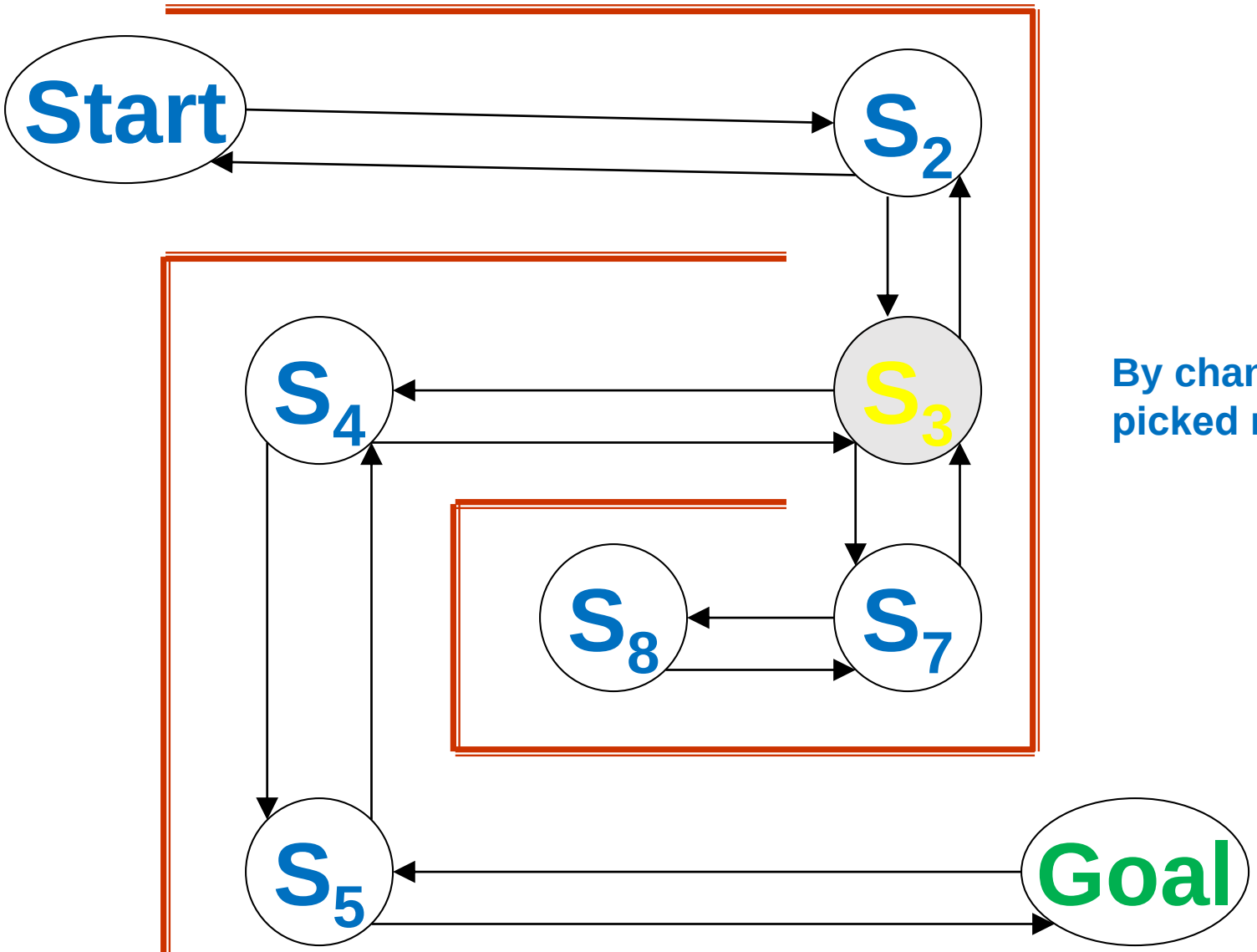


Suppose the randomly sampled response leads to S3 …

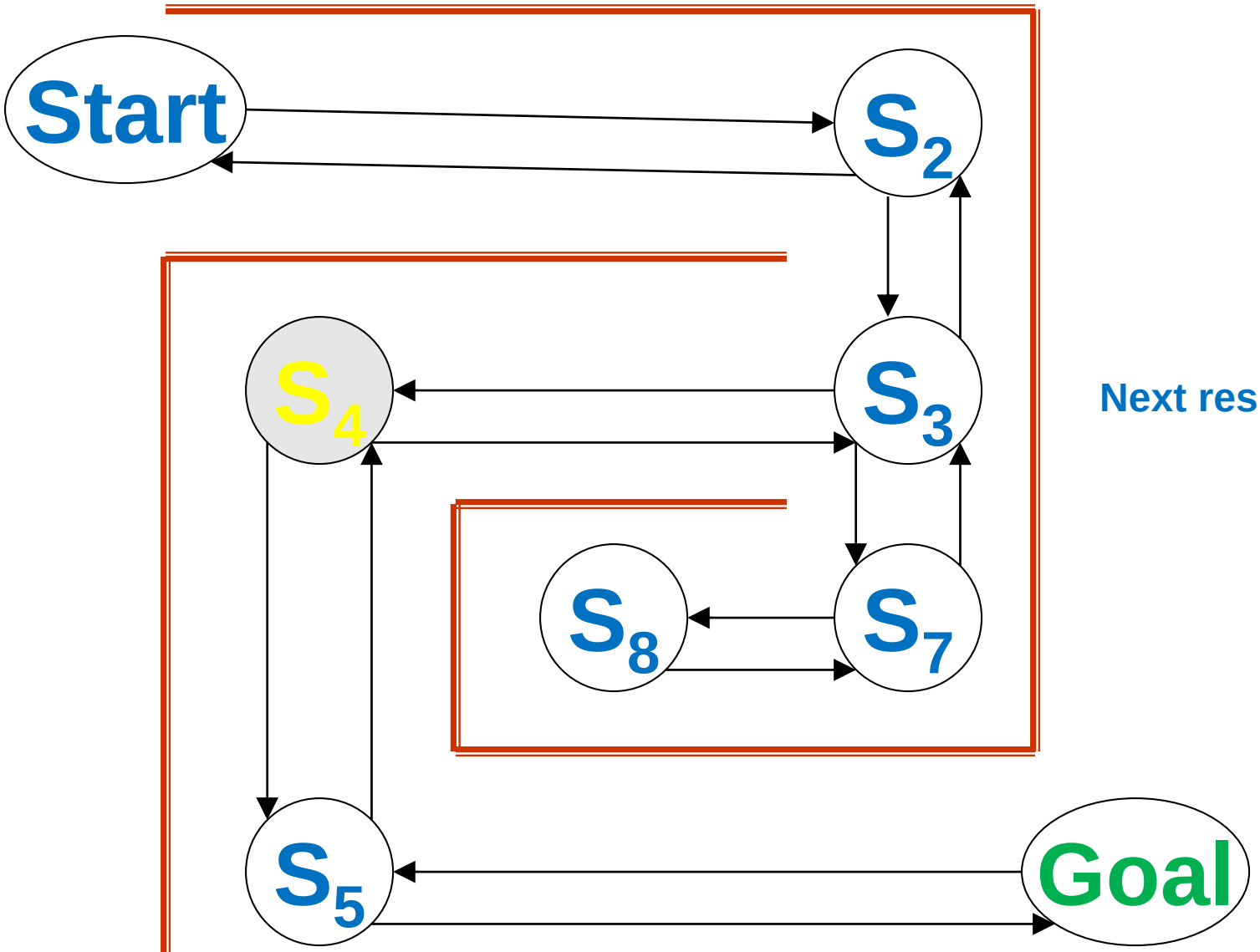


At S3, choices lead to either S2, S4, or S7.

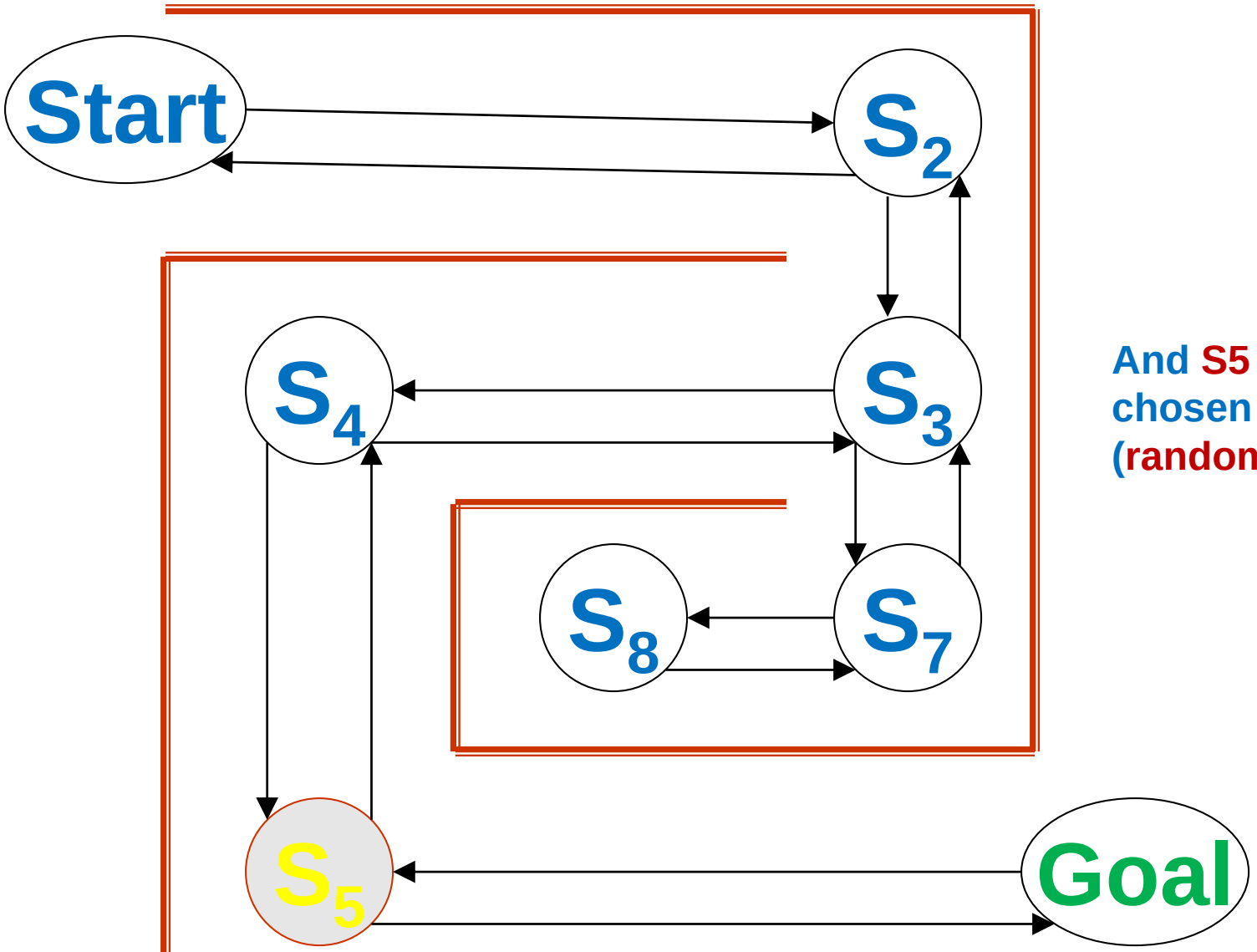
S7 was picked (randomly)



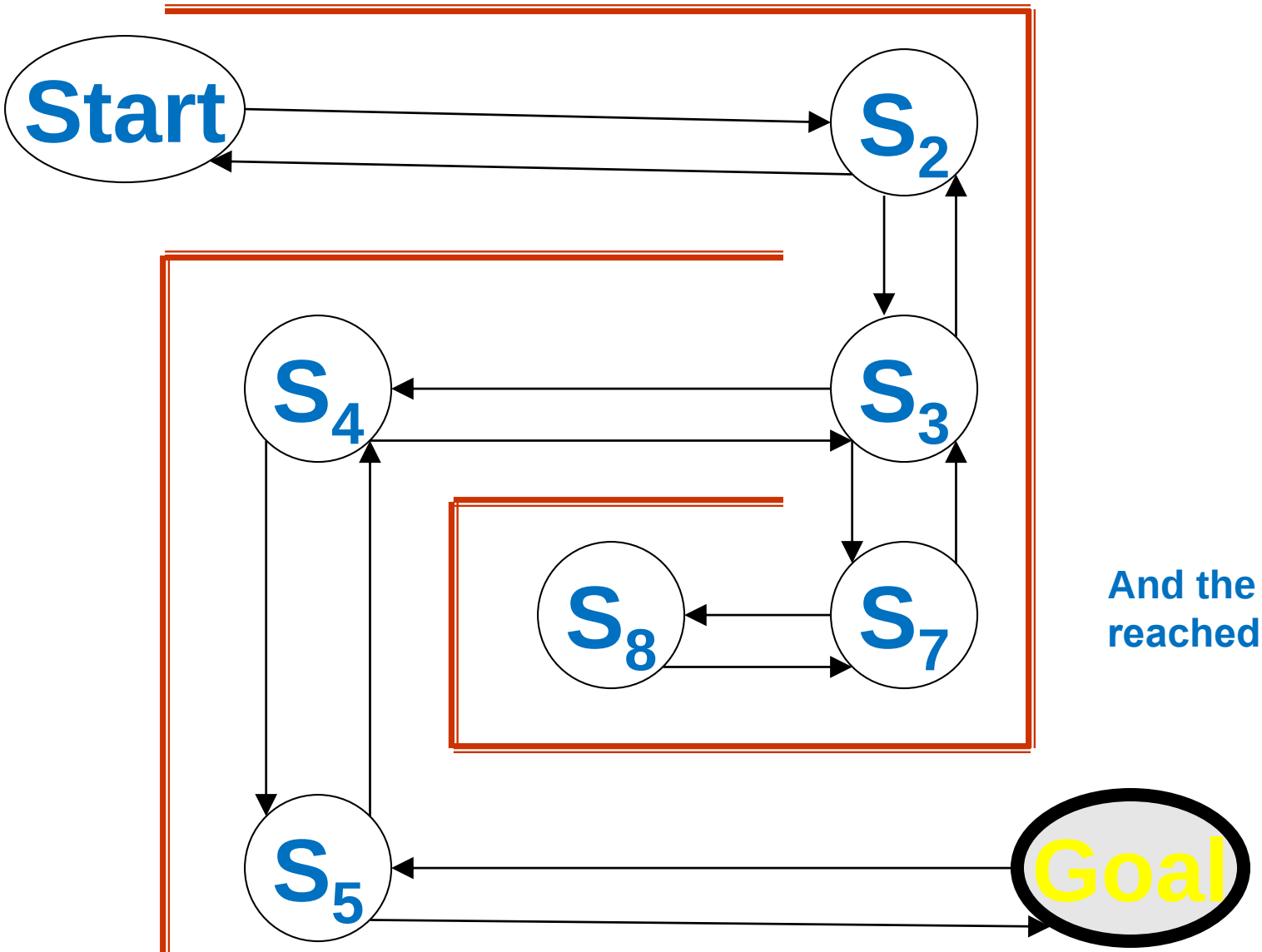
By chance, S3 was picked next…



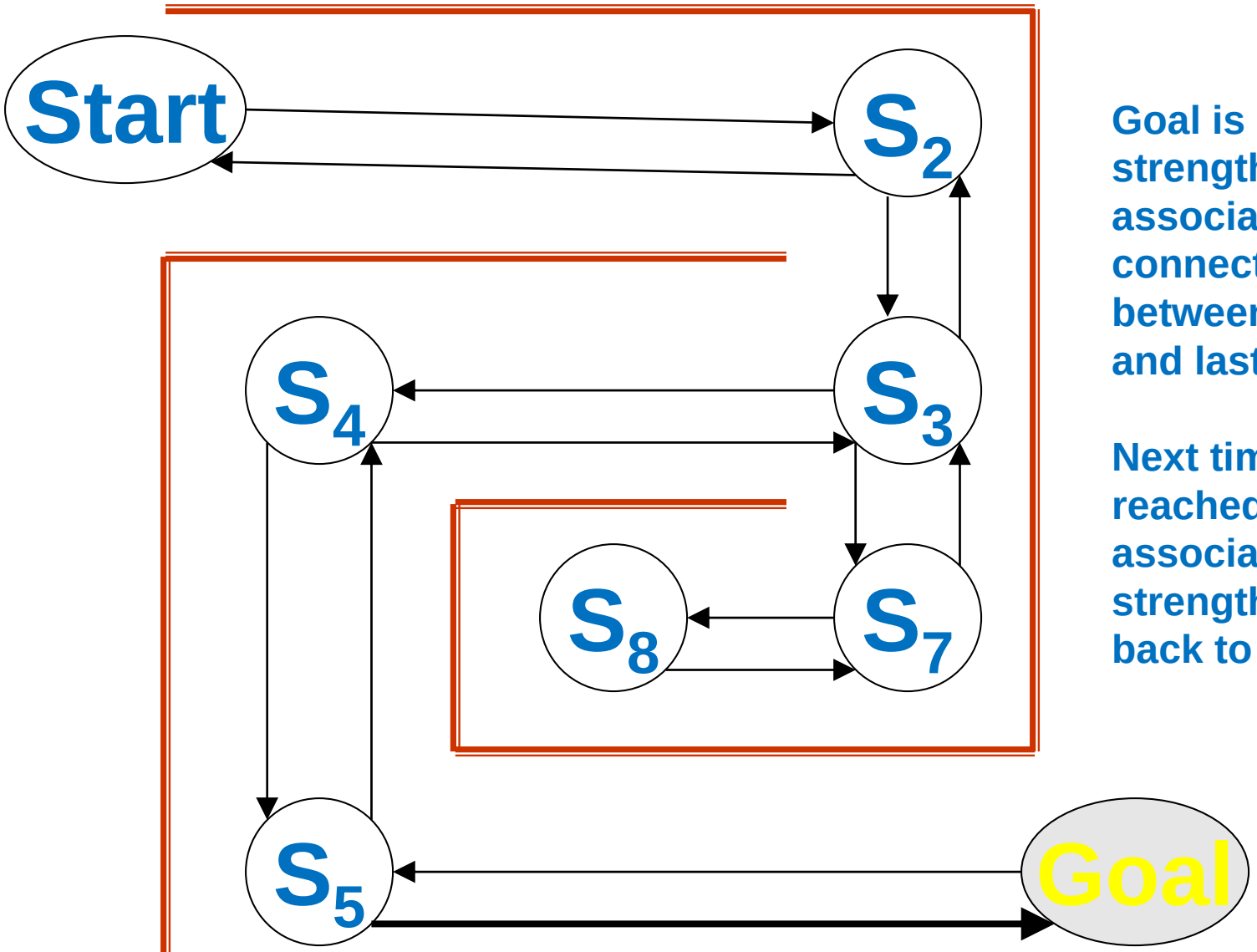
Next response is S4



And S5 was chosen next (randomly)

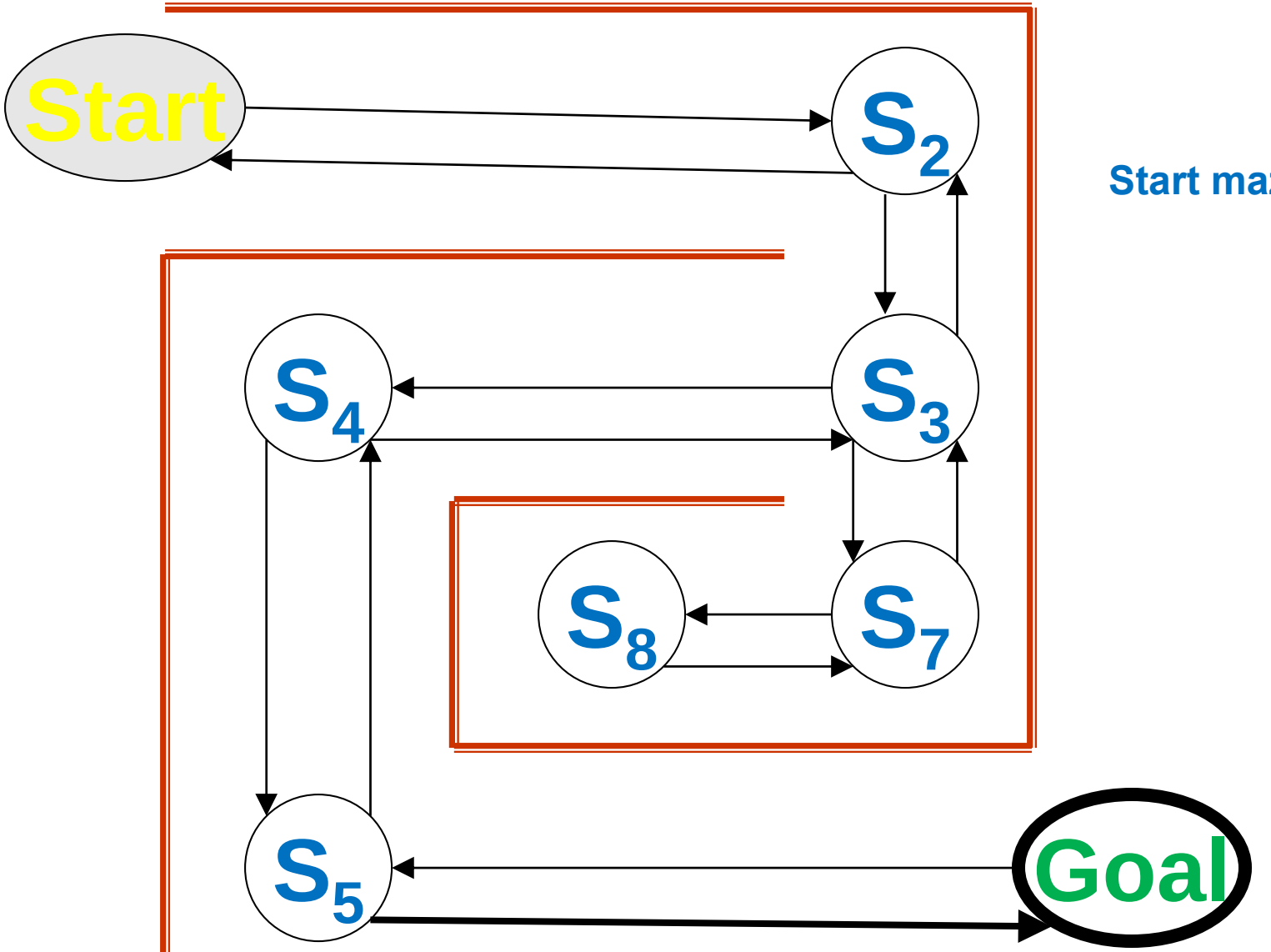


And the goal is reached …

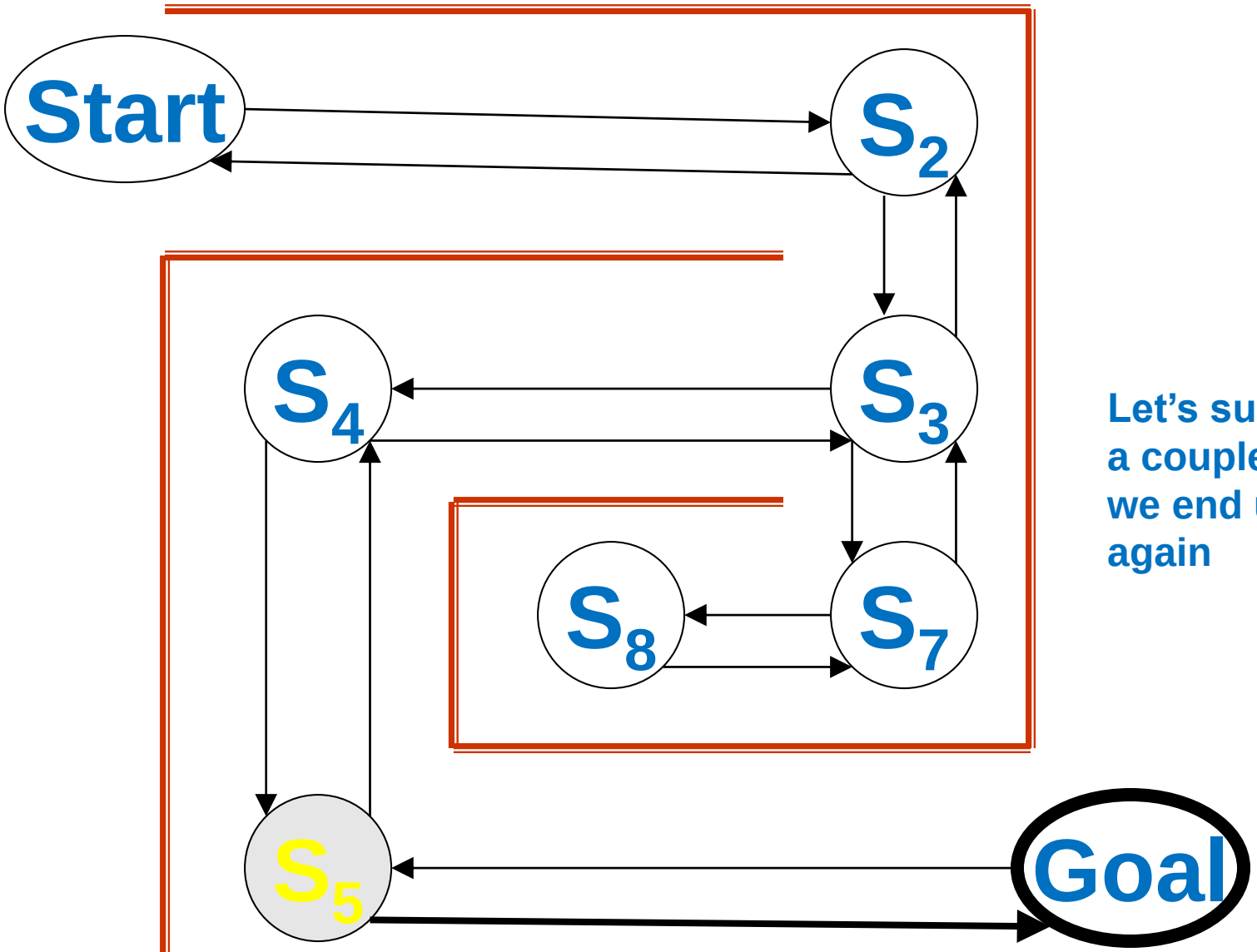


Goal is reached, strengthen the associative connection between goal state and last response

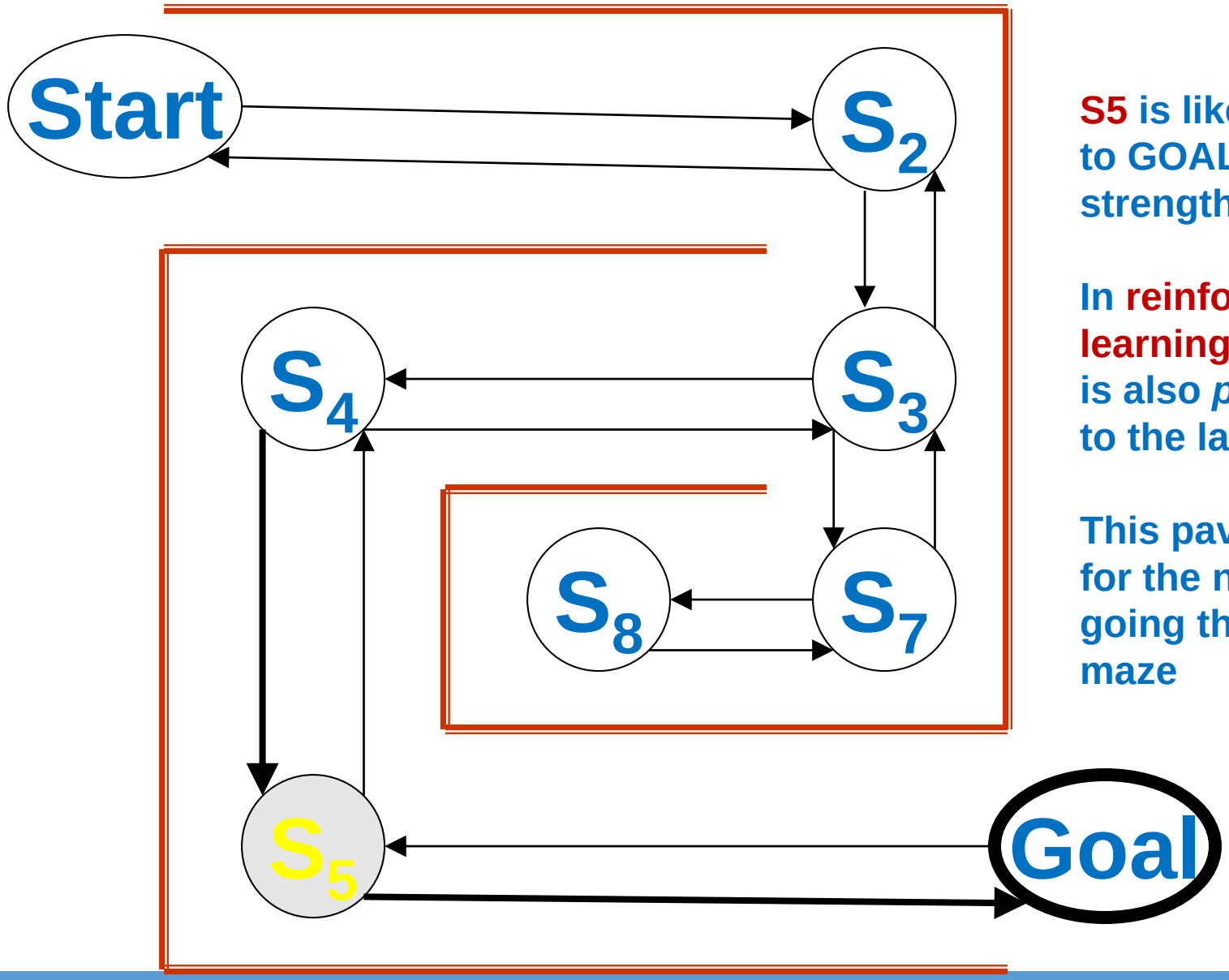
Next time S5 is reached, part of the associative strength is passed back to S4...



Start maze again…



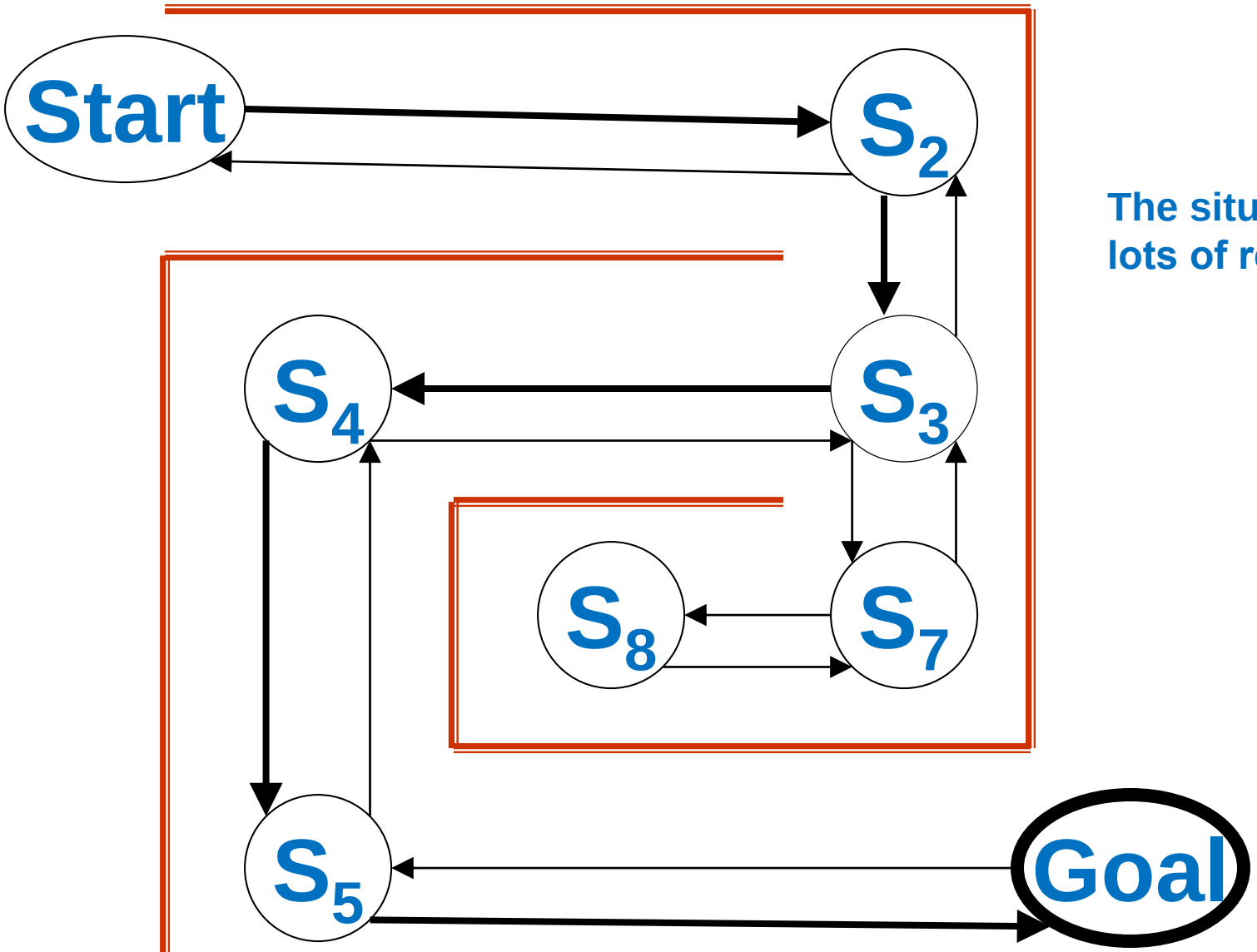
Let's suppose after a couple of moves, we end up at **S5** again



S5 is likely to lead to GOAL through strengthened route

In **reinforcement learning**, strength is also *passed back* to the last state

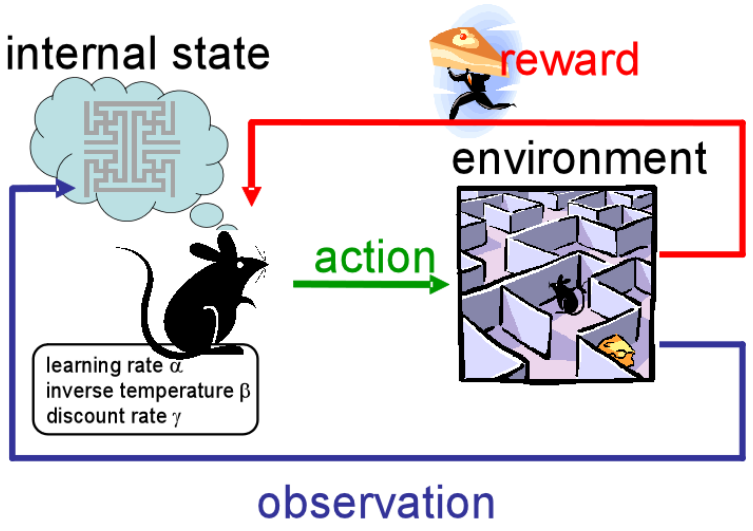
This paves the way for the next time going through maze

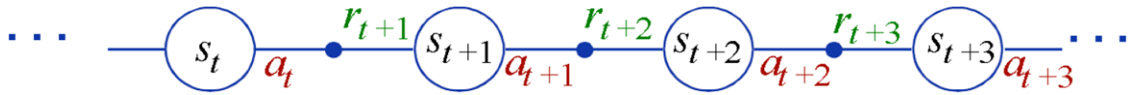


The situation after lots of restarts …

Markov Decision Process (MDP)

- An MDP $M =< \mathcal{S}, \mathcal{A}, P, C >$
  - **State space** $\mathcal{S}$: $\mathbf{s}^{(k)}$
  - **Action space** $\mathcal{A}$: $\boldsymbol{a}^{(k)}$
  - **Transition probability** P
  - **Cost/Reward function** C
- A **strategy** π, from state $\mathbf{s}^{(k)}$ to an action $\pi(\mathbf{s}) = \boldsymbol{a}^{(k)}$ to minimize the value function starting from the state $\mathbf{s}$





Modeling the State-Value Function

- **Infinite Horizon Model**: Discounted Accumulative Cost
  - $V^{\pi}(s) = E_{\pi}\left[\sum_{k=0}^{\infty} \gamma^{k} C(s^{(k)}, \pi(s^{(k)}) | s^{(0)} = s)\right] = E_{\pi}\left[C(s, \pi(s)) + \right.$

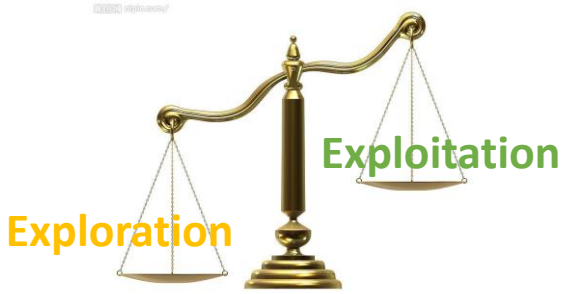
Bellman Equation and Optimal Strategy: The Methodology

- Bellman equation and optimal strategy π^*

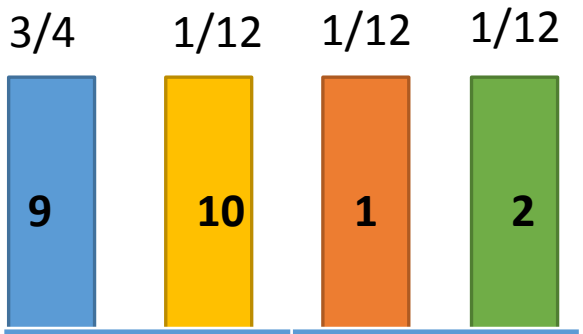
$$V^*(s) = V^{\pi^*}(s) = \min_{a \in \mathcal{A}} \{ E_{\pi^*} [C(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V^{\pi^*}(s')] \}$$

- Two important sub-problems to find the optimal strategy and the value function
  - **Action Selection**
  - **Value Function Approximation**

Action Selection



- Action Selections is actually a tradeoff between exploration and exploitation.
  - Exploration: Increase the agent's knowledge base;
  - exploitation: Leverage existing but under-utilized knowledge base
- Assume that the agent has N actions to select
  - Greedy
  - ϵ-Greedy
    - Choose the action with the largest reward with a probability of $1-\epsilon$
    - Choose others with the largest reward with a probability of $\epsilon/(N-1)$
  - Gibbs or Boltzmann Distribution
    - $\frac{e^{p(s^k,a^k)/\tau}}{\sum_{a\in\mathcal{A}} e^{p(s^k,a)/\tau}}$
    - Temperature $\tau \to 0$: Greedy algorithm;
    - $\tau \to \infty$: Uniformly selecting the action



Selected Action	Reward/Cost
1	11
2	9
1	9
1	10
2	10

Action Selection: From the Point of Game

- **Best response**

$$\text{argmax}_{a \in \mathcal{A}} E[C(s,a)]$$

The discontinuities inherent in this maximization present difficulties for adaptive processes.

- **Smooth best response**

$$\text{argmax}_{a \in \mathcal{A}} E[C(s,a)] + \tau\mu(s,a)$$

$\mu(s,a)$ is a smooth, strictly differentiable concave function.

- **If $\boldsymbol{\mu(s,a) = -\sum p(s,a) log p(s,a)}$, we can obtain the Boltzmann distribution**.
  - By Lagrange Multiplier Algorithm, it equals that $\max_{a \in \mathcal{A}} E[C(s,a)]$, subject to $\mu(s,a) = C$.

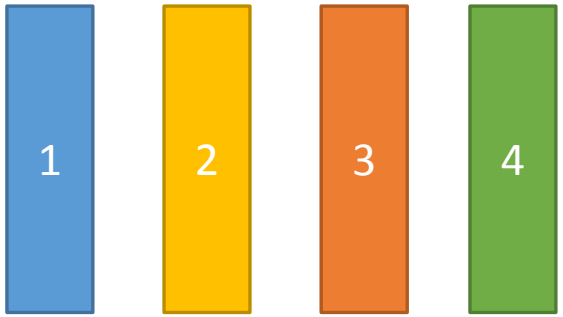
State-Value Function Update/Approximation

- Iterations: The way to obtain a strategy
  - Policy Update
  - State-Value Function Update
- Temporal Difference (TD) Error
- Example:

$$Q_{k+1} = \frac{1}{k+1}\sum_{i=1}^{k+1} r_i = \frac{1}{k+1}(r_{k+1} + \sum_{i=1}^{k} r_i)$$

$$= \frac{1}{k+1}(r_{k+1} + kQ_k) = Q_k + \frac{\mathbf{1}}{\mathbf{k+1}}[\boldsymbol{r_{k+1}} - \boldsymbol{Q_k}]$$

- Newton's Gradient Decent Method

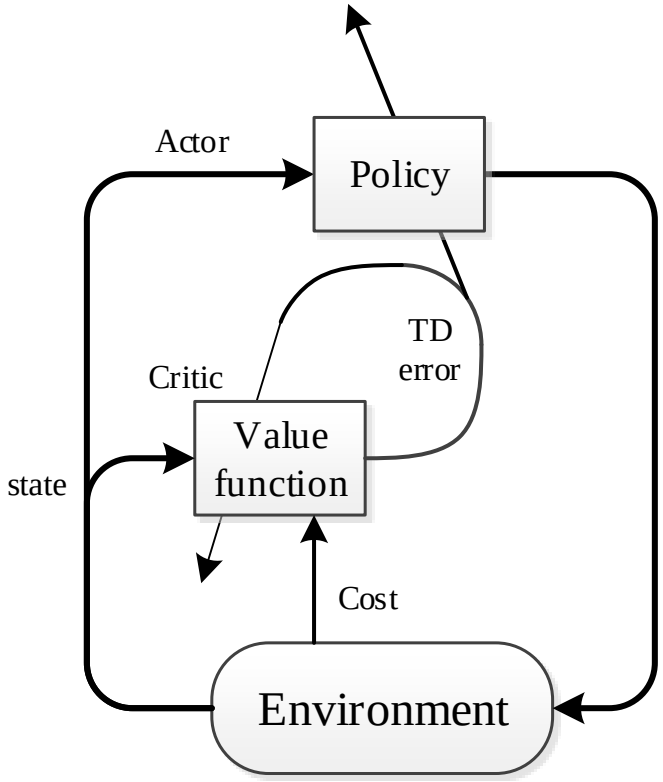


Selected Action	Reward/Cost
1	11
2	9
1	9
1	10
2	10

Actor-Critic Algorithm

- The actor-critic algorithm encompasses three components: actor, critic, and environment.
- **Actor**: According to Boltzmann Distribution, select an action in a stochastic way and then executes it.
- **Critic**: Criticizes the action executed by the actor and updates the value function through TD error. (TD(0) and TD(λ))

$$\delta^{(k)}(s^{(k)}, a^{(k)}) = C^{(k)}(s^{(k)}, a^{(k)}) + \gamma \cdot V^{(k)}(s^{(k+1)})$$



A Comparison among the Typical RL Algorithms

Nam

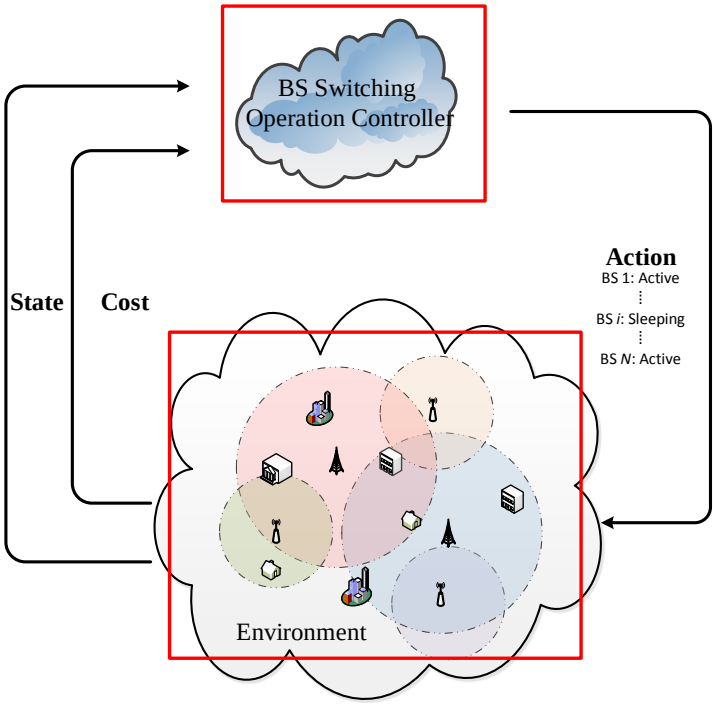
An Actor/Critic Algorithm that is Equivalent to Q-Learning

Robert H. Crites
Computer Science Department
University of Massachusetts
Amherst, MA 01003
crites@cs.umass.edu

Andrew G. Barto
Computer Science Department
University of Massachusetts
Amherst, MA 01003
barto@cs.umass.edu

RL Architecture for Energy Saving in Cellular Networks

- Environment: a region $\mathcal{L} \in \mathbb{R}^2$ served by a set of BSs $\mathcal{B} = \{1, \dots, N\}$
- Controller: a BS switching operation controller to turn on/off some BSs in a centralized way;
- A traffic load density as $\gamma(x) = \frac{\lambda(x)}{\mu(x)} < \infty$: arrival rate per unit area $\lambda(x)$ and file size $\frac{1}{\mu(x)}$.
- Traffic load within BS i's coverage: $\Gamma_i = \int_{\mathcal{L}} \gamma(x) I_i(x, \mathcal{B}_{on})\, dx$
  - $I_i(x, \mathcal{B}_{on}) = 1$ denotes location x
  - is served by BS$i \in B_{on}$, vice versa.



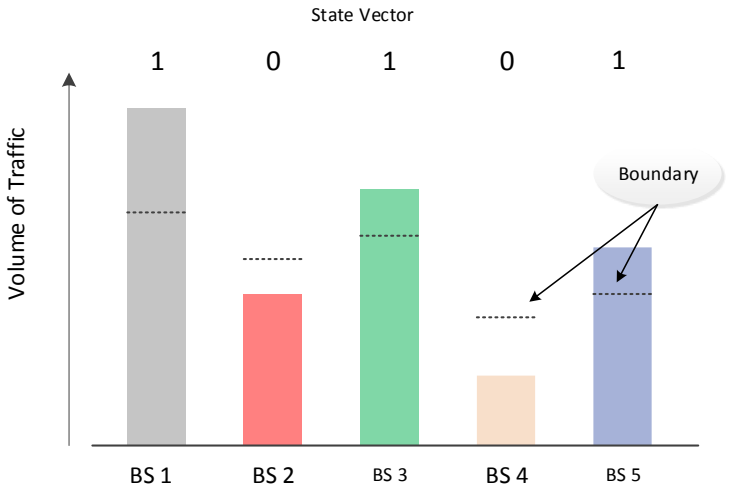
- Rongpeng Li, Zhifeng Zhao, Xianfu Chen, Jacques Palicot, and Honggang Zhang, "TACT: A Transfer Actor-Critic Learning Framework for Energy Saving in Cellular Radio Access Networks," summited to IEEE Transactions on Wireless Communications (Second Round Review).
- Rongpeng Li, Zhifeng Zhao, Xianfu Chen, and Honggang Zhang, "Energy Saving through a Learning Framework in Greener Cellular Radio Access Networks," in Proceedings of IEEE Globecom 2012, Anaheim, California, USA, Dec. 2012.

Power Consumption Model and Problem Formulation

- All active BSs consumed power
- $\psi(\boldsymbol{\rho}, \mathcal{B}_{on}) = \sum_{i \in \mathcal{B}_{on}} \left[(1 - q_i)\rho_i P_i + q_i P_i \right]$
  - $q_i \in [0,1]$: the portion of constant power consumption for BS i;
  - P_i: the maximum power consumption of BS i when it is fully utilized.
  - System load for BS $i \in \mathcal{B}_{on}$: $\rho_i = \int_{\mathcal{L}} \varrho_i(x) I_i(x, \mathcal{B}_{on}) dx$
    - System load density is defined as the fraction of time required to deliver traffic load $\gamma(x)$ from BS $i \in \mathcal{B}_{on}$ to location x, namely $\varrho_i(x) = \gamma(x)/c_i(x, \mathcal{B}_{on})$.
- The delay optimal performance function
  - $\omega(\boldsymbol{\rho}, \mathcal{B}_{on}) = \sum_{i \in \mathcal{B}} \frac{\rho_i}{1-\rho_i}$
- Objection function $\min_{\mathcal{B}_{on}, \boldsymbol{\rho}} \{\psi(\boldsymbol{\rho}, \mathcal{B}_{on}) + \varsigma\omega(\boldsymbol{\rho}, \mathcal{B}_{on})\}$
  - Subject to $\rho_i \in [0,1) \forall i \in \mathcal{B}$

BS Traffic Load State Vector

- Finite state Markov process (FSMC) to demonstrate the traffic load variation condition;
- Traffic load Γ_i for BS i is partitioned into several parts by a boundary point Γ_b;



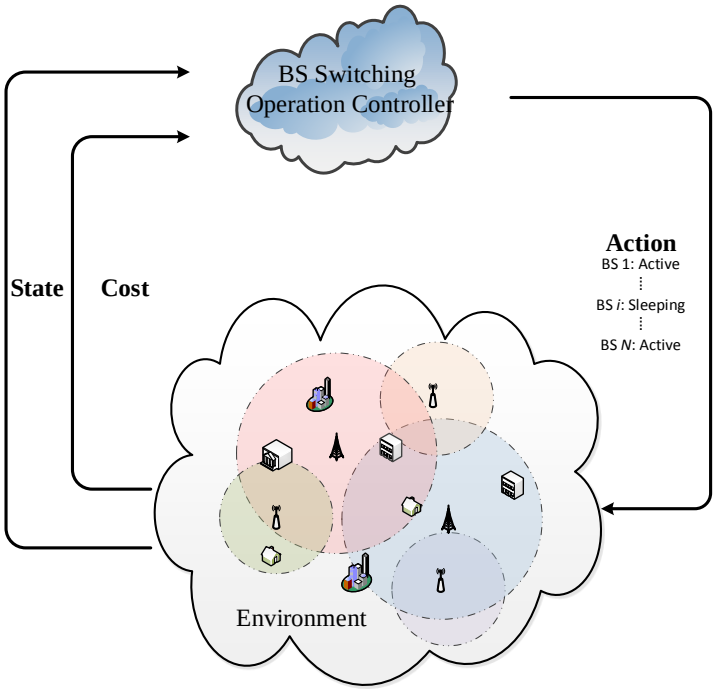
Bellman Equation

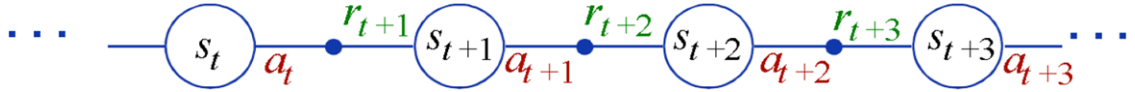
- Accumulative cost

$$\boldsymbol{V^{\pi}(\mathbf{s})} = \sum_{\boldsymbol{k=0}}^{\infty} \boldsymbol{\gamma^{k}}\, \boldsymbol{C^{(k)}}\left(\mathbf{s}^{(\boldsymbol{k})}, \boldsymbol{\pi}(\mathbf{s}^{(\boldsymbol{k})}) | \mathbf{s}^{\mathbf{0}} = \mathbf{s}\right)$$
$$= \boldsymbol{C}(\mathbf{s}, \boldsymbol{\pi}(\mathbf{s})) + \boldsymbol{\gamma} \sum_{\mathbf{s}' \in \mathbb{S}} \boldsymbol{p}\,(\mathbf{s}' | \mathbf{s}, \boldsymbol{\pi}(\mathbf{s})) \boldsymbol{V^{\pi}}(\mathbf{s}')$$

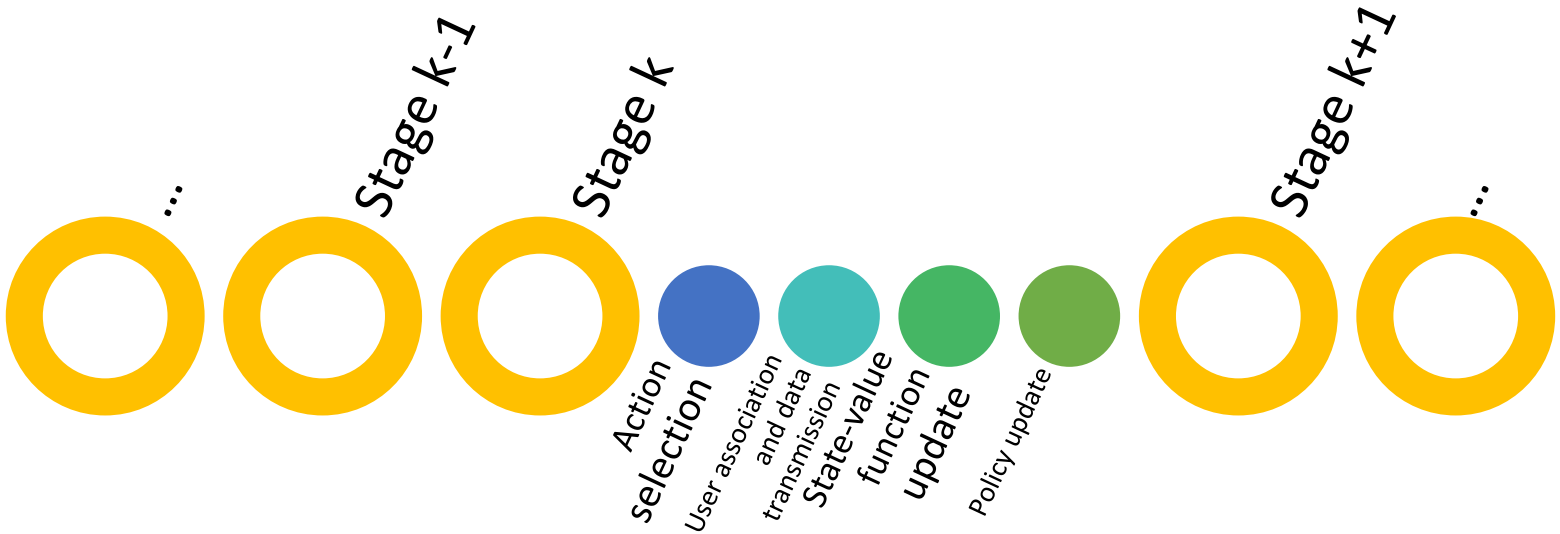
- Bellman equation and optimal strategy π^*

$$V^*(\mathbf{s}) = V^{\pi^*}(\mathbf{s})$$
$$= \min_{\mathbf{a} \in \mathbb{A}} \left\{ C(\mathbf{s}, \mathbf{a}) + \gamma \sum_{\mathbf{s}' \in \mathbb{S}} p\,(\mathbf{s}' | \mathbf{s}, \mathbf{a}) V^{\pi^*}(\mathbf{s}') \right\}$$





Learning Framework based Energy Saving Scheme Detail



Learning Framework based Energy Saving Scheme Detail

Learning Framework based Energy Saving Scheme Detail

Take the assumption that the system is at the beginning of stage k, while the traffic load state is $\mathbf{s}^{(k)}$.

Learning Framework based Energy Saving Scheme Detail

Action selection: the controller the controller selects an action $\mathbf{a}^{(k)}$ in state $\mathbf{s}^{(k)}$ with the probability (Boltzmann distribution)

$$\pi^{(k)}(\mathbf{s}^{(k)}, \mathbf{a}^{(k)}) = \frac{\exp\{p(\mathbf{s}^{(k)}, \mathbf{a}^{(k)})\}}{\sum_{\mathbf{a}^{(k)} \in \mathbb{A}} \exp\{p(\mathbf{s}^{(k)}, \mathbf{a}^{(k)})\}}$$

After that, the corresponding BSs turns into sleeping mode.

Learning Framework based Energy Saving Scheme Detail

User association and data transmission: the users at location x choose to connect one BS i according to the following equation and start the data communication slot by slot.

$$i^*(x) = \text{argmax}_{j \in \mathcal{B}on} \frac{c_j(x, \mathcal{B}_{on})}{\left(1 - q_j)P_j + \varsigma\left(1 - \rho_j^{k-1}\right)^{-2}}$$

Learning Framework based Energy Saving Scheme Detail

State-value function update: after the transmission part of stage k, the traffic loads in each BS will change, and system will transform to state $\mathbf{s}^{(k+1)}$.

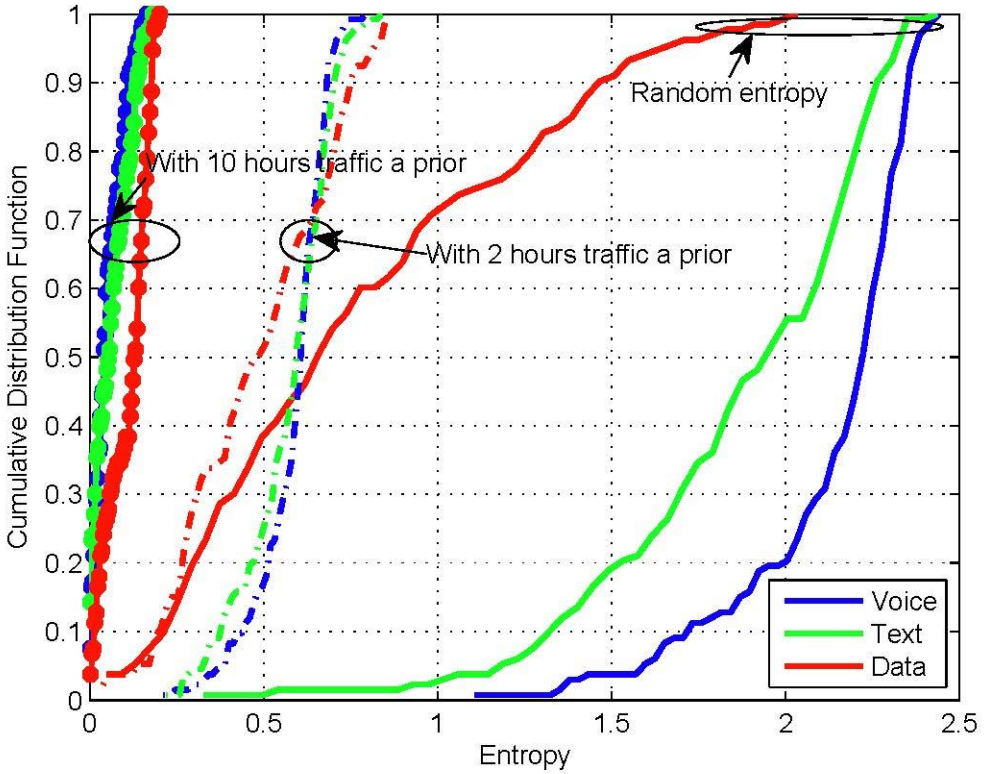
A temporal difference error $\delta(\mathbf{s}^{(k)}) = C^{(k)}(\mathbf{s}, \mathbf{a}) + \gamma \cdot V(\mathbf{s}^{(k+1)}) -$

Learning Framework based Energy Saving Scheme Detail

Policy update: At the end of stage k, "criticize" the selected action by $p(\mathbf{s}^{(k)}, \mathbf{a}^{(k)}) \leftarrow p(\mathbf{s}^{(k)}, \mathbf{a}^{(k)}) - \beta \cdot \delta(\mathbf{s}^{(k)})$.

Remark: one action under a specific state can be selected with higher probability if the ``foresighted'' cost it takes is comparatively smaller.

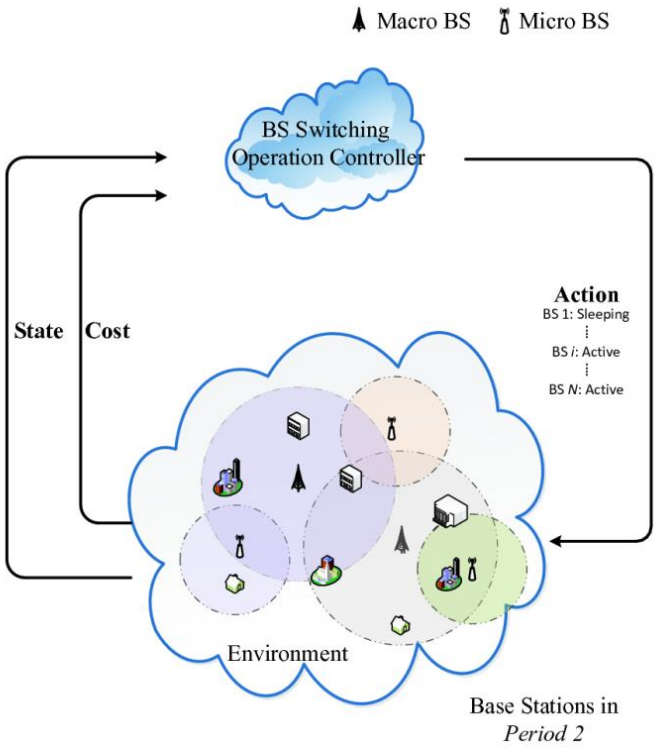
A Rethink of the Traffic Characteristics



- Rongpeng Li, Zhifeng Zhao, Xuan Zhou, Jacques Palicot, and Honggang Zhang, "The Prediction Analysis of Cellular Radio Access Networks Traffic: From Entropy Theory To Network Practicing," summited to IEEE Communications Magazine (Second Round Review).

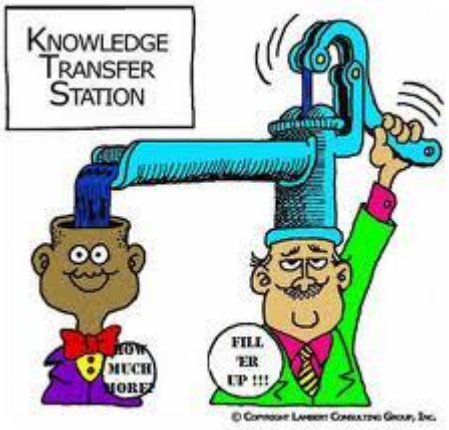
Transfer Actor-Critic Algorithm: Motivation

The concept of transfer learning

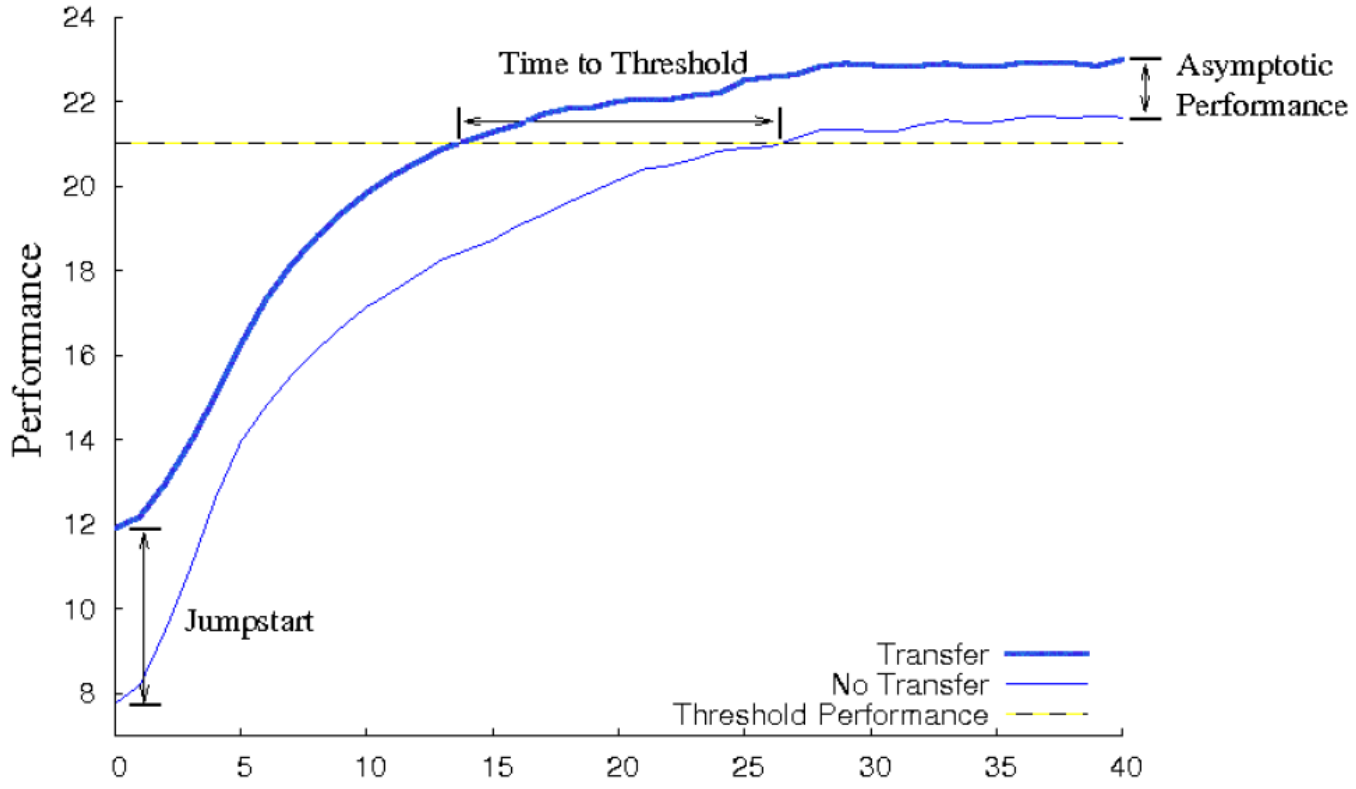


- Rongpeng Li, Zhifeng Zhao, Xianfu Chen, Jacques Palicot, and Honggang Zhang, "TACT: A Transfer Actor-Critic Learning Framework for Energy Saving in Cellular Radio Access Networks," summited to IEEE Transactions on Wireless Communications (Second Round Review).

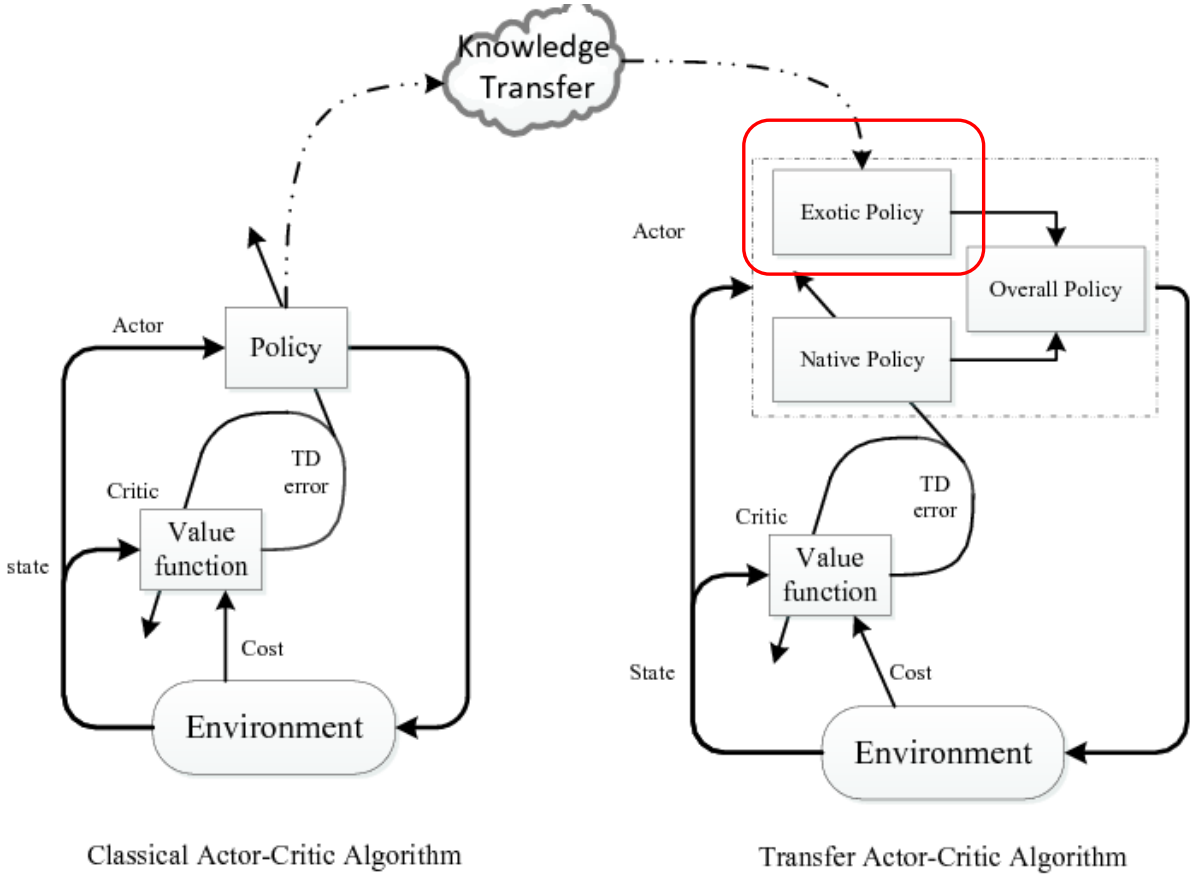
Examples of Transfer Learning



Advantages of Transfer Learning



Transfer actor-critic algorithm: methodology



Transfer Policy Update

$$p_o^{(k+1)}(\mathbf{s}^{(k)},\mathbf{a}^{(k)})$$

$$=\Big[(1-\zeta(\nu_2(\mathbf{s}^{(k)},\mathbf{a}^{(k)},k)))\,p_n^{(k+1)}(\mathbf{s}^{(k)},\mathbf{a}^{(k)})+\zeta(\nu_2(\mathbf{s}^{(k)},\mathbf{a}^{(k)},k))\,p_e(\mathbf{s}^{(k)},\mathbf{a}^{(k)})\Big]^{p}$$

Transfer Rate

Native Policy

Exotic Policy

- **Transfer rate** $\nu_2(\mathbf{s}^{(k)},\mathbf{a}^{(k)},k)$
  - Incrementally decreases as the iterations run.
  - Diminishes the impact of exotic policy once the controller masters certain amount of information.

Initialization:
for each $\mathbf{s} \in \mathbb{S}$, each $\mathbf{a} \in \mathbb{A}$ **do**
Initialize state-value function $V(\mathbf{s})$, native policy function $p_n(\mathbf{s}, \mathbf{a})$, exotic policy function $p_e(\mathbf{s}, \mathbf{a})$, and strategy function $\pi(\mathbf{s}, \mathbf{a})$;
end for
Repeat until convergent

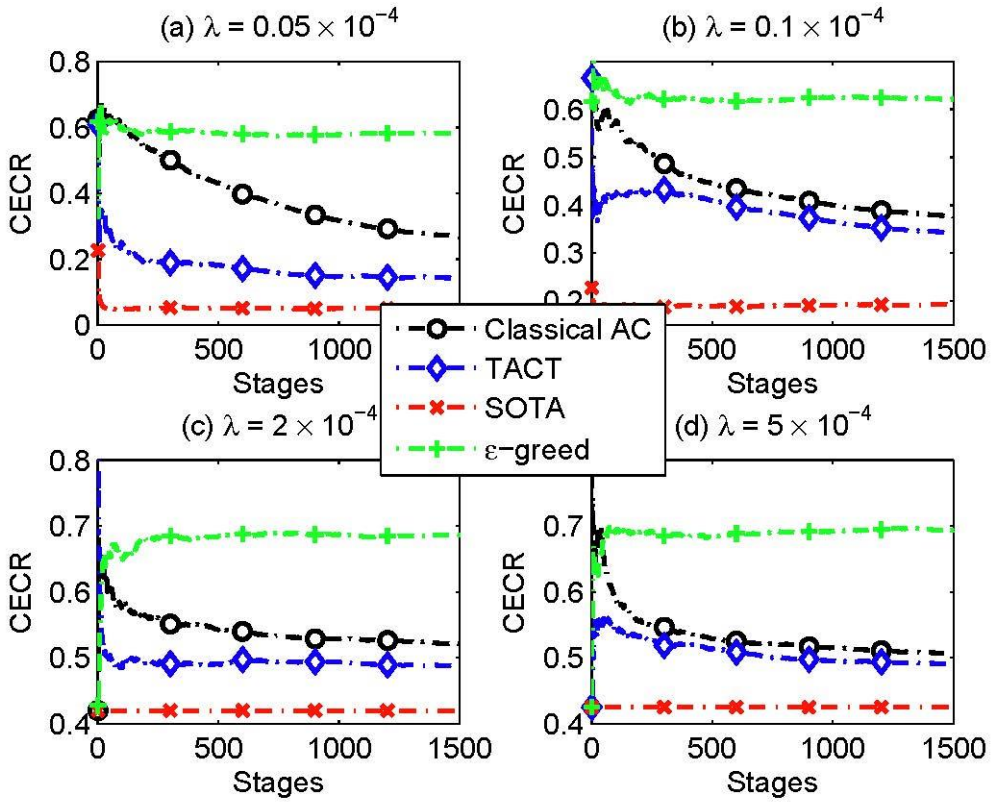
1. Choose an action $\mathbf{a}^k$ in state $\mathbf{s}^k$ according to $\pi(\mathbf{s}^k, \mathbf{a}^k)$;
2. Users at location x connect one BS i by $i^*(x) = \arg\max_{j \in \mathcal{B}_{on}} \frac{c_j(x, \mathcal{B}_{on})}{(1-q_j)p_j + \varsigma(1-\rho_j^k)^{-2}}$, $\forall x \in \mathcal{L}$ and then start data transmission;
3. If $\rho_i \leq 1, \forall i \in \mathcal{L}$, the chosen action is feasible. The cost function $C(\mathbf{s}^k, \mathbf{a}^k)$ is calculated by $\psi(\rho, \mathcal{B}_{on}) + \varsigma\omega(\rho, \mathcal{B}_{on})$; otherwise, an emergent response paradigm starts by quickly turning on some BSs in the hotspot.
4. Identify the traffic loads and accordingly update state $\mathbf{s}^k \to \mathbf{s}^{k+1}$ and compute the TD error by $\delta^k(\mathbf{s}^k) = C^k(\mathbf{s}^k, \mathbf{a}^k) + \gamma \cdot V^k(\mathbf{s}^{k+1}) - V^k(\mathbf{s}^k)$;
5. Update the state-value function $V(\mathbf{s}^k)$ by $V^{k+1}(\mathbf{s}^k) = V^k(\mathbf{s}^k) + \alpha(\nu_1(\mathbf{s}^k, k)) \cdot \delta^k(\mathbf{s}^k)$;
6. Update the native tendency function $p_n(\mathbf{s}^k, \mathbf{a}^k)$ by $p_n^{k+1}(\mathbf{s}^k, \mathbf{a}^k) = p_n^k(\mathbf{s}^k, \mathbf{a}^k) - \beta(\nu_2(\mathbf{s}^k, \mathbf{a}^k, k)) \cdot \delta^k(\mathbf{s}^k, \mathbf{a}^k)$, and update the function $p_o(\mathbf{s}^k, \mathbf{a}^k)$ by $p_o^{k+1}(\mathbf{s}^k, \mathbf{a}^k) = \Big[(1 - \zeta(\nu_2(\mathbf{s}^k, \mathbf{a}^k, k)))p_n^{k+1}(\mathbf{s}^k, \mathbf{a}^k) + \zeta(\nu_2(\mathbf{s}^k, \mathbf{a}^k, k))p_e(\mathbf{s}^k, \mathbf{a}^k)\Big]_{-p_t}^{p_t}$;
7. Update the strategy function $\pi^{k+1}(\mathbf{s}^k, \mathbf{a}) = \frac{\exp\{p_o^{k+1}(\mathbf{s}^k, \mathbf{a})/\tau\}}{\sum_{\mathbf{a}' \in \mathbb{A}} \exp\{p_o^{k+1}(\mathbf{s}^k, \mathbf{a}')/\tau\}}$, for all $\mathbf{a} \in \mathbb{A}$.

Proof of Convergence

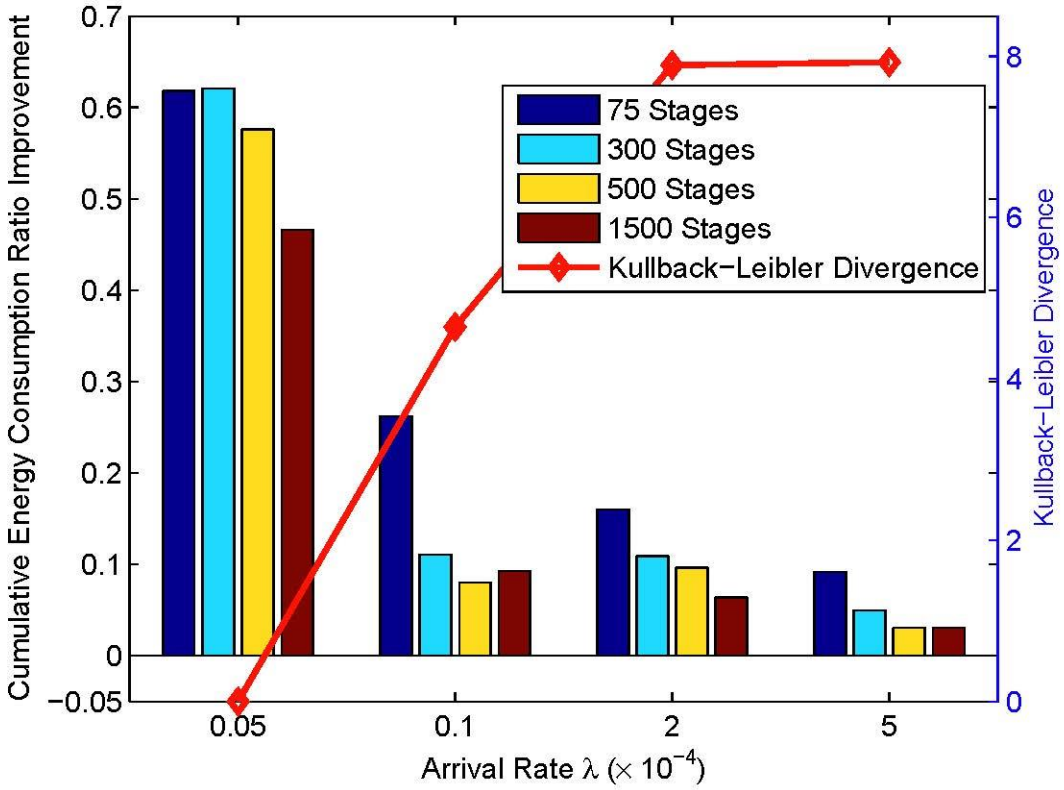
Theorem 4. Regardless of any initial value chosen for $p_{\mathrm{n}}(\boldsymbol{s}, \boldsymbol{a})$, and transferred knowledge $p_{\mathrm{e}}(\boldsymbol{s}, \boldsymbol{a})$, if the learning rate $\alpha(k)$, $\beta(k)$ and the transfer rate $\zeta(k)$ meets the required conditions meanwhile p_{t} and τ are sufficiently large, the Algorithm 1 converges.

Proof. The proof is the direct application of Theorem 2, which establishes the convergence given two conditions. First, the policy $p_{\mathrm{o}}(\boldsymbol{s}, \boldsymbol{a})$ tracks the solution of an ODE. by Theorem 1. Second, the tracked ODE has a strict Lyapunov function, by Theorem 3. Therefore, the learning process in Algorithm 1 converges. ■

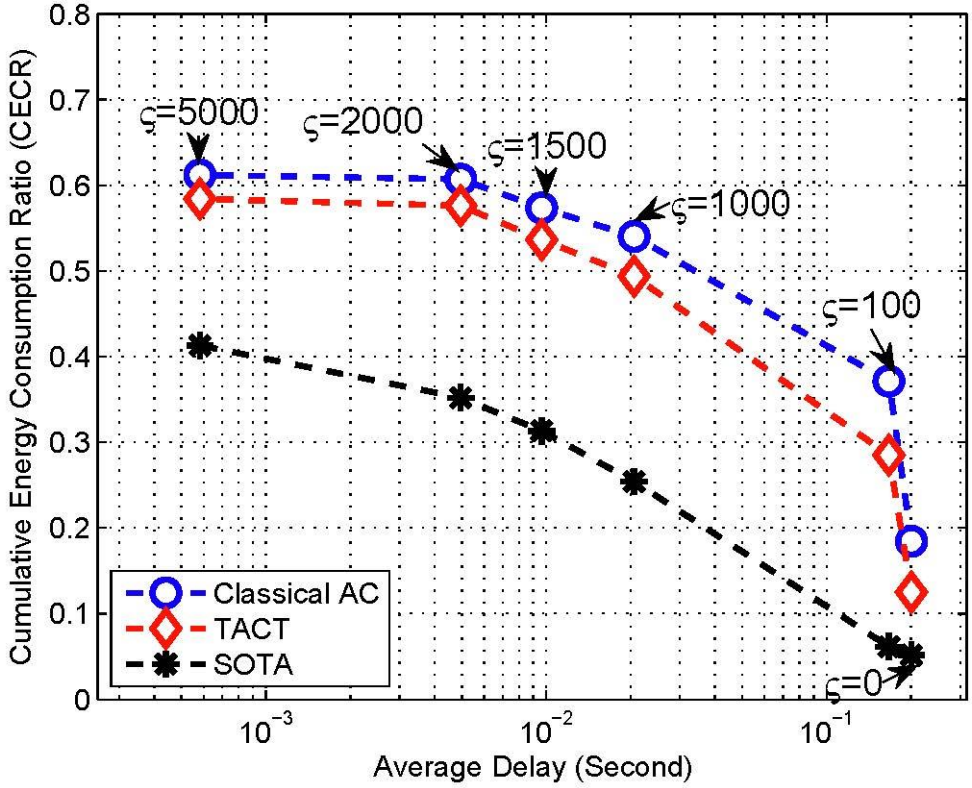
Performance: Different Traffic Arrival Rates



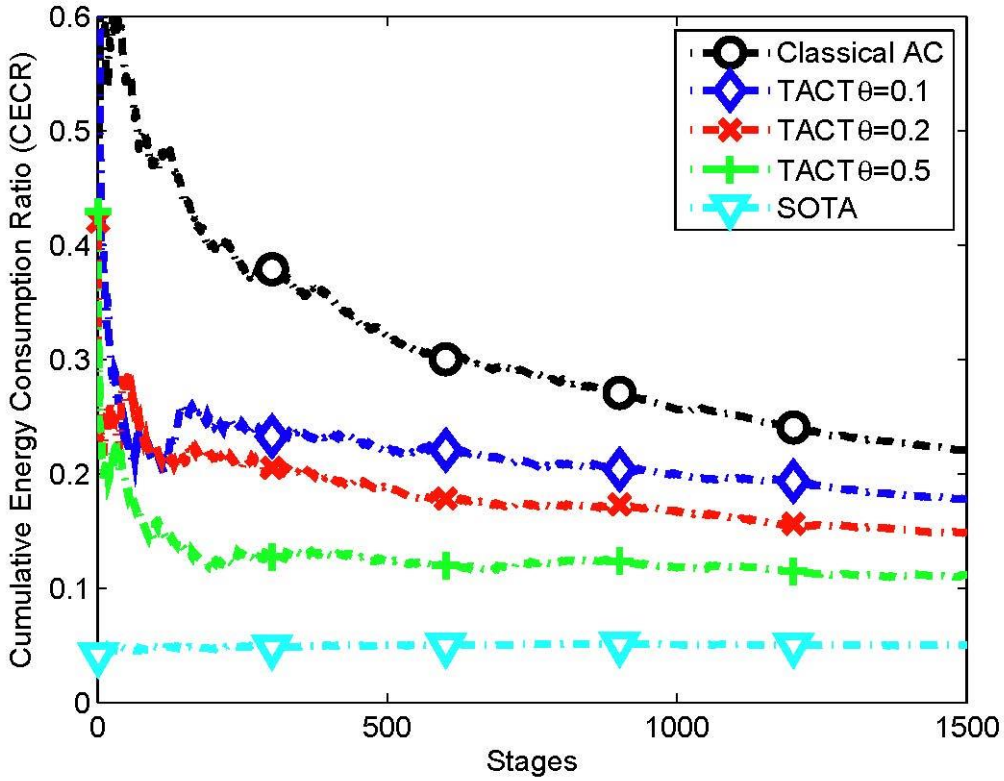
Performance Improvement of TL and KL Divergence



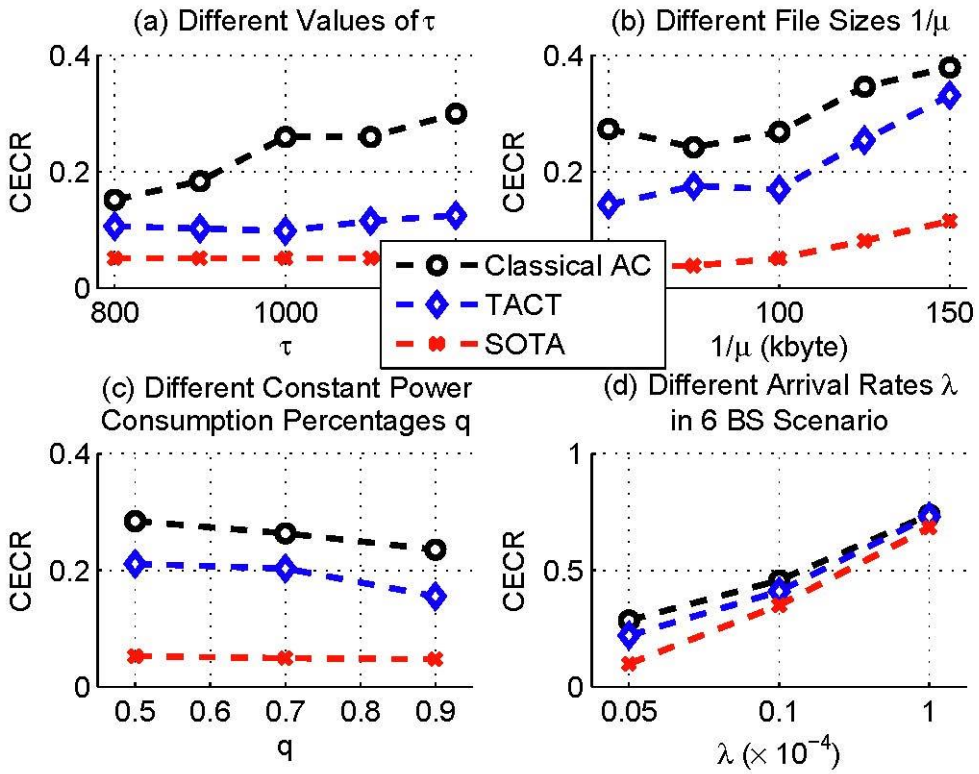
Performance: The Tradeoff between Energy and Delay



Performance: Different Transfer Rates



Performance: Sensitivity Analysis



Q&A

LI Rongpeng

Zhejiang University

Email: lirongpeng@zju.edu.cn

Web: http://www.Rongpeng.info