CNRS - INP - UT3 - UT1 - UT2J

Institut de Recherche en Informatique de Toulouse

Learning@SEPIA : Machine learning pour l'efficacité énergétique

Journée Deep Learning @ IRIT

1er juillet 2021

Table of contents

Using ANN for power consumption estimation

Improving energy-efficiency of ICT

Knowledge is Power

Power consumption knowledge is crucial

- Datacenter operators
  - Power capping
  - Billing in colocation centers
  - Improving energy efficiency using autonomous systems
- Users and developers
  - Increasing their awareness
  - Choosing the best application or library

But power meters are costly

and

Linear classical model: error ~ 10-15%

Methodology

Context:

- power consumption knowledge
- watt meters are costful
- Linear classical model: error ~ 10-15%

1. Data preprocessing
   a. Multiple parallel time series with imperfect hardware

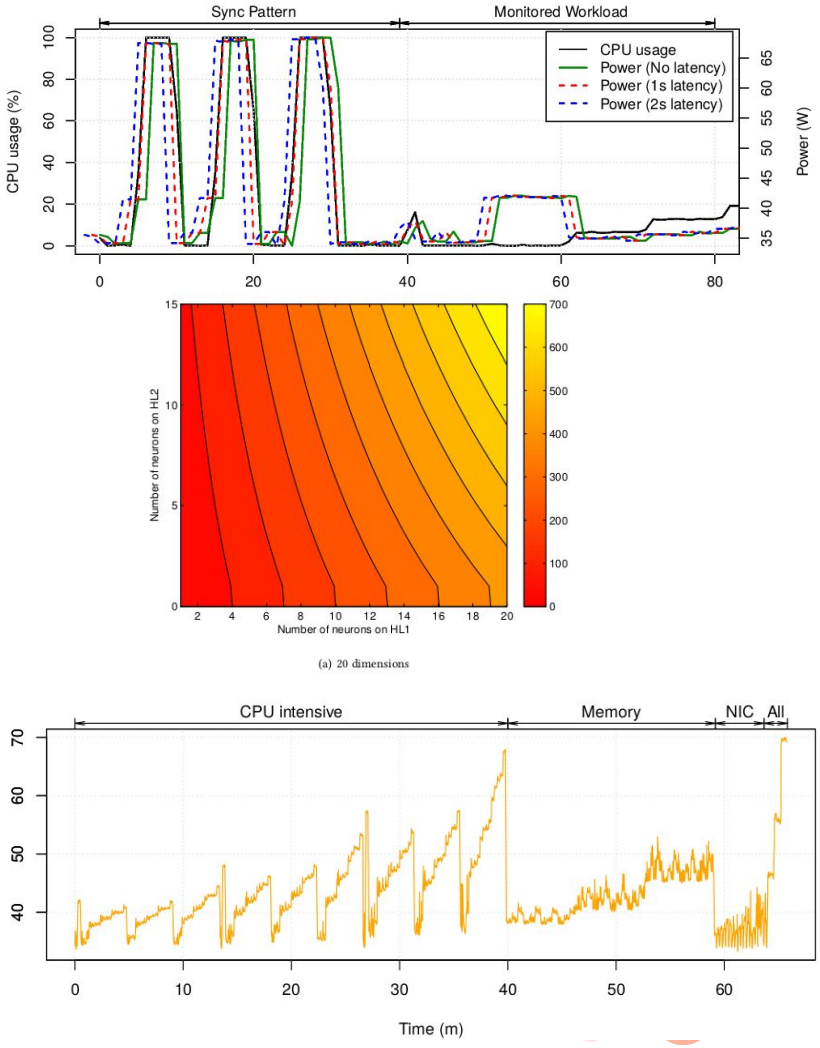
      Da Costa, G., Pierson, J. M., & Fontoura-Cupertino, L. (2017). Mastering system and power measures for servers in datacenter. SUSCOM

2. Network topology
   a. Impact of number of neurons on 2 hidden levels (fully connected)

      Costa, G. D., Pierson, J. M., & Fontoura-Cupertino, L. (2020). Effectiveness of Neural Networks for Power Modeling for Cloud and HPC: It's Worth It!. TOMPECS

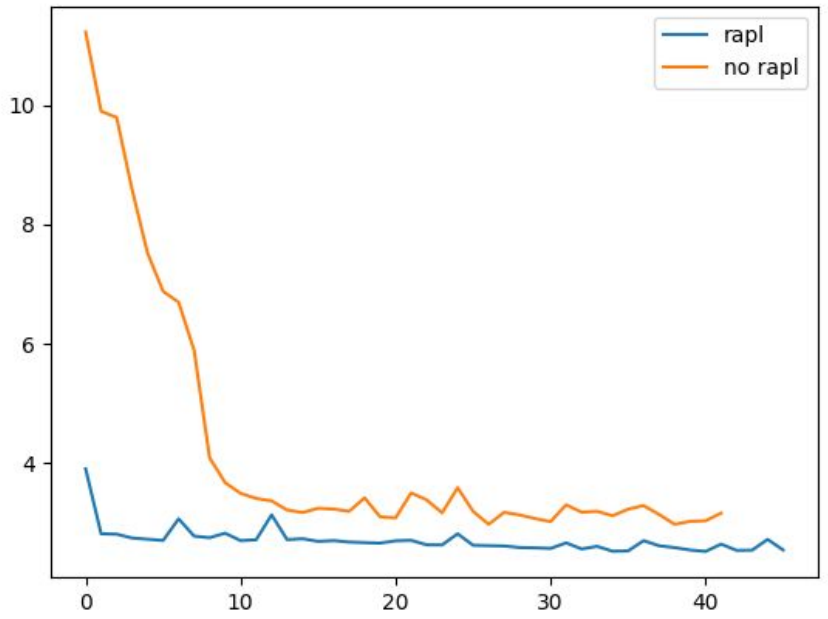
3. Large scale of experiments
   a. One experiment : 1.3 Go/node
   b. Need of variable reduction
   c. Specific benchmark for learning

      Da Costa, G., Pierson, J. M., & Fontoura-Cupertino, L. (2020). Fast maximum coverage of system behavior from a performance and power point of view. CCPE



Current work and difficulties

- Feature reduction (hundreds of features)
  - Correlation matrix
  - Residual method
  - PCA
  - …
- Explore other models
  - Larger networks
  - Polynomial models
  - On-line learning
- Change the point of view
  - Impact of particular features on speed and quality



Smart DVFS

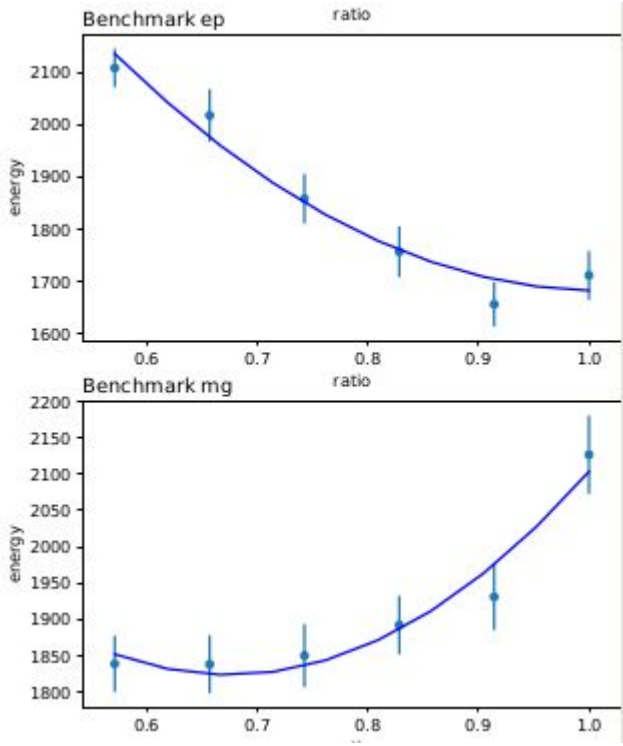
Each application has its preferred frequency

Applications are black boxes

1. Detect the current resources consumed
   a. Hundreds of features
2. Evaluate which is the best frequency for this particular resource consumption

Example:

- Lots of memory access
  - Reduce the processor frequency
- Lots of CPU usage
  - Increases the processor frequency



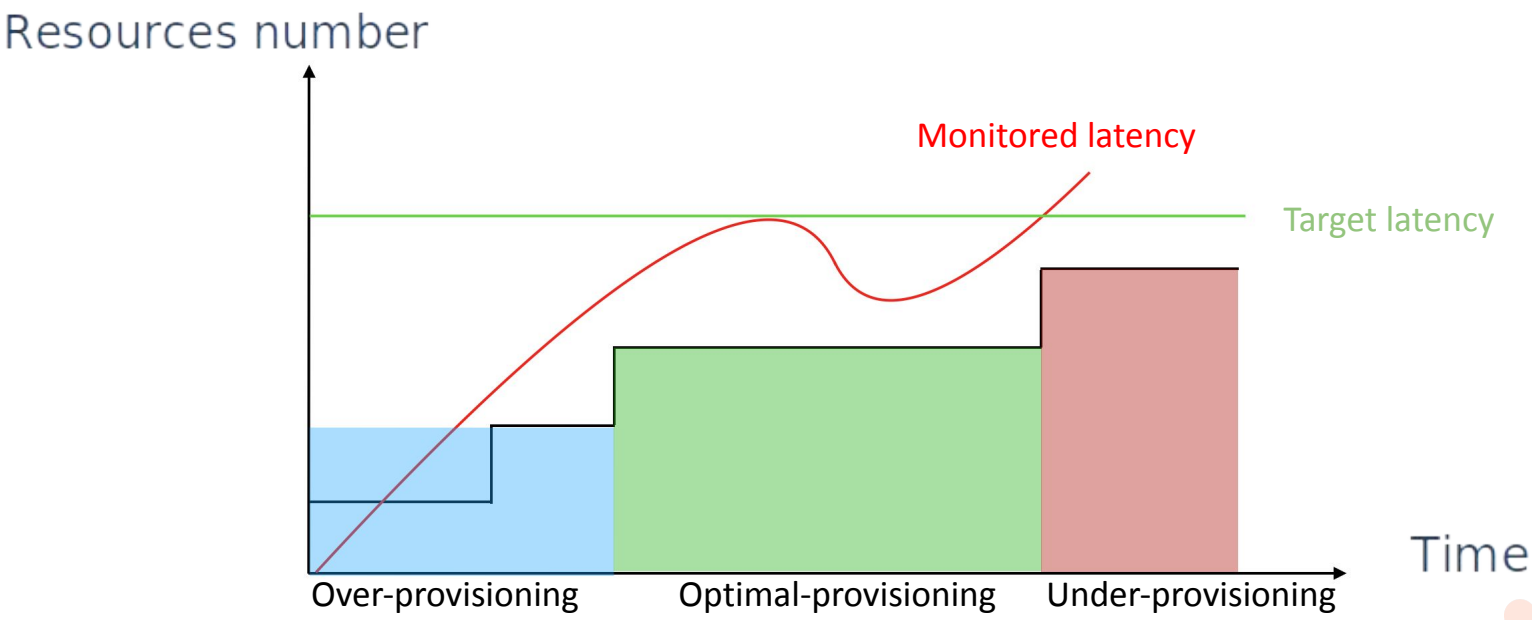
Current state

- Large scale monitoring of features
  - Done but slow
- Feature reduction
  - Needed to increase precision and reduce overhead
  - Correlation matrix
  - PCA
  - Not satisfactory
- Model to find the best frequency
  - Decision tree
  - Random Forest
  - ANN

Reinforcement learning

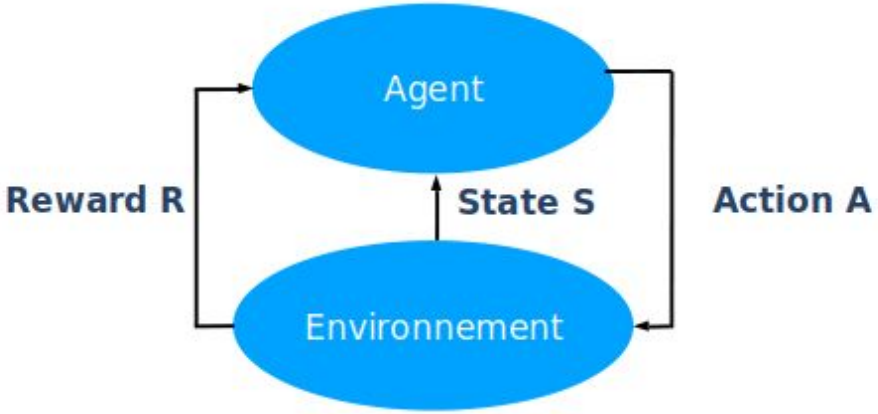
Context and problem

- Cloud elasticity
- Context CAAS
- An auto-scaler autonomously takes decisions to
  - Optimize the latency
  - Under the constraint of minimizing the number of allocated resources



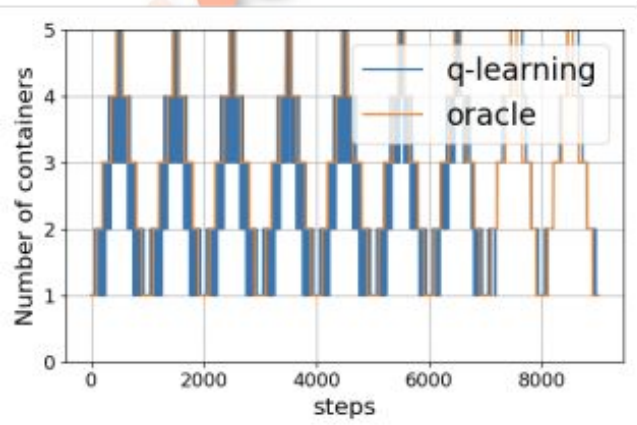
Reinforcement learning approach

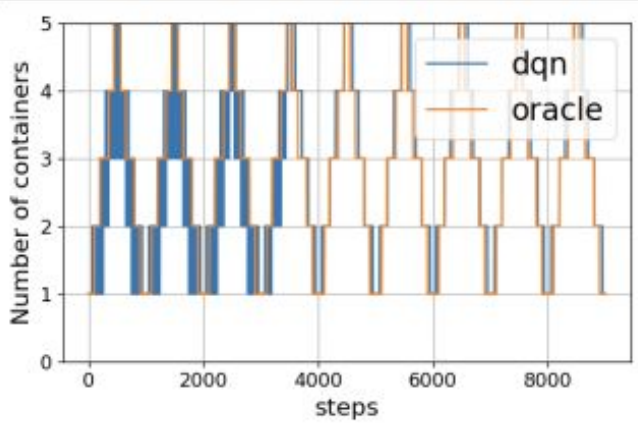
- Actions:
  - u: increases the number of containers by one (reactive)
  - s: stays with the same number
  - d: decreases the number by one
- States:
  - Number of containers
  - 95th percentile latency
  - Requests per second
  - cpu, ram mean
- Algorithms
  - HSLinUCB (Contextual Bandits family algorithm)
  - Q-Learning
  - Deep Q-learning

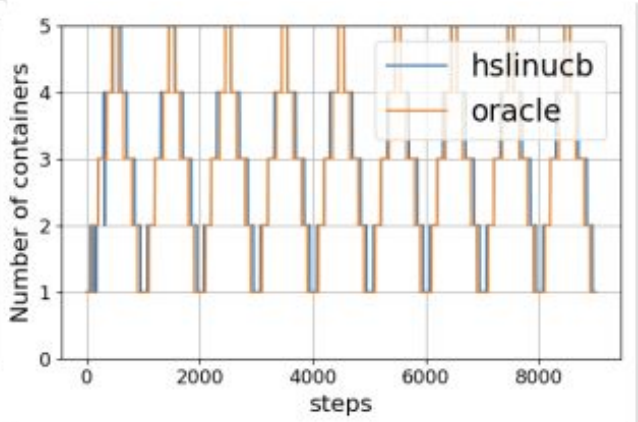


Some results

David Delande, Patricia Stolf, Raphaël Féraud, Jean-Marc Pierson, André Bottaro. HSLinUCB: Horizontal Scaling in Cloud Using Contextual bandits. Europar 2021







- Q-Learning
  - Learning-rate = 0.1 ; discount-factor = 0.9 ; increases by 0.000166 at each step
- Deep Q-Learning
  - Context scaled between [0,1]
  - 2 hidden layers (24 and 12 neurons) with ReLu activation function
  - Learning-rate = 0.01 ; discount-factor = 0.9 ; batch size = 50
  - For simulated experiments: target network update every 100 steps ; increases by 0.000333 at each step
- Comparison between Hot-start and cold-start: how the agent performs on cold-start while learning a new environment ? how the agent performs on an already learnt environment ?

Some conclusions

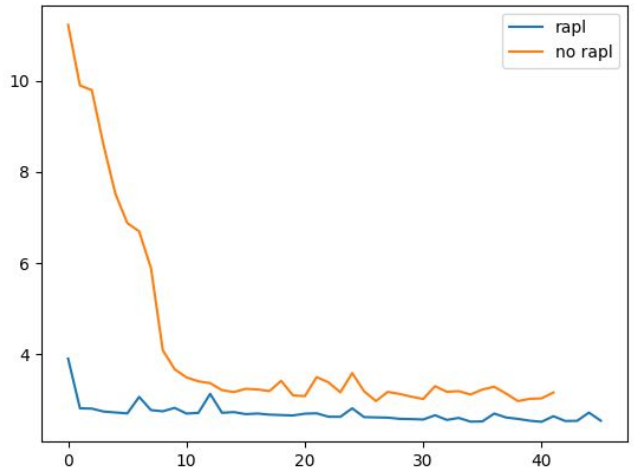
- In the case of cold start, HSLinUCB outperforms the other algorithms by quickly converging

- In the case of hot start, for seen and unseen contexts HSLinUCB obtains quite similar results than DQN (high learning performances of Deep Neural Networks). Q-learning obtains worse results on unseen contexts while it obtains similar results on seen contexts.

- Current work : Study workload and Cloud stationary changes

Difficulties, challenges

- Size of dataset to obtain
- Feature reduction (correlation, PCA….)
- Data quality (sometimes imprecise measurements)
- Robustness and bias
- Many data = time series with variable time steps (aggregation relevant ?)
- Feedback measurement online for RL
- Study the impact of the structure of the neural network
- Need of online learning (re-learn sometimes)...



What about deep learning approaches energy consumption ?

- Tradeoff between benefits and consumption for learning (exploration, exploitation)
- IA is consuming but also permits some progress
- Compare different network architecture (performance and consumption)
- Sepia team has tools to monitor applications

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

https://arxiv.org/pdf/1906.02243.pdf

Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell, Ananya Ganesh, Andrew McCallum


IRIT