Method of Minimizing Costs in Consideration of System Backup Intervals and Expected Costs for Information Infrastructure of University

SHINYA MIZUNO[†1] MIKI SHINOHARA[†2] HARUKI INOUE[†3]
TAKAHIRO HASEGAWA[†3] NAOKAZU YAMAKI[†3]

Abstract: In addition to the evolution of today's digital society, which includes the spread of the Internet and the invasion of smart devices, we are now clearly in the age of big data and are, therefore, seeing an explosion in the volume of unstructured data. Corporate data is growing by approximately 60% per year, and is expected to grow further in the future. Viewed from another angle, the huge size of this data means that massive damage will occur if the data is lost. If data loss occurs due to human error, computer viruses, machine failure, natural disasters, or terrorism, significant losses will occur, either due to payment of compensation or lost opportunities due to the need to suspend operations. In order to prevent this, an appropriate backup system is required. This does not mean just saving data, but maintenance and replacement of servers and network devices is also necessary. However, whereas reliability will certainly be increased if these devices are regularly replaced, costs will also skyrocket. It is necessary to use not only experience, but appropriate indicators when maintaining information devices. At this time, systems should be designed by estimating a backup interval that will minimize expected costs based on the relationship between system loss probability, system loss costs, and system backup costs. In this study, we will model the relationship between system loss, system loss costs, and backup costs, and derive a recommended value for the backup interval as an optimized solution. Furthermore, we will add the probability distribution for system loss and verify this from multiple angles including system configuration. Current information devices rarely operate alone, and are often in multi-device configurations such as RAID. In this study, we will also model cases involving multi-device configurations. Finally, we propose the management of backup for information infrastructure of University.

Keywords: Simulation, Optimization, Backup, Reliability, Manegement

1. Introduction

TIn addition to the evolution of today's digital society, which includes the spread of the Internet and the invasion of smart devices, we are now clearly in the age of big data and are, therefore, seeing an explosion in the volume of unstructured data. Corporate data is growing by approximately *60%* per year, and is expected to grow further in the future[3]. Viewed from another angle, the huge size of this data means that massive damage will occur if the data is lost. If data loss occurs due to human error, computer viruses, machine failure, natural disasters, or terrorism, significant losses will occur, either due to payment of compensation or lost opportunities due to the need to suspend operations[6]. In order to prevent this, an appropriate backup system is required[7].

This does not mean just saving data, but maintenance and replacement of servers and network devices is also necessary[4]. However, whereas reliability will certainly be increased if these devices are regularly replaced, costs will also skyrocket. It is necessary to use not only experience, but appropriate indicators when maintaining information devices. At this time, systems should be designed by estimating a backup interval that will minimize expected costs based on the relationship between system loss probability, system loss costs, and system backup costs.

In this study, we will model the relationship between system loss, system loss costs, and backup costs, and derive a recommended value for the backup interval as an optimized solution. Furthermore, we will add the probability distribution for system loss and verify this from multiple angles including system configuration. Current information devices rarely operate alone, and are often in multi-device configurations such as RAID[1]. In this study, we will also model cases involving multi-device configurations.

Several optimal backup policy is presented [2], [5]. Sandeh[5] is shown that the backup policy when we have N-job. They formulate the backup time over an infinite time span. Cunhua[2] shows that the backup model treat as cumulative damage model. They didn't refer to the cost. And these don't assume the environment of the big data. So we consider the management of backup and we optimize the cost balance.

We introduce the management of backup for Shizuoka University. And we propose a structure of a backup in information infrastructure of a university.

2. Modeling For System Backup

Among the costs that should be taken into consideration, the known one is backup costs, with system loss probability and system loss costs being unknown. Whereas it cannot be denied that if statistical information can be accumulated in these areas, greater accuracy can be achieved moving forward, we can fully imagine progress in IT and drastic innovations in the framework, so it is unclear whether alluding to the accuracy of these parameters itself has any meaning. Therefore, it is more meaningful to decide the backup interval by deriving the relationship between these parameters and by analyzing their sensitivity. We will take these as our objectives as we proceed with our modeling.

Here, system backup costs are referred to as S, system loss cost is h/day, system loss probability is p, and backup interval as

T(days).

2.1 Modeling for a single device system

First, we make model for a single device system. If we use a geometric distribution for the probability of a system loss occurring within t days of the final system backup, this is $(1-p)^{t-1}p$. At this time, the system loss cost is t days of loss, and will be th. However, we will ignore the system recovery costs. Therefore, the system loss costs expected value *H(T)* occurring within backup interval T is as expressed in equation (1).

$$H(T)=\sum_{t=1}^{T} th(1-p)^{t-1}p \qquad (1)$$

$$H(T)=\frac{T(1-p)^{T+1}-(T+1)(1-p)^{T}+1}{p}h$$

Based on this, the expected value *C(T)* for total costs per day for T days after the final backup can be expressed as

$$C(T)=\frac{1}{T}(H(T)+S)$$

$$C(T)=\frac{1}{T}\left[\frac{h\{1-(1-p)^{T}\}}{p}+S\right]-h(1-p)^{T}$$

Although calculations were performed by using a geometric distribution this time, if we change the probability distribution in (1), it is possible to calculate *C(T)* in the same way.

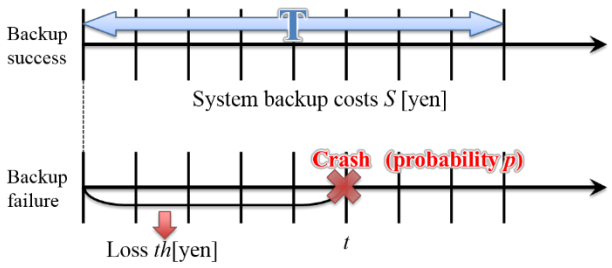

Figure 1 Process for backup

2.2 Modeling for multi device system

We consider the case for multi device system. We usually use multi devices for necessity of redundancy. Similarly single device system, we assume that failure rate of each device follows a geometric distribution with probability p. We have n device and $X_1, X_2, \ldots, X_n$ are the random variable that mean the time between failures for each device. Here, $X_1, X_2, \ldots, X_n$ are independent and identically distributed. So, we get following equations.

$$P(X_i=k)=p(1-p)^{k-1}$$

$$F_i(k)=P(X_i\le k)=1-(1-p)^{k}$$

We set $Y=\min_{1\le i\le n} X_i$, we get

$$P(Y\le k)=1-(1-F_i(k))^{n}=1-(1-p)^{kn} \qquad (3)$$

$$P(Y=k)=P(Y\le k)-P(Y\le k-1)$$

$$=(1-p)^{(k-1)n}\{1-(1-p)^{n}\} \qquad (4)$$

So, we get *H(T)* in case of The minimum value distribution of X_n from following equation.

$$H(T)=\sum_{t=1}^{T} th(1-p)^{(t-1)n}\{1-(1-p)^{n}\}$$

Next, we set $Z=\max_{1\le i\le n} X_i$, we get

$$P(Z\le k)=(F_i(k))^{n}=\{1-(1-p)^{k}\}^{n} \qquad (6)$$

$$P(Z=k)=P(Z\le k)-P(Z\le k-1)$$

$$=\{1-(1-p)^{k}\}^{n}-\{1-(1-p)^{k-1}\}^{n} \qquad (7)$$

We can get this case's *H(T)* like (5). We can analyze the case of a multi-device using these formulas.

3. Numerical Simulation

Here, we will introduce cases where the geometric distribution is used and cases where Weibull distribution is used. First, we consider the case of a single device.

3.1 Numerical Simulation using Geometric Distribution with single device

Here, as we calculate using the geometric distribution in (1), (2) is used as is. If we assume that system loss occurs once every *1000* days, we can suppose that *p=0.001*. Furthermore, the backup costs *S* are considered to be *100,000* yen. h fluctuates from *100,000* yen to *1* million yen, and the minimum backup interval is calculated as in Table 1 as *C(T)*. The relationship between *h* and *T* is shown in Figure 2. From Table 1, the backup interval with the minimum costs is calculated when system loss cost is applied. Furthermore, from Figure 2, we can see that as the loss costs increase, the backup interval with the lowest costs becomes shorter.

Table 1 System loss costs and optimal backup interval

h	T	C(T)
10	46	0.445
20	32	0.635
30	26	0.78
40	23	0.91
50	20	1.02
60	18	1.12
70	17	1.21
80	16	1.30
90	15	1.38
100	14	1.46

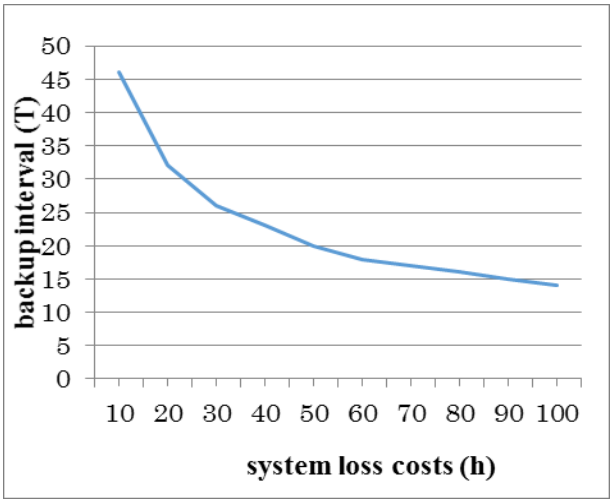

Figure 2 System loss costs and optimal backup interval

3.2 Numerical Simulation using the Weibull distribution

Next, we performed calculations using the Weibull distribution. The Weibull distribution is used to rate system reliability and is known from defect rate graphs and bathtub curves. Here, the calculation divides the initial defect period, the random failure, and the wear-out failure period. The Weibull distribution function is shown in (8).

$$F(t) = 1 - e^{-(\frac{t}{\beta})^{\alpha}} \quad t > 0 \qquad (8)$$

The parameters α, β used for each period are shown in Table 2. Here, m indicates the average defect interval.

Figure 3,Figure 4and Figure 5 show *C(T)* values corresponding to the changes in T when applying values of h from *10* to *100*. In all periods, when *h* is small, the *C(T)* transitions to a low value. Furthermore, in the initial defect period, the minimum *C(T)* value is higher than in the other periods, and from this we can see that this is a period in which care must be taken to reduce costs.

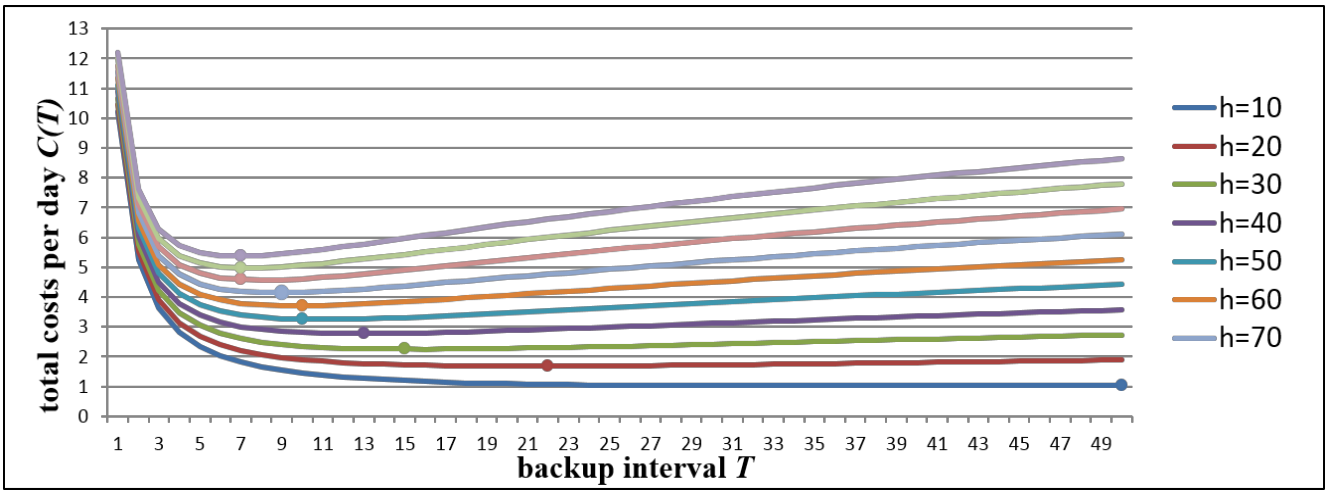

Figure 3 Transition in *C(T)* during initial defect period

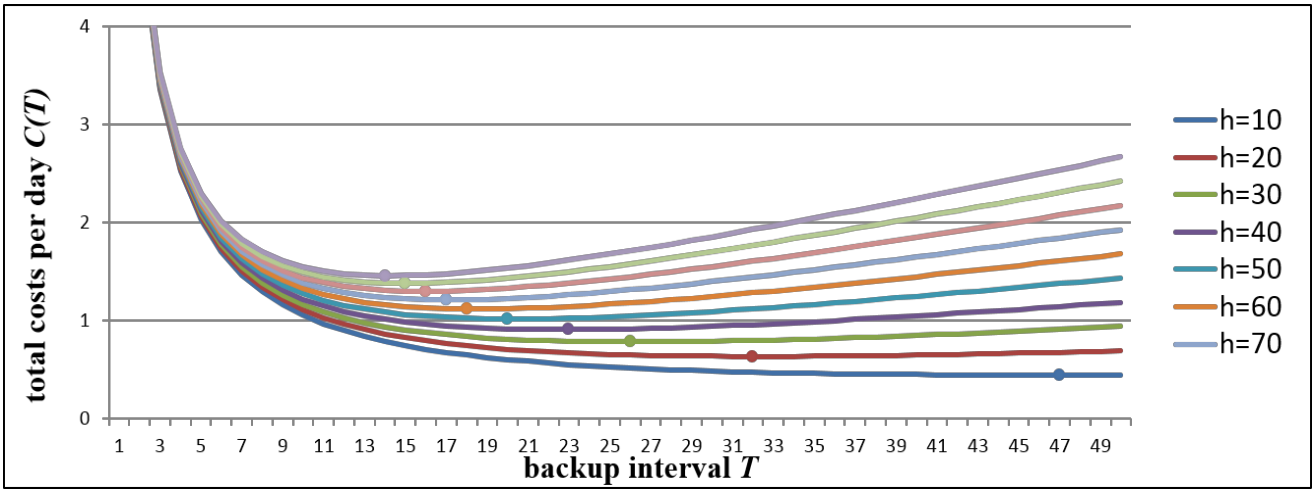

Figure 4 Transition in *C(T)* during random defect period

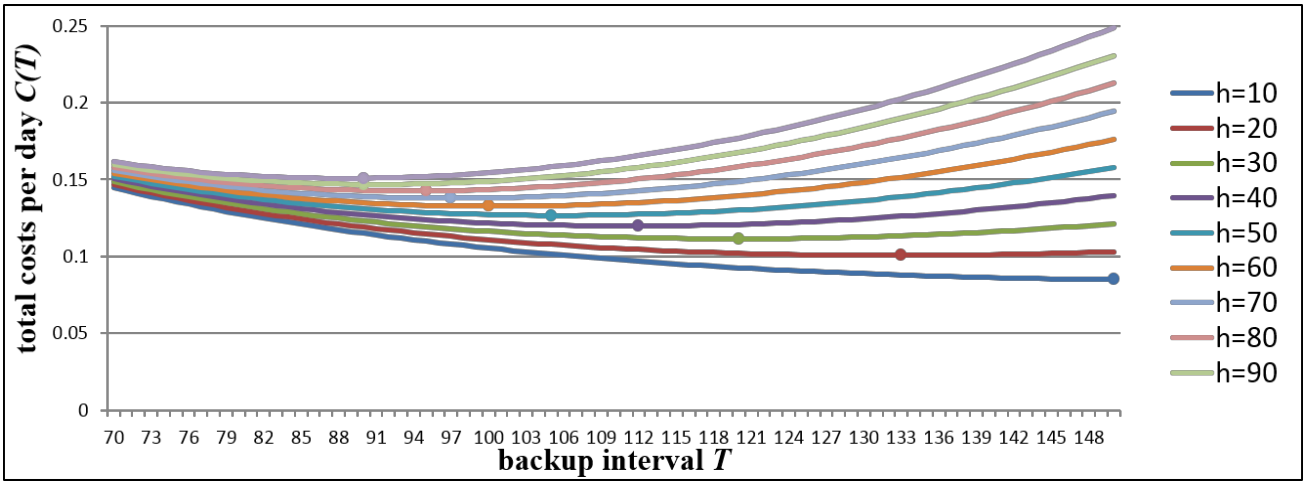

Figure 5 Transition in *C(T)* during wear-out defect period

Table 2 Parameters when using the Weibull distribution

	Initial defect period	Random failure period	Wear-out failure period
α	0.5	1	3
β	500	1000	1120
m	1000	1000	1000

Table 3 System loss costs and optimal backup interval for each period

h	Initial defect period		Random failure period		Wear-out failure period	
	T	C(T)	T	C(T)	T	C(T)
10	64	0.507	46	0.445	158	0.085
20	21	1.686	32	0.635	133	0.101
30	16	2.255	26	0.783	120	0.111
40	13	2.772	23	0.907	111	0.120
50	11	3.254	20	1.018	105	0.127
60	10	3.712	19	1.119	101	0.133
70	9	4.149	17	1.211	97	0.138
80	8	4.569	16	1.298	94	0.143
90	7	4.979	15	1.379	91	0.147
100	7	5.373	14	1.457	90	0.151

3.3 Numerical Simulation using Geometric Distribution with multi devices

Next, we compute the case of multi devices system using geometric distribution. If we have multi devices, before the 1st breakdown occurs, we would like to get a backup. So, we use (5). We assume that the values of parameters are same as the occasion of the single device. Table 4 shows the optimized backup period and cost for multi devices system.

Table 4 The optimized backup period and cost for multi devices system

Number of device	Optimized *T* (day)	Optimized *C(T)* (yen)
1	46	4454.563070775519
2	33	6286.995924667928
3	27	7687.549839761031
4	24	8863.276624820473
5	22	9897.703950342928

Increasing in the number of devices, we find out that a backup period is shorter. Table 5 and Figure 6 show transition for optimized *C(T)* in case of *2* devices changed value of h from *100,000* yen to *1,000,000* yen.

Table 5 Transition for optimized *C(T)* changed value of *h* with *2* devices

h (yen)	*T* (day)	*C(T)* (yen)
100000.0	33	6286.995924667928
200000.0	23	9006.957705542483
300000.0	19	11118.119567433048
400000.0	16	12912.15199325225
500000.0	14	14510.398534392893
600000.0	13	15954.94331661747
700000.0	12	17296.44481995731
800000.0	11	18559.096095227254
900000.0	11	19742.619470767022
1000000.0	10	20863.41983606384

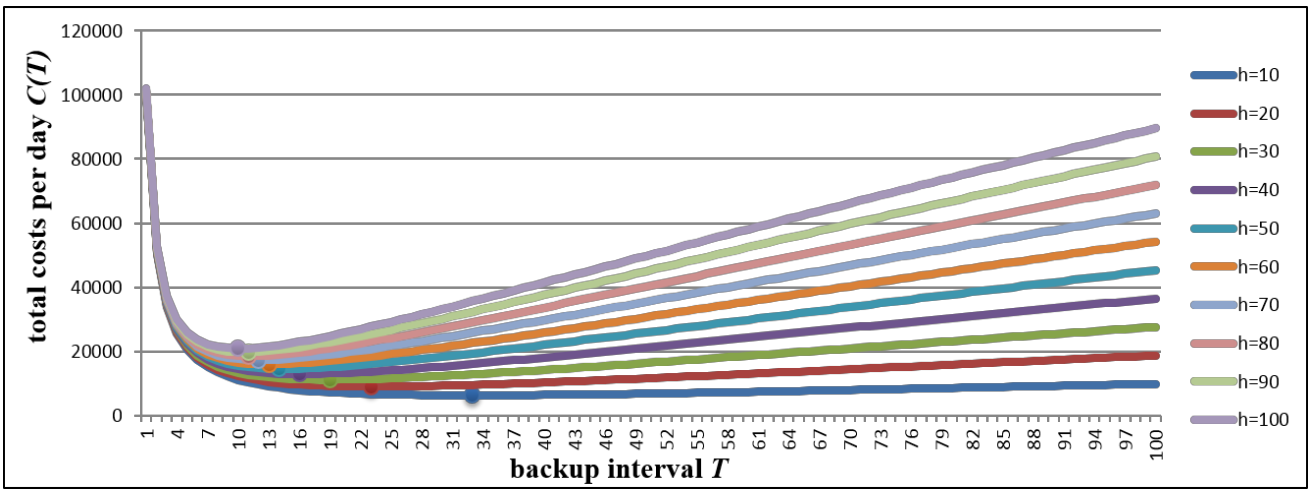

Figure 6 Transition for *C(T)* changed value of *h* with *2* devices

4. Application To Shizuoka University Information Infrastructure

It is important for a university to plan for BCP(Business Continuity Plan). For Shizuoka University, there were a lot of servers in campus before 2010. So we started to use the Cloud center. We have *2* types Cloud centers, Private Cloud Center and Public Cloud Center in Figure 7. In Private Cloud Center, there is a mission-critical system for University. In Public Cloud Center, there are a lot of Virtual Private Server(VPS) to use each researcher. It increases in the number of servers every year in Figure 8, now we use *334* Virtual Private Server(VPSs) as of April, 2014. VPS is installed in Public Cloud Center, VPS is used as each Homepage and calculation servers[4]. We have to support backup for VPSs. We calculate the most suitable backup period using the following condition. Generally, cloud enterprises indicate the rate of operation for VPSs, *99.999%* as SLA.

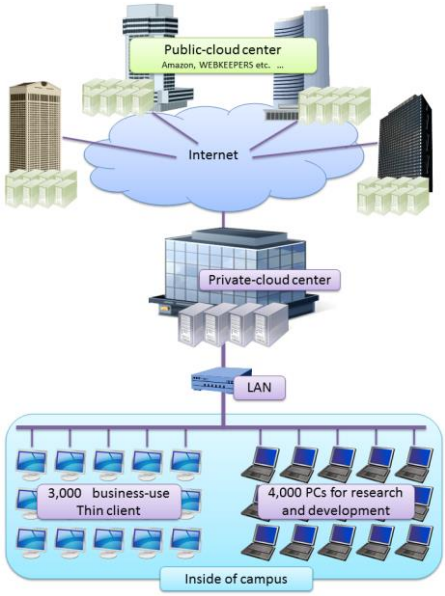

Figure 7 Cloud-information infrastructure of Shizuoka University

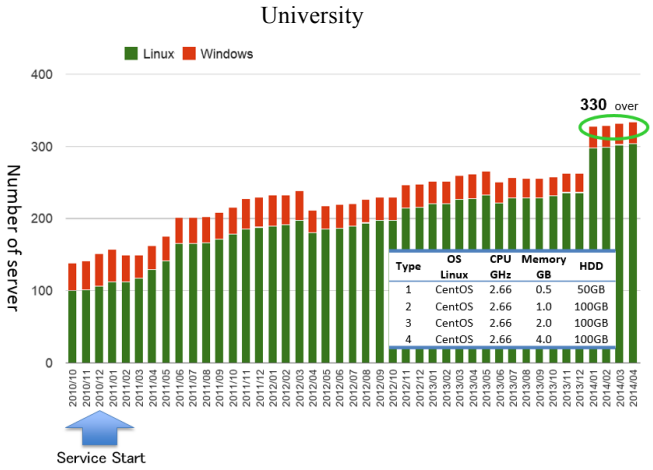

Figure 8 Number of the introductory transitions of VPS at Shizuoka University

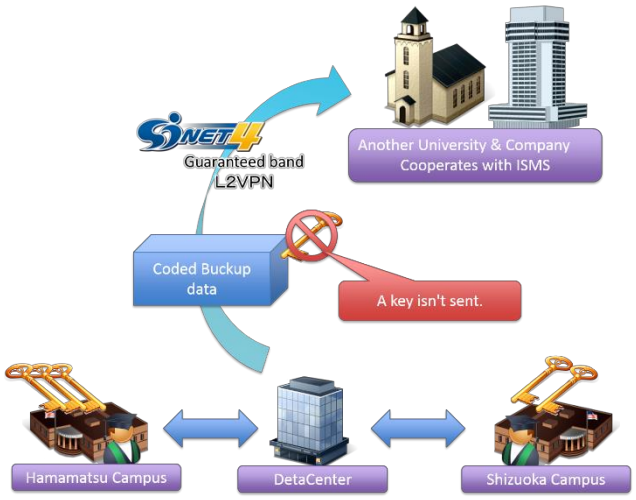

Figure 9 Backup flow for Shizuoka University using Sinet

Figure 9 shows that the flow of backup for Private Cloud Center. Shizuoka University joins Sinet4. So we use the environment for L2VPN, we send backup to a distant organization that they have ISMS cooperation.

For Public Cloud Center, we could not use the Sinet network.

So the problem to support backup is to take much time. And systems engineer pays attention to easily operation for backup. In the environment for big data, it takes much time to backup for lower network bandwidth. It's necessary to put a backup in the distant place. So we have to observe the network bandwidth.

We manage the backup management cycle in Figure 10.

1. Assessment for value of information: *h/day*.
2. Evaluate for availability of system: *p* and *S*.
3. Execute backup: calculate for the optimized backup duration and cost.
4. Assessment for backup: check possibility of restore and backup time. Return to 1.

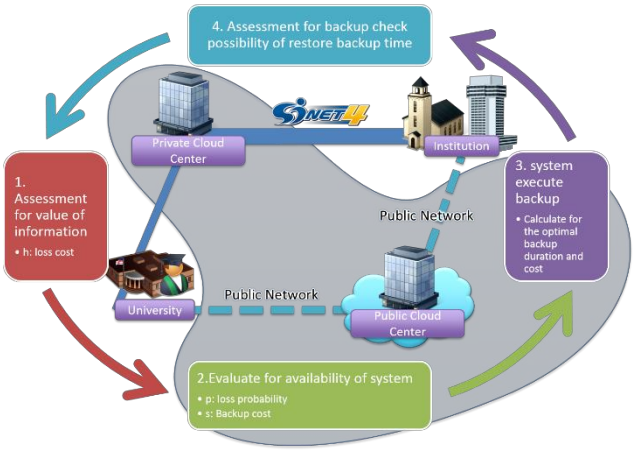

Figure 10 Backup management cycle

In each phase, a parameter is calculated from the present information. And we have the mathematical expression to optimize backup duration like (4) and (5). It is to cooperate with a backup management cycle, it is effective process. With the backup management cycle, we calculate the optimized backup period using geometric distribution with multi devices. We assume the parameter of backup as Table 6. So we get the result that the optimized backup period T is 47 day and the optimized backup cost is *13611.620565228011* yen. Figure 11 shows the transition for *C(T)* with *334* devices. As a result, we can get the policy for backup of 334 devices. We need to check if the backup is more effective and reassessment for value of information.

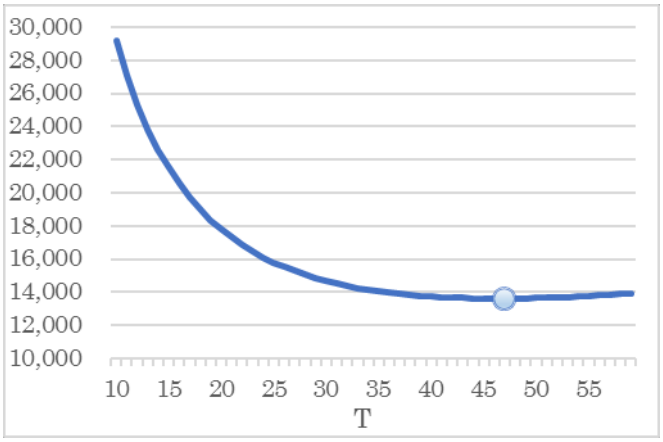

Figure 11 Transition for *C(T)* with *334* devices

Table 6 Parameter of backup for *334* VPS in Shizuoka University

System loss probability *p*	0.00001
System loss cost *h*/day	100,000 (yen)
System backup costs *S*	300,000 (yen)
Number of devices	334

5. Conclusion

In this study we calculated the backup interval in which costs can be kept to a minimum. Here, we performed calculations using the geometric distribution and Weibull distribution methods. But it is also possible to use other methods of probability distribution. By estimating parameters and calculating cost based on actual data, a more meaningful interval can be proposed.

We Propose the cost clearly that it requires in the most suitable backup as well as backup duration. And this model can correspond to multi devices. At Shizuoka University, 334 VPS are used. So it is very important to decide the backup interval, considering a balance of safety and cost. From this model, we get the policy for backup. We think it is important for university to execute backup with this management of backup.

Reference

1) Chen,P.M. ,Lee,E.K. ,Gibson,G.A. ,Katz,R.H.and Patterson,D.A.: RAID: High-Performance, Reliable Secondary Storage. ACM Computing Surveys, vol. 26, pp. 145-188.(1994
2) Cunhua Qian. Nakamura, S. and Nakagawa, T.: CUMULATIVE DAMAGE MODEL WITH TWO KINDS OF SHOCKS AND ITS APPLICATION TO THE BACKUP POLICY, Journal of the Operations Research Society of Japan 42(4
3) Ericsson AB.: Ericsson Mobility Report. EAB-14:061078 Uen, Revision A (2014.11
4) Mizuno.S, Nagata.M, Seki.M, Inoue.H, Hasegawa.T, Yamaki.N.: Constructing, evaluating, and applying an automatic verification system for Virtual Private Servers, Japan Industrial Management Association, Special English Issue,, Vol.64, NO.4E, pp.601-613.(2014
5) Sandeh,H. And Kawai,H.: AN OPTIMAL N-JOB BACKUP POLICY MAXIMIZING AVAILABILITY FOR A HARD COMPUTER DISK, Journal of the Operations Research Society of Japan 34(4
6) Seagate Technology,: If data disappears. Seagate Technology LLC. (2012
7) Toshio Nakagawa.: Maintenance Theory of Reliability Springer Series in Reliability Engineering Springer Science & Business Media, (2006