Application of Multi-agent Reinforcement Learning to CELSS Material Circulation Control

Tomofumi Hirosaki[a], Nao Yamauchi[b], Hiroaki Yoshida[c], Yoshio Ishikawa[b], and Hiroyuki Miyajima[d]

[a] *Fujitsu Limited, Earth Science Systems Department, Science Systems Division*
9-3, Nakase 1-Chome, Mihama-ku, Chiba City, Chiba 261-8588, Japan
Tel: +81-43-299-3250, Fax: +81-43-299-3013, E-mail: hirosaki@ssd.ssg.fujitsu.com

[b] *Nihon University, Department of Aerospace Engineering, College of Science and Technology*
7-24-1 Narashinodai, Funabashi, Chiba, 274-8501, Japan
Tel: +81-47-469-5415, Fax: +81-47-467-9569, E-mail: {nao, yishi}@stone.aero.cst.nihon-u.ac.jp

[c] *Nihon University, Department of Precision Machinery Engineering, College of Science and Technology*
7-24-1 Narashinodai, Funabashi, Chiba, 274-8501, Japan
Tel: +81-47-469-5245, Fax: +81-47-467-9504, E-mail: yoshida@mosquito.eme.cst.nihon-u.ac.jp

[d] *Tokyo Jogakkan Junior College, Department of Information and Social Studies*
1105 Tsuruma, Machida, Tokyo, 194-0004, Japan
Tel: +81-42-796-9464, Fax: +81-42-799-2652, E-mail: miyajima@m.tjk.ac.jp

Abstract

A Controlled Ecological Life Support System (CELSS) is essential for man to live for a long time in a closed space such as a lunar base or a Mars base. Such a system may be an extremely complex system that has a lot of facilities and circulates multiple substances. Therefore, it is a very difficult task to control the whole CELSS. Thus by regarding facilities constituting the CELSS as agents and regarding the status and action as information, the whole CELSS can be treated as a multi-agent system (MAS). If a CELSS can be regarded as MAS, the CELSS can have three advantages with the MAS. First, the MAS need not have a central computer. Second, the expendability of the CELSS increases. Third, its fault tolerance rises. However, it is difficult to describe the cooperation protocol among agents for MAS. Therefore in this study, we propose to apply reinforcement learning (RL), because RL enables an agent to acquire a control rule automatically. To prove that MAS and RL are effective methods, we have created the system in Java, which easily gives a distributed environment that is the characteristic feature of an agent. In this paper, we report the simulation results for material circulation control of the CELSS by the MAS and RL.

Keywords:

CELSS, multi-agent system, reinforcement learning, cooperation protocol

Purposes and Background

A Controlled Ecological Life Support System (CELSS) allows man to live for a long time in a closed space such as a lunar base or a Martian base [1]. The CELSS is an extremely complex system that provides a lot of facilities and circulates multiple substances [2]. Therefore controlling operation over the whole CELSS is quite a hard task. As methods to solve the task, applications of Fuzzy [3][4], AI [5], and Intelligent Automated Control [6] have been proposed. However, these methods have the necessity for clearly describing the control law of the entire system or among facilities. This means all facilities must be operated by the commands from the master computer. Therefore, when new facility is added to the system, it must be necessary to reconstruct the entire control law, because of dependence of each facility. Moreover, to get ready for sudden malfunction of the system, it must be necessary to define the control law in the full pattern that can be supposed. Consequently, there is a problem to need very complex software.

On the contrary, we propose an operation management method of the CELSS as follows.

(a) Regarding each facility as an agent, and its state and condition as information.

(b) Each agent distributed solves problems through the cooperation among other agents.

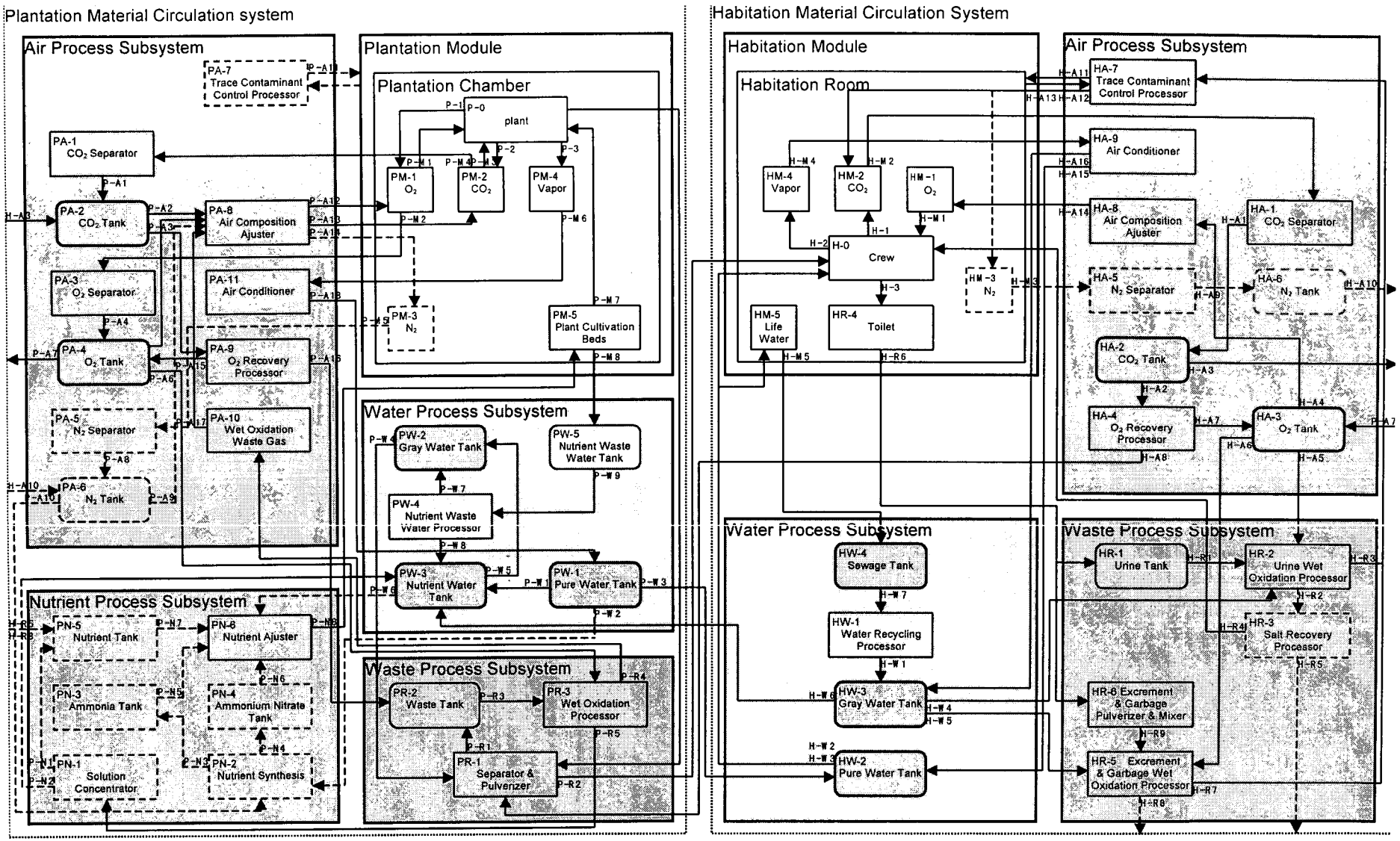

Fig.1 CEEF (CPEF & CAHF) Material Circulation System

(c) As the result, the optimum material circulation status is maintained in the whole system.

There is a multi-agent system (MAS) as a technique that treats multiple agents. The MAS has the following three advantages.

(1) Each agent controls material circulation cooperating with surrounding agents. Therefore, no control law for the entire system needs to be described. Also, no central control computer needs to be existed.

(2) Each agent is independent, so that the extension and composition change of system can be made easily by an addition/change of agents. Moreover, no system shutdown is necessary, thus excellent extendibility and flexibility is ensured.

(3) Even if fault occurs to a facility, the remaining facilities (agents) can shift to a reduced operation autonomously, because control laws only may be described among facilities mutually connected. In fact, the system can be reconstructed autonomously. As the result, the fault tolerance of the system can rise.

On the other hand, realizing cooperation protocol is a key issue in the MAS technique. Furthermore, in order to let an agent perform an autonomous action, a mechanism to adapt itself to an inexperienced environment should be established. Hence we have aimed at that an agent itself can acquire a cooperation protocol by learning, by applying the reinforcement learning theory [7].

In a large-scale system like CELSS, it is possible to describe the control law of each facility. However, since it is difficult to describe the control law of whole material circulation, it is thought very effective to make each facility acquire a control law autonomously by learning.

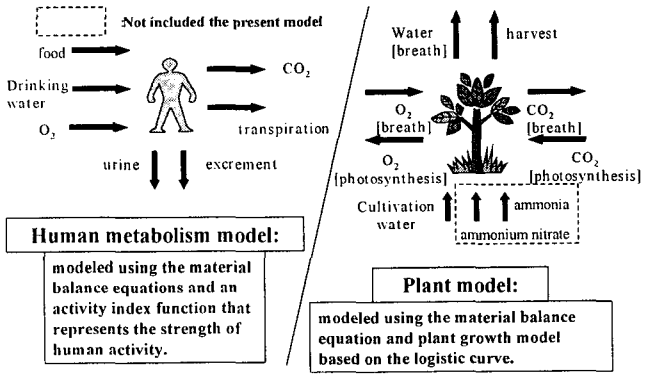

Fig.2 Model of Human and Vegetable (Plant)

Modeling

As a model of CELSS, Closed Ecology Experiment Facilities (CEEF) of Institute for Environmental Sciences in Japan was referred. As a research stage of a MAS application, only the circulation of O_2, CO_2, and H_2O were considered for simplification. The circulation of N_2, ammonia (NH_3), and ammonium nitrate (NH_4NO_3) is not included in the model (the solid line portion of Fig.1 is included in the model while the dotted line portion is not). In addition, the plant growth model and human activity model were based on literature [2].

Reinforcement Learning Theory

Outline

The MAS has three advantages, stated in section 1. On the other hand, to control the whole CELSS as MAS, cooperation protocol must be built into MAS. As a technique to let agents themselves acquire the cooperation protocol, we have introduced reinforcement learning (RL). RL enables agents to advance learning with evaluation (reward) as a result of actions, without a correct output (teaching signal) that is essential for the neural network. It must be effective to introduce RL into the problem that the control law cannot be described in advance.

For this problem, Q Learning, the most famous among RL theories, is applied. Q Learning estimates weight to the set of states and actions, that is a rule. This weight is referred to as Q value. A rule is defined as the combination of a sensible state x and a selectable action a by an agent, then the Q value is described as $Q(x,a)$. Q value will be updated by eq.(1), when an agent in the state x_t selects an action at in time t, and as a result, the state is changed to x_{t+1} and reward r_t is obtained.

$$Q_{t+1}(x_t,a_t) = (1-\alpha)Q_t(x_t,a_t) + \alpha(r_t + \gamma \max_{k\in A} Q_t(x_{t+1},a_k)), \quad (1)$$

where α represents the learning rate and $0<\alpha\leq 1$. γ denotes the discount rate and a_k is the action which gives maximum Q value at the state x_{t+1}.

The first term of the right side in eq.(1) represents the value obtained by the previous learning experience. The second term expresses present learning result (reward). The third term implies the optimal action in the future.

Therefore, it is assumed that a better action has higher Q value, while a worse action has lower Q, in a group of actions selectable at a certain state.

Definition of State, Action, and Reward

The definition for a tank as an example is as follows,

State: danger level of the tank,

Action: quantity of gas transfer between tanks and module gas concentration adjustments, and

Reward: degree of improvement at the danger level of the tank.

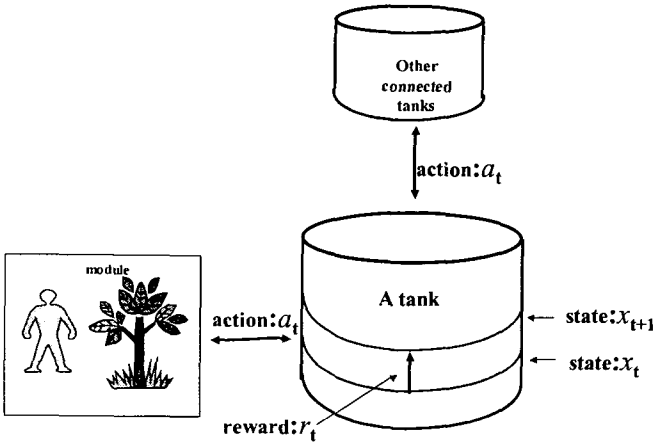

Fig.3 The concept of state, action, and reward

Table1 Definition of dangerous level

Gas quantity in the tank		Danger level
Lower limit	Upper limit	
Maximum allowable quantity	Maximum quantity	2
(target+Maximum allowable quantity) /2	Maximum allowable quantity	1
(target+Minimum allowable quantity) /2	(target+Maximum allowable quantity) /2	0
Minimum allowable quantity	(target+Minimum allowable quantity) /2	-1
Minimum quantity	Minimum allowable quantity	-2

The Action Selection Rule

The following procedure has been adopted as an action selection rule.

(a) Calculate $\tau = \exp(5 \times n_i / N)$, where $0 < \tau < 1$, and n_i is the number of a learning at present and N is the target number of learning practices. Unique n_i is defined about each state,

(b) Generate a random number between 0 and 1. If it is less than , further generate another to choose an action at random,

(c) Otherwise select an action for the maximum Q value.

It is expressed in the above that learning is carried out with random action selection in less experience stage, and that an action is selected based on learning results after much experience. In addition, the number of learning practices is defined one by one for all states. If a disturbance turns less experienced state from much experienced state, learning is more activated. Therefore it is thought that more experience in various disturbances can enhance the ability to adapt to wider range of states.

Implementation Environment

The simulation software is built in Java language, which realizes easily a distributed environment that is the characteristic feature of an agent. In the material circulation model, O_2 tank, O_2 recovery processor, CO_2 tank and so on are regarded as an agent, respectively. The specification of a facility is described in a remote object imitating the real world. On the other hand, the cooperation protocol (the control law) is described in each agents. The remote object is operated based on the control plan that is built up by the agent. In order to enable the placement of all these processes on discrete computer, Java RMI™ is used for a means of communications.

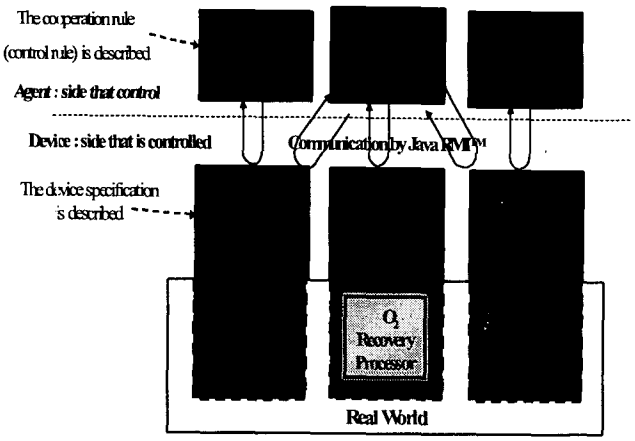

Fig.4 The implementation environment overview

Simulation Conditions

Simulation was carried out using the built software. The given conditions are shown below.

- As for plants, rice, soybean, lettuce, tomato, sweet potato, and sesame are grown. The amount of cultivation is determined so that one person can live. The length of stay is assumed as 800 days.

- The plantation module is divided into some partitions for each species. Shifting harvest time ensures stable ration of foods.

- Since the tank capacity might be exceeded until the material circulation reaches a steady state, the outer buffer tanks are prepared. That is, when the outer buffer tanks are used, the operation of CELSS becomes the open system. After that, when the outer buffer tanks becomes unused, it becomes the closed system.

- The lighting system of the plantation module provides light/dark periods in a day to fluctuate the degree of plant growth.

- The RL parameter are determined as follows: learning rate α = 0.4, a discount rate γ =0.85, and target number of learning practices = 5000.

- Simulation is carried out in the following pattern,

a. With no disturbance,

b. With disturbance,

c. With no disturbance (taking over Q value of a),

d. With disturbance (taking over Q value of b),

"Taking over Q values of a" shows that simulation is carried out once again using learning result of "a" from the beginning.

- As a disturbance, the leakage of 2400g/day is added to the oxygen tank for habitation module for three days at the three timings (300,400,500th day).

Simulation Results

The history of content in the O_2 tanks for habitation module are shown in Fig. 5 - 8.

Fig.6 is the simulation result with disturbance. It can be shown that the variation of oxygen with disturbances added on the 300th, 400th, and 500th day has become smaller in order of 300th, 400th, 500th day. This shows that the control law to adapt to the situation of the disturbance occurrence could be obtained autonomously.

Fig.7 and 8 are the results of the repeated simulation by using the result of learning (a set of Q value) obtained through the simulations shown in Fig.5 and 6 respectively. Though the phenomenon that the amount in the oxygen tank in the habitation module increases rapidly on about 270th day is seen in Fig.7, such a phenomenon is not seen in Fig.8. We consider that a better control law could be obtained by the learning of the simulation result with the disturbances.

Conclusion

The above results lead to the following conclusions: material circulation of CELSS can be simulated with the application of MAS and RL. It has been shown that a control rule and a cooperation rule are autonomously gained by the learning.

Furthermore, in order to express MAS, the Java RMI™ function is used. Consequently, facilities and agents can be treated as independent processes, which enables the addition and reconstruction of a facility easily. Thus, a possibility to construct a highly expandable system has been demonstrated by applying MAS to CELSS.

Future Subject

- Since the present Q table has a fixed size, it is not easy to succeed a learning result when a facility is appended. Therefore it is essential to consider the management method of learning results so that the existing learning results can be succeeded.

- We will make CELSS model more precise so that facilities of the whole CEEF are included.

References

[1] Peter Eckart. 1994. *Spaceflight Life Support and Biospherics.* Microcosm Press.

[2] H. Miyajima, Y. Ishikawa 2000. Development of Simulation Model and Its Application to an Integration Test Project of CEEF. Presented at the ICES, SAE No.2000-01-2334.

[3] H. Miyajima, H. Yoshida, Y. Ishikawa 1999. New Sequential Fuzzy Linear Programming Method Using "m-operator" and its Application to a Closed Environmental Life Support System. *Journal of Japan Society for Fuzzy Theory and Systems* 11(6): 1049-1058.

[4] A. Ashida, K. Nitta, M. Takatsuji, K. Matsumoto 1991. Application of the Fuzzy Algorithm to the CELSS. *CELSS JOURNAL* 13(1): 53-60.

[5] H. Miyajima, Y. Ishikawa 2000. Development of Simulation Model and Its Application to an Integration Test Project of CEEF. SAE No.2000-01-2334.

[6] Debra Schreckenghost, Daniel Ryan, Carroll Thronesbery, Peter Bonasso, Daniel Poirot 1998. Intelligent Control of Life Support Systems for Space habitats. Proceeding national Conference Artificial Intelligence 15th: 1140-1145

[7] William Little: Approaching Gaia 1999. Intelligent Automated Control in Bioregenerative Life Support Systems. Presented at the ICES, SAE No.1999-01-2082.

[8] Richard S. Sutton, Andrew G. Barto 1998. *Reinforcement Learning*: An Introduction. MIT Press.

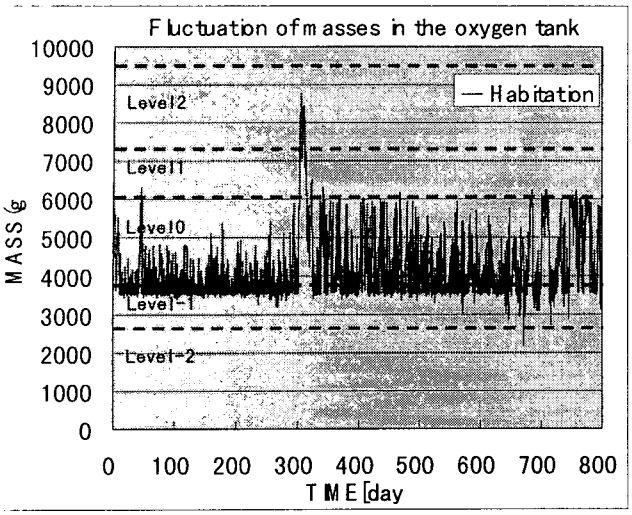

Fig.5 The history of [a] O_2 tank content (with no disturbance)

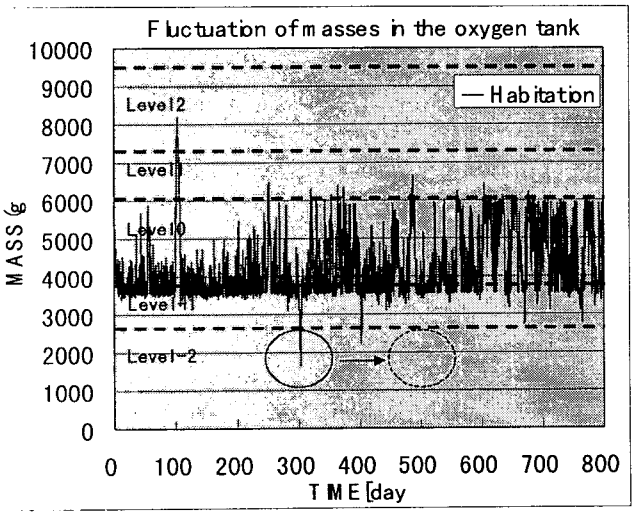

Fig.6 The history of [b] O_2 tank content (with disturbance)

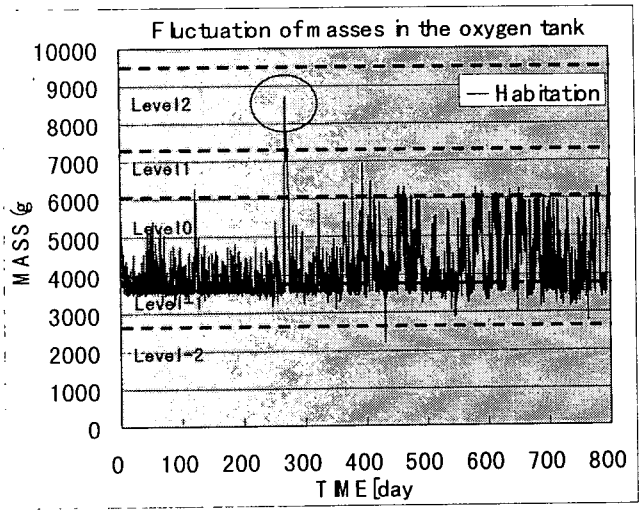

Fig.7 The history of [c] O_2 tank content(with no disturbance, taking over Q value of a)

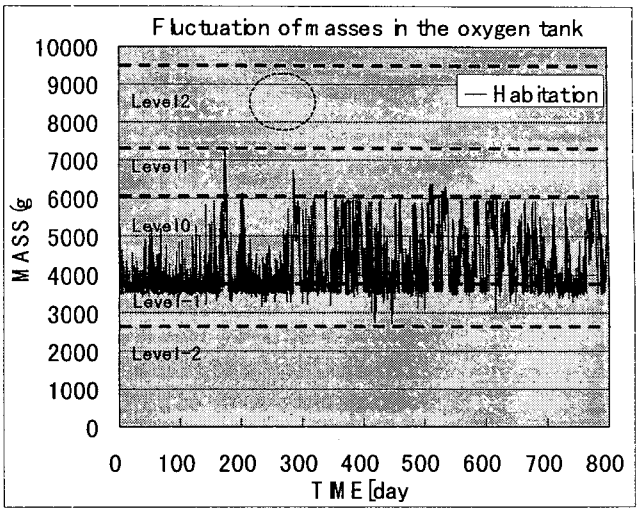

Fig.8 The history of [d] O_2 tank content(with disturbance, taking over Q value of b)