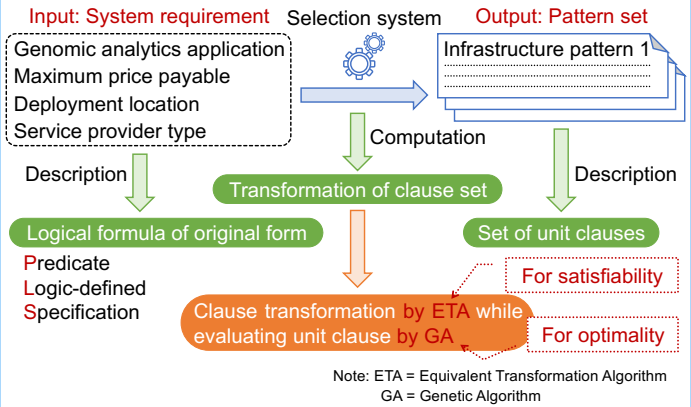
Cloud Resource Selection based on PLS Method for deploying Optimal Infrastructure for Genomic Analytics Application

Katsunori Miura (Kitami Institute of Technology)
Courtney Powell (Hokkaido University)
Masaharu Munetomo (Hokkaido University)

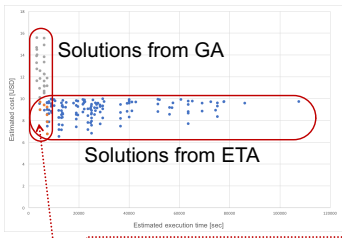
1. INTRODUCTION

We propose a method for determining cloud resources that concurrently meet system infrastructure satisfiability and optimality requirements.

Overview



Importance of combining ETA and GA



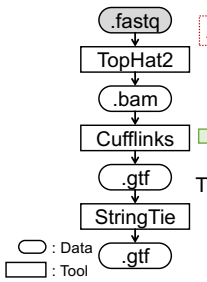
ETA
Finds all valid infrastructures, but cannot determine their optimality.

GA
Finds all optimal infrastructures, but cannot determine their satisfiability.

Orange points represent system infrastructures that are both valid and optimal with regards to system requirements.

2. GENOMIC ANALYTICS APPLICATION

A genomic analytics application is a workflow composed of various tools such as TopHat2, HISAT, and StringTie.



A workflow is described by list format in our method.

Description

$$[\,[step1, \text{“}tophat2\text{”}, [2.0]]$$
$$[step2, \text{“}cufflinks\text{”}, [2.0]]$$
$$[step3, \text{“}stringtie\text{”}, [2.0]]\,]$$

The meaning of the elements in the list is as follows:
First element: Step number
Second element: Tool name
Third element: Input file size given to a tool [GB]

3. CLOUD RESOURCE SELECTION BASED ON PLS METHOD

The PLS method finds all feasible infrastructures that completely satisfy system requirements by transforming clause sets.

Cloud resource selection process

Step 1: Construct a PLS reflecting system requirements
A PLS has the following form:

$$\forall_{\bar{v}}\{E_1 \wedge E_2 \wedge \cdots E_n\},$$

where E is an atomic formula (atom), $\bar{v}$ is a set of variables appearing in the formula.

Step 2: Generate a definite clause cl from the PLS
A definite clause cl has the following form:

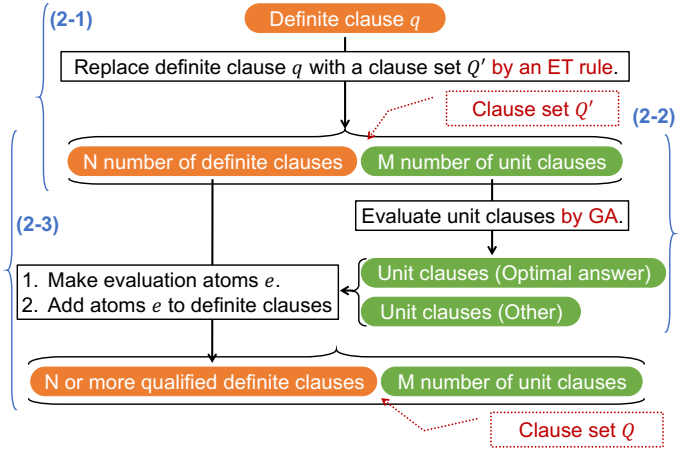
$$ans(v_1, v_2, \cdots v_i) \leftarrow E_1, E_2, \cdots E_n,$$

where the term of an ans atom is all variables in $\bar{v}$.

Step 3: Transform clause set $\mathbb{D}$
Clause set $\mathbb{D}$ is transformed while preserving the declarative meaning of the initial state in accordance with the following procedure, where the initial state is $\{cl\}$.

(1) Select a definite clause q from clause set $\mathbb{D}$, where $\mathbb{D} = \{q\} \cup D$.
A definite clause q is randomly selected from clause set $\mathbb{D}$. *(NB: From the viewpoint of efficient computation, the definite clause selection method is important, but this method is left for future work.)*

(2) Replace clause q with a clause set Q, where $\mathcal{M}(\{q\}) = \mathcal{M}(Q)$.



(3) Make new clause set $\mathbb{D}$ by joining clause sets D and Q.

Termination condition of the transformation

- Clause set $\mathbb{D}$ is a unit clause set or an empty set.
- A unit clause determined to be optimal by GA is obtained.

4. SYSTEM RQUIREMENTS AS INPUT DATA

To describe system requirements with respect to genomic analytics application deployment, four (4) atoms are utilized:

$$\begin{aligned}\forall_{\{*structure\}}\{&Envrionment(*\,structure, [[step1, \text{“}tophat2\text{”}, [2.0]]\\ &\qquad [step2, \text{“}cufflinks\text{”}, [2.0]]\\ &\qquad [step3, \text{“}stringtie\text{”}, [2.0]]]) \wedge\\ &Cost(*\,structure, 10.0) \wedge\\ &Location(*\,structure, [\text{“}Virginia\text{”}, \text{“}Tokyo\text{”}]) \wedge\\ &Policy(*\,structure, \text{“}public\text{”}, \text{“}single\text{”})\}\end{aligned} \quad (1)$$

- *Environment* atom: Genomic analytics application to deploy
- *Cost* atom: Maximum price payable
- *Location* atom: Deployment location or region
- *Policy* atom: Service type and number of cloud service providers

5. OPERATION EXPERIMENT OF PLS METHOD

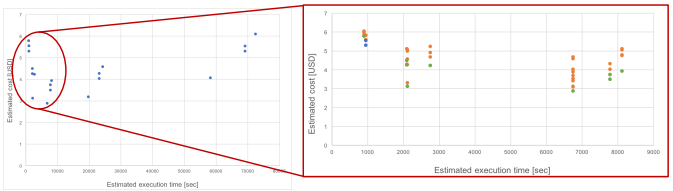
Purpose

Find optimal infrastructures that satisfy (1).

Computation

As a result of transformations of clause set $\mathbb{D}$,
sixty (60) definite clauses and nineteen (19) unit clauses
are obtained from the initial state.

In process (2-2) of Step 3, the unit clauses are evaluated by GA.
Green points represent optimal solutions.



Example of an optimal solution

$$\begin{aligned}ans(\,&[\text{“}aws\text{”}, \text{“}r4.16xlarge\text{”}, \text{“}Virginia\text{”}, [step1, \text{“}tophat2\text{”}, [2.0], 892.49]],\\ &[\text{“}aws\text{”}, \text{“}m4.xlarge\text{”}, \text{“}Virginia\text{”}, [step2, \text{“}cufflinks\text{”}, [2.0], 1.0]],\\ &[\text{“}aws\text{”}, \text{“}m4.xlarge\text{”}, \text{“}Virginia\text{”}, [step3, \text{“}stringtie\text{”}, [2.0], 1.0]]) \leftarrow\end{aligned}$$

6. CONCLUSION

- The proposed method makes it possible to find infrastructures that concurrently meet system satisfiability and optimality requirements.
- The proposed method can be extended to deploy other applications in addition to genomic analytics, without modifying the resource selection algorithm, simply by adding new atoms and their applicable ET rules.