International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 5 (2018) pp. 2589-2594

Automate Backup Using Workgroup Distributed File System

Ravi Upra[1] and Roungsan Chaisricharoen[2]

School of Information Technology, Mae Fah Luang University, Chiang Rai, Thailand.
E-mail: [1]ravi@technocom.co.th, [2]roungsan.cha@mfu.ac.th

Abstract

Within the network of Small or Medium Businesses (SME), unused storage spaces inside a device or computer are scattered and available. The objective of this paper is to utilize those resources by pooling all available free spaces to form a personal cloud storage. Implementing a Peer-to-peer (P2P) for distributing and managing of information to create a reliable automate backup system. P2P, nodes to nodes is the key concept of this conceptual. Nodes are decentralize and are fully automate without administrators. This paper involves in conceptual steps from selecting available spaces, selecting reliable nodes, distributing information to selected nodes and restoring back the require information from client's nodes.

Keywords: Peer-to-Peer ,Distributed File System, Personal Cloud Backup System, Decentralize Manage, Storage, Restoring, nodes

INTRODUCTION

Having a good backup system is still a major problem for Small and Medium businesses (SME). The main source of problems is lacking budget and knowledge in implementing the backup system, such as investing in Network Attach Storage (NAS) or using schedule backup software, etc. When storage fails or data crisis occurs, most of the small businesses fail to have a decent backup of business's related documents such as financial accounting files, business database, customer records, customer credit records, etc. This could lead to enormous waste of time and money in recovering and rebuilding the system back to the previous state. This paper proposes a concept of using unused storage resource, which is scattered within the domain of small or medium businesses. This could benefit in saving of extra budget in spending on the backup system and protecting the privacy of information as information is still within the local network.

A survey was conducted in 2016 from 96 companies in the local Area [1]. 76% of the surveyed companies were worried about their data and more than 50% do not have a proper backup. They find it is too complicated to learn about the backup software. The company's owners are aware that more than 50% of storage resources are not being used. Lost data can lead to costly downtime, lower productivity, and also company's ability to compete in this digital era. A recent study from a security software company shows that small business owner pays little attention to data backup when compare to changing their password for security[2].Data breaches refer to a situation where sensitive or confidential data is lost, stolen or put at a risk. The recent study by Ponemon Institute [3] found that average cost of confidential data breaches per record increases nearly 23% from year 2013 from $145 to $154 in year 2015.

The objective of this paper is how to utilize unused storage resources that are scattered within the domain of small business areas and supports a good and automate backup system for small business areas using Peer-to-peer (P2P) automate backup, which support build-in function, such as compressing, decompressing, encryption and decryption

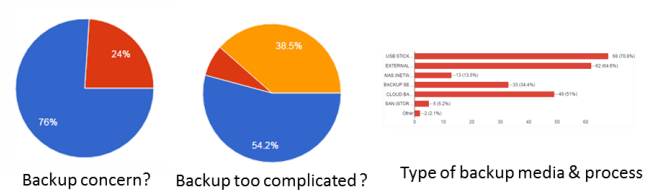

Figure 1. Survey of local small business in year 2016 [1]

BACKGROUND THEORY

P2P algorithms were introduced base on a concept of sharing resources among connected nodes. Most of the P2P algorithm adopts the use of Distributed Hash Table (DTH) [4], making it faster to routes from the source node to destination node for retrieving, updating, or adding resources such as files. The keys concept of P2P is Decentralize system, all nodes are self-administrator. Nodes need to be connected and registered, to become part of a P2P network. Fault-tolerance and redundancy is another key advantage for P2P. No single point of failure, sometimes nodes are unavailable or inactive, resources still could be retrieving from neighbour's node due to multiple and redundancy of resources. Algorithm related to P2P like Chord [5], Pastry [6], Gnutella[7], Napsters are a common algorithm and are applied in many P2P application like Bittorrent [8], Spotify, Napster, Utorrent, Tor, etc. This paper did focus on using some basic key concepts of P2P like decentralize of nodes, data replicas for fault-tolerance. Unlike using Chord [5] or Pastry [6] algorithm, this paper, nodes are selected base on the reliability of nodes, not by fastest route as in most P2P algorithm. Scope or range of network is within local network (Workgroup). For communication, the node will broadcast to all related nodes within domain instead of using a complex algorithm like chord, as not a large number in term of nodes or peers for Small and Medium Enterprise Businesses network.

Backing up sensitive or confidential data is a crucial step for small and medium businesses. Data loss may occur from failure of hardware, software, or malware, which could lead to

enormous spending for recovery of information. At First, we look at today's trends in the technology of backing up information. The easiest way is to use external storage devices, such as a USB stick or USB external hard drive. The problem with these types of devices is the tedious work of copying and pasting, and repeating it every day Small businesses often tend to bypass this process, which in turn leads to crisis. NAS (Network-Attached Storage) [9] could prove to be useful, but it is again too complicated for small businesses to learn or have extra budgets for hiring an implementer or administrator for maintaining the system. NAS could also be a large investment for small businesses. Another solution is using cloud backup[10] systems, such as Dropbox. Many companies offer a free trial with a little space for personal use. Storing information is convenient but vulnerable as well [11]. Businesses need to pay an extra cost for more spaces. The possibility of a cloud service provider is ceasing to exist. Like in the past, the Megaload website has just quit and ceased giving information. Another solution is SAN (Storage Area Network), which uses a high bandwidth network, such as fiber optics, connect to multiple servers and multiple storages. It utilizes redundancies in both hard drives and power supplies. Their price ranges from $100,000 to a million dollars. SAN requires a real specialist to install as well as maintain the system. SAN does not really fit well in a small business area due to large investments and knowledge required for maintaining the system. Compression can help in reducing the transfer time and is accomplished by looking for repeated patterns or predictability, or common areas of information. Compression can be dividing into two categories: Lossless and Lossy compression [12] Lossy compression is mostly using in multimedia sectors, such as digital photo, digital audio or digital video. This type of compression reduces the size of the original files and discards [13] unnecessary data while maintaining nearly the same quality of that digital photo or video. The problem with lossy compression is that when it decompresses, the recovered files are nearly original but not the same as the original. Lossless [14] compression, on the other hand, helps in reducing the original file or information, but at the same time, these files or information can be recovered to their original state without losing any information. The drawback of using lossless compression is its decompressing factor. It takes a longer time to compress as well as to decompress information back to their original. Examples of this are LZW, DEFLATE, BZIP2, LZMA, and LZO

There are two types of encryption available: Asymmetric encryption [15], where two keys are used, one for encryption (Public Key) and one for decryption (Private Key). Asymmetric encryption is widely used in protecting the communication of information from one end to the other end. This process is secure and fit for applications that need to transfer information back and forth privately. On the other hand, symmetric encryption [16] or a secret key is mostly used for encrypting files. Files will be encrypted and decrypt using one single key. There are many encryption algorithms in symmetric encryptions, such as Advanced Encryption Standard (AES) [17] encryption to protect data that is transferred onto other machines. AES [18] was introducing in January 1997, and The National Institute of Standards and Technology's (NIST) has decided to propose Rijndael as the Advanced Encryption Standard (AES). Other types of encryption are Twofish, symmetric, block cipher, Serpent, DES, and 3DES (Data Encryption Standard).

METHODOLOGY

An agent is a service that is installed on devices. The agent collects all the related information like device's on time, free storage spaces which will be used in the reliable calculation. Node can be of 3 Types

Client Node is a node that shares free spaces to other nodes.

Server node is a node that uses free spaces from other nodes to keep important information.

Hybrid Node is a node that plays both roles. Hybrid Node can be both, Server Node as well as Client Node

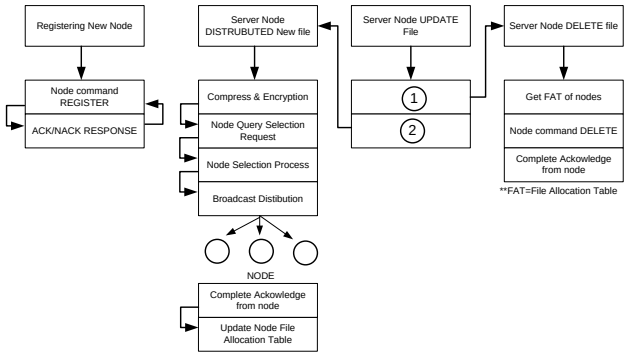

Figure 2. Displays all the related function for Peer-to-peer automate backup system

Attributes Definitions Collected By Agent

Active (A): denoted by A is the Node or device on/off status. Active is more concern on the average on time of the node. Active has more weighted than other attributes, as reliability will be based more on the device on time. Finding Active attribute has 2 parts.

1. Find the average on time of node of that particular day

At be the average On-time of the day at that specific time in percentage

f be a frequency in seconds that agent will need to be run

Tl be the total loop for a day that agent will have to achieve 100 Percent

$$\mathrm{Tl} = \frac{86000}{\mathrm{f}} \quad \text{where } 86000 \text{ is seconds in a day}$$

Tc be the total number of count at that specific time

$$\mathrm{At} = \left(\frac{100}{\mathrm{Tl}}\right) * \mathrm{Tc}$$

2. Recalculate the history average on time

A be the average On-time of the node in percentage or Active

Ag be the history average On-time of that node

D be the date different from Ag last date to At date

W1 be the weight of attribute A

$$A = \left(\frac{At + Ag}{D}\right) W1 \quad \text{Or}$$

$$\left(\frac{\left(\left(\left(\frac{100}{Tl}\right) * Tc\right) + Ag\right)}{D}\right) W1 \qquad (1)$$

Processor Power (P), Memory (M), Bandwidth (B), use the same calculation

P,M,B be the average free processing, memory, bandwidth remain of nodes.

Pt,Mt,Bt

be the free available percentage of processor, memory, bandwidth at that specific time.

Pg,Mg,Bg

be the history average processing power, memory, bandwidth remain of the Node.

W2,W3,W4

be the weight of attribute P,M,B

$$P = \left(\frac{(Pt + Pg)}{2}\right) W2 \qquad (2)$$

$$M = \left(\frac{(Mt + Mg)}{2}\right) W3 \qquad (3)$$

$$B = \left(\frac{(Bt + Bg)}{2}\right) W4 \qquad (4)$$

Score(Sc)

Score is the value that indicates the average score of particular node at a given time and is derived from

$$Sc = \frac{1 + 2 + 3 + 4}{\sum Wi} \quad \text{or} \quad \frac{A + P + M + B}{\sum Wi} \quad \text{or}$$

$$\frac{\left(\left(\frac{At + Ag}{D}\right) W1\right) + \left(\left(\frac{(Pt + Pg)}{2}\right) W2\right) + \left(\left(\frac{(Mt + Mg)}{2}\right) W3\right) + \left(\left(\frac{(Bt + Bg)}{2}\right) W4\right)}{\sum Wi} \qquad (5)$$

∑wi = Total weighted mean which is 9 in which we focus more on on- time as well as Bandwidth .W1 = 3; W2 = 2; W3=1; W4=3

Storage (St)

Be the remaining free spaces in percentage and Ss are the real size in byte of a node.

Selection of node for distribution-I

When the Server Node is ready to distribute files to other nodes, the Server Node will broadcast Request Query Selection Command (SQSC). It waits for a certain period to get all the feedback from all the active client nodes. It then creates a table for decision-making based on the information replied. The table will be sorted based on the Score (Sc) value received from nodes

Sc is the score value of the Client node

St is the remaining free spaces in percentage

Ss is the real size in bytes of remaining free space of a node

Table 1. Sorted table based on the score recieved from client nodes

Node Name	IP Address	Response Time (Rt)	Score (Sc)	St%	Ss (Kb)
Sak	192.168.0.78	7	67	18	23773
Ravi	192.168.0.12	5	68	70	8192
Jeff	192.168.0.99	7	55	67	4101005
Upra	192.168.0.45	4	53	25	1205316
Thida	192.168.0.13	4	51	40	851
Phot	192.168.0.10	8	42	39	422202
John	192.168.0.15	4	32	89	325736

Selection of node for distribution-II

At first, nodes which do have less remaining spaces than the required spaces will be deducted from the list. Later, adding weighted means to focus on certain attributes based on reliability of information and then re-sort the list again. The required attributes are as follow

Current mean Score (Sm)

The more score means then more secure of information as more weight were added to On/Off state (Active attribute) .

Percentage Of files transferred to those nodes base on past history (Fp)

The more files being transferred to that specific node means the more risk. Files should be equally distributed to many nodes as possible. A Percentage of Files Transferred to a given node can

be calculated by

Ft be the total file size transferred out

Fn be the total file size transferred to a given node

Fp total file transferred to a given node in percentage

$$Fp = \left(\frac{100}{Ft}\right) * Fn \quad (6)$$

Ifp = 100 − Fp (the more Fp mean the less Ifp)

Remaining Free Space (St)

The more percentage means more room spaces for files to be distributed.

Response means time of Node (Rn)

Faster response time can ensure that node is closer or having a good bandwidth.

Rf be the fastest response time

Rt be the response time of Reliability query of the node

$$Rn = \left(\frac{100}{Rf}\right) * Rt \quad (7)$$

Final Score (Fs)

All the node will be sorted base on their final score

$$s = \frac{(Sm * W1) + (Ifp * W2) + (St * W3) + (Rn * W4)}{\sum Wi} \text{ or}$$

$$s = \frac{(Sm * W1) + \left(\left(100 - \left(\left(\frac{100}{Ft}\right) Fn\right)\right) W2\right) + (St * W3) + (Rn * W4)}{\sum Wi} \quad (8)$$

∑wi = Total weighted mean which is 9 and W1 = 3 ; W2 = 3 ;W3=1; W4=2

Table 2. Show the sample sorted node based from Final Score (Fs)

Node Name	IP Address	Response Time(Rt)	Score (Sc)	St%	Ss (Kb)
Sak	192.168.0.78	7	67	18	23773
Ravi	192.168.0.12	5	68	70	8192
Jeff	192.168.0.99	7	55	67	4101005
Upra	192.168.0.45	4	53	25	1205316
Thida	192.168.0.13	4	51	40	851
Phot	192.168.0.10	8	42	39	422202
John	192.168.0.15	4	32	89	325736

Data reliability

History of distributed data to a given node

This history is used in the above equation Percentage Of files transferred to those nodes base on past history (Fp) (2) to make sure not to much of the information can be distributed to a given node. By using weighted mean and giving less weight for this attribute so that the node that holds more data will be less likely to be chosen.

Replication for data reliability

This model will be using data replication-based for data reliability, 3 Replica is to be the basic number for data reliability. The number can be more or less base on the total free spaces collected from all the nodes as well as the Self-Reliability of all the Server-Nodes.

Server-Node Self-Reliability

If more devices tend to be more in off state or not reliable, it may be good to increase replica set from 3 to 4, but this all depends on the remaining spaces. Current Average Server-Node Reliability (Cg)

$Cg = \sum_{i=1}^{n} Smi$, where Smi is the current mean score value of that server node form 1 to n node.

History Average Server-Node Reliability (Hg).If Cg < Hg by 50% then consider adding replica.

EXPERIMENT

Two Simulations were setup based on the fixed attributes and variable attributes. Fixed attributes were assigned the best attribute's value, such as the response time < 4ms or available bandwidth up to 90%.

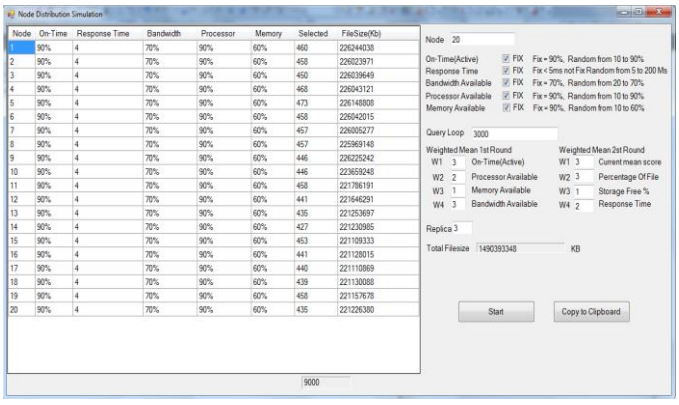

Figure 3. Node Distribution simulation software

From Fig. 4. The left side graph shown by the file size, and fix attributes variable. File is distribute to all the nodes with the same attributes variable value and the right-hand side graph

uses the same pattern except the attributes changes in the real environment, such as the processing power, bandwidth, active time, etc. The files are equally distributed to all the nodes (blue line) when the attributes are the same, and the nodes are also equally selected. When the attributes changes in a real environment, such as the processing power, the bandwidth, active time, node selection also changes. In a real environment, selection is based more on reliability of information on the client nodes not on the fastest route.

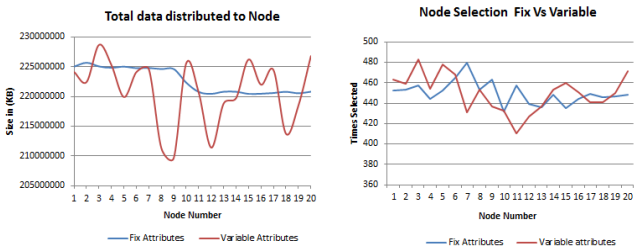

Figure 4. Fix Attributes Vs Variable Attributes Node Selection

Another Two Simulations were setup based on fix attributes and variable attributes. Fix attributes are assign a best attribute's value such as response time < 4ms or available bandwidth up to 90%.Left side Graph 1, shows by file size distributed to all node and right hand side shows times of node being selected. When all the attributes are the same (blue line), files are equally distributed to all the nodes and nodes are equally selected. Where else when attributes changes in real environment like processing power, bandwidth, active time, node selection also changes. Selection will be based on reliability of data on nodes

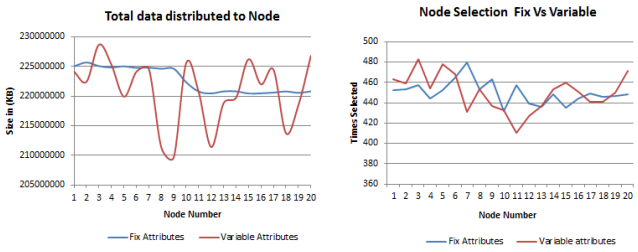

Figure 5. Fix Attributes Vs Variable Attributes Node Selection

Another experiment, files were distributed back and forth between nodes to get average transfer speed, total size of 25.3 GB. Two types of network speed, 100 Mb and 1000 Mb. From Graph 2: Network Speed 1000 Mb can perform better than 100 Mb around 5-8 times faster depend on network traffic. Transferring speed doesn't have much effect if file size is lower than 100 MB. If more than 100 M , much different in transfer time can be notice between these two networks

Figure 6. File size compare to Transfer time and also Transfer Rate

CONCLUSION

Resources tend to expand its capacity year by year due to advance in technology. This paper proposes conceptual peer-to-peer distributed file system for small and home offices, which could prove to be useful way to pool and utilize left over storage resources.. WDFS use a batch file process, running in background mode in which it complete task one by one and it is not bound to timing. This paper only introduced the initial phase for selecting reliable node before distributing data to that particular node. All the remaining and complete work will be prepared and publish in the future work.

REFERENCES

[1] Upra, R. (2016). Survey Local Small and Medium Business 8-12 October 2016. Resource document. https://tinyurl.com/y8z6zhue accessed 06 January 2018.

[2] Eddy, N. (2013). Small Businesses Unprepared for Data Loss, Lack Backup Policies, eWeek, November 14. Resource document. Eweek. http://www.eweek.com/security/small-businesses-unprepared-for-data-loss-lack-backup-policies accessed 06 January 2018.

[3] LLC, Benchmark research sponsored by IBM Independently conducted by Ponemon Institute (2015). Cost of Data Breach Study:Global Analysis. Resource document. https://www.ibm.com/security/data-breach accessed 06 January 2018.

[4] Stoica, I.,Morris, R.,Jiben-Nowell, D.,Kanger, D R., Kasshoek M F., Dabek, F. & Balakrishnan, H. Chord:(2001) A Scalable Peer-to-peer Lookup Protocol for Internet Applications. Proceedings of the 2001 SIGCOMM 01' conference on Applications, technologies, architectures, and protocols for computer communications.149-160

[5] Rowstron, A. & Druschel, P.(2001). Pastry: Scalable, Decentralized Object Location, and Routing for Large-Scale Peer-to-Peer Systems. IFIP/ACM International Conference on Distributed Systems Platforms and Open Distributed Processing. 329-350.

[6] Matei, R., Lamnitchi, A. & Foster,P. (2002). Mapping the Gnutella Network. IEEE Internet Computing 6(1), 50-57.

[7] Parameswaran, M., Susarla, A. & Whinston, A.B. (2001). P2P networking: an information sharing alternative. Computer, the flagship publication of the IEEE Computer Society,34 (7),31-38.

[8] Guo, L., Chen, S. & Xiao, Z. (2007). A performance study of BitTorrent-like peer-to-peer systems. IEEE Journal on Selected Areas in Communications.25(1).

[9] Katz, R.H. (1992). Network-Attached Storage Systems .Proceedings Scalable High Performance Computing Conference SHPCC-92.

[10] Qiu, S., Zhou, J. & Yang, T. (2013). Versioned File Backup and Synchronization for Storage Clouds. 2013 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing

[11] *Katal, A., Gupta, N., Sharma, S. (2012).* Information Storage on the Cloud: 2012 Students Conference on A Survey of. Engineering and Systems (SCES).

[12] Pinho, A. & Neves, A. (2006). Lossy-to-lossless Compression of Images Based On BinaryTree Decomposition. 2006 International Conference on Image Processing.

[13] Rufai, M.A., Anbarjafari, G., Demirel, H. (2013). Lossy medical image compression using Huffman coding and singular value decomposition. 2013 21st Signal Processing and Communications Applications Conference (SIU)

[14] Robert, L. & Nadarajan,R. (2009). Simple lossless preprocessing algorithms for text compression. IET Software, 3(1), 37-45.

[15] Fanfara.,P., Danková.,E. & Dufala., M. (2012). Usage of asymmetric encryption algorithms to enhance the security of sensitive data in secure communication. 2012 IEEE 10th International Symposium on Applied Machine Intelligence and Informatics (SAMI).

[16] Bharadwaj, Y. & Chakraverty, S.(2013). A design pattern for symmetric encryption. 2013 International Conference on Control Computing Communication & Materials (ICCCCM).

[17] Guo,G.,Qian, Q. & Zhang, R. (2015). Different Implementations of AES Cryptographic Algorithm. 2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems.

[18] Lu, C. &Tseng, S.(2002). Integrated design of AES (Advanced Encryption Standard) encrypter and decrypter. Proceedings IEEE International Conference on Application- Specific Systems, Architectures, and Processors.