INTERNATIONAL JOURNAL OF SOCIAL SCIENCE, INNOVATION AND EDUCATIONAL TECHNOLOGIES (ONLINE) - ISSN: 2717-7130

Vol:1, Issue: 4 pp: 588-594

JEL Codes: O3, O33

AMIN A., (2020). "**GS RichCopy 360: New Multi-Threaded Technology for Data Migration**", Vol: 1 Issue: 4 pp: 588-594

Keywords: data migration, TCPcoy, multi-threaded technology

Article Type Research Article

GS RichCopy 360: New Multi-Threaded Technology for Data

Arrived Date	Accepted Date	Published Date
10.09.2020	18.09.2020	31.10.2020

Ahmed Amin[*]

ABSTRACT

The need to copy billions of files and folders from one server to another is continiusly increasing in the world of Internet of Things (IoT). IT administrators are facing this challenge in quartely basis and normally, they would either write scripts or purchase migration software that can be super expensive due to premium features. This paper represents a new solution to transfer files from one server to another with reduced computational time and cost. The developed software GS RichCopy 360 performs the data transfer in a mutlithreaded way. Two versions are available GS RichCopy 360 Standard and GS RichCopy 360 Enterprise. The used methods and the key features of the two versions are detailed in the following sections.

INTRODUCTION

Data migration is decribed to be the process to copy data between computers, storage devices, servers or computing environments. Organisations consider data migration in several cases including 1) server maintenance, replacement or upgrade 2) database relocation 3) website maintenance 4) storage devices upgrades. With the increase of the amount of data, called Big Data, companies are relying more and more on cloud-based storage and applications infrastructure.

(Taylor.C, 2019) identifies three major risks related to data migration: Data loss, issues related to comapatibily and several impacts on the business (missed deadlines, exceeded budgets).

The main challenge for entreprises is to ensure the data transfer in a secure and cost-effective manner using an efficient method (IBM cloud education, 2019). The diversity of the data, applications and their corresponding requirements has been continuously motivating well known IT companies to develop file copy tools.

Robocopy is a Microsoft product known as Robust File Copy for Windows. It was created to support Information Technology professionals to perform robust data migrations. The file copy tool has the advantages of copying large data sets and logging the complete tasks and encountred errors but lacks on flexibility and adds significant overhead when managing multiple instances. (Lavelle, Konrad, 2007).

Microsoft RichCopy is a successor of Robocopy with a friendly Graphic User Interface GUI, faster copying process using multi-thread technology. This tool fails to copy long file path. RichCopy is discontinued.

GuruSquad is commited to providing costumers worldwide with innovative solutions for data replication. In this paper we present the key features of two powerful software GS RichCopy 360 Standard and GS RichCopy 360 Enterprise created by GuruSquad developers. We describe in the next section different existing data transfer methods.

Overview

[*] itahmed@hotmail.com/ USA

According to the literature, a successful data migration requires a flexible plan designed, executed and monitored to support the change (Iqbal, Colomo-Palacios, 2019). The assessment of data quality is fundamental step prior to planning the data migration in order to select the right migration technique (Oracle.2011). The diversity and volume of the data define the complexity of the process and additional steps including transformation, compression and synchronization are required.

Extract Transform Load (ETL) is the most used technique by IT Organizations for its abilities to extract, clean the raw data, correct errors/ missing data, transfom raw data into the appropriate formate or structure and load it into the target destination, (Xaviera.C; Moreira.F,2013)

(Souibguia.M; Atiguib.F; Zammalia.S; Cherfib.S; Ben Yahia.S, 2019) described data quality challenges when using ETL process: naming conflicts, real-time detection and duplication. The authors highlited the importance of developing a solution that handles Big Data complexity related to volume, variety and velocity.

Business users and leading IT organizations are more and more relying on Cloud-based solution to store massive data (Nahar,P; Joshi,A ;Saupp,A. 2012). Cloud computing offers four main services: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS) and Function as a Service (FaaS). This fact allows companies select the appropriate service meeting their challenges and pay only for that service. The effiency and low cost of the listed services encouraged companies to move, not only data, but also their development environments to the Cloud. Data loss or unavailability remain a major threat to the popularity of a cloud-based solutions. In order to provide data durability and availability, Amazon's S3, Google's Bigtable implement data replication (Abadi, 2009). Data Replication consists on generating an identical copy of the source data and store it at a single or various destinations or sites (Abawajy, J; Deris, M. 2014). Replication is considered to be the key factor in improving the availability of data stored at multiple sites. This gives the user the ability to access data even in the event of some sites failure (Son, 1987).

Proposed Solution

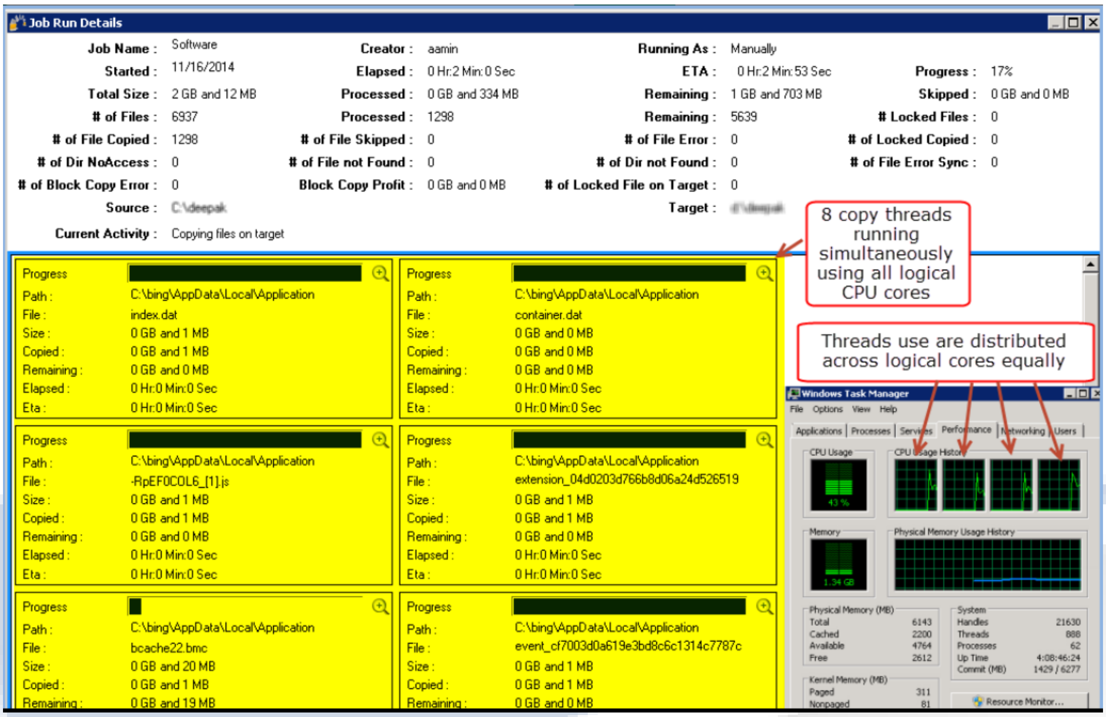
Purpose of the Work

GS RichCopy 360 solutions have been developed by GuruSquad Company in order to solve IT problems related to data replication, migration, and synchronization. The proposed solution transfers data between servers at reduced computational time and cost (GuruSquad, 2020) using a unique multi-threaded technology.

Used Method

The GS RichCopy 360 utilises a multi-Threading technology, resulting in 100% parallel computing operation. The presented patent-pending technology distributes multiple threads through virtual cores providing a robust level of maximum performance during file copy operations and reduced time. The maximum number of threads is set to 256 threads. The recomended maximum number of used threads is double the number of the logical cores on the system. The default number of threads set by GuruSquad is set to 4.

Figure1 shows an example of 8 copy threads running simultaneously using all logical CPU cores.



Job Run Details

Job Name :	Software	Creator :	aamin	Running As :	Manually		
Started :	11/16/2014	Elapsed :	0 Hr:2 Min:0 Sec	ETA :	0 Hr:2 Min:53 Sec	Progress :	17%
Total Size :	2 GB and 12 MB	Processed :	0 GB and 334 MB	Remaining :	1 GB and 703 MB	Skipped :	0 GB and 0 MB
# of Files :	6937	Processed :	1298	Remaining :	5639	# Locked Files :	0
# of File Copied :	1298	# of File Skipped :	0	# of File Error :	0	# of Locked Copied :	0
# of Dir NoAccess :	0	# of File not Found :	0	# of Dir not Found :	0	# of File Error Sync :	0
# of Block Copy Error :	0	Block Copy Profit :	0 GB and 0 MB	# of Locked File on Target :	0		
Source :	[redacted]			Target :	[redacted]		
Current Activity :	Copying files on target						

Progress
Path : C:\bing\AppData\Local\Application
File : index.dat
Size : 0 GB and 1 MB
Copied : 0 GB and 1 MB
Remaining : 0 GB and 0 MB
Elapsed : 0 Hr:0 Min:0 Sec
Eta : 0 Hr:0 Min:0 Sec

Progress
Path : C:\bing\AppData\Local\Application
File : container.dat
Size : 0 GB and 0 MB
Copied : 0 GB and 0 MB
Remaining : 0 GB and 0 MB
Elapsed : 0 Hr:0 Min:0 Sec
Eta : 0 Hr:0 Min:0 Sec

Progress
Path : C:\bing\AppData\Local\Application
File : -RpEF0COL6_[1].js
Size : 0 GB and 1 MB
Copied : 0 GB and 1 MB
Remaining : 0 GB and 0 MB
Elapsed : 0 Hr:0 Min:0 Sec
Eta : 0 Hr:0 Min:0 Sec

Progress
Path : C:\bing\AppData\Local\Application
File : extension_04d0203d766b8d06a24d526519
Size : 0 GB and 1 MB
Copied : 0 GB and 1 MB
Remaining : 0 GB and 0 MB
Elapsed : 0 Hr:0 Min:0 Sec
Eta : 0 Hr:0 Min:0 Sec

Progress
Path : C:\bing\AppData\Local\Application
File : bcache22.bmc
Size : 0 GB and 20 MB
Copied : 0 GB and 1 MB
Remaining : 0 GB and 19 MB

Progress
Path : C:\bing\AppData\Local\Application
File : event_cf7003d0a619e3bd8c6c1314c7787c
Size : 0 GB and 1 MB
Copied : 0 GB and 1 MB
Remaining : 0 GB and 0 MB

8 copy threads running simultaneously using all logical CPU cores

Threads use are distributed across logical cores equally

Windows Task Manager

Physical Memory (MB)	
Total	6143
Cached	2200
Available	4764
Free	2612

Kernel Memory (MB)	
Paged	311
Nonpaged	81

System	
Handles	21630
Threads	888
Processes	62
Up Time	4:08:46:24
Commit (MB)	1429 / 6277

Figure.1. Example of 8 Copy Threads Running Simulataneously Using All Logical CPU Cores.

The enterprise version GS RichCopy 360 Enterprise offers a copy method called TCP Copy. This method requires an agent to be installed on the destination machine (has to be windows) then the client from the source connects to the destination over a single TCP port.

Given the advantage of having a client on the source side and an agent on the destination, comparison of files can be reduced by as much as 95% as the client on the source no longer has to reach remotely to the destination to see if a file exists, outdated, or identical as the remote agent on the destination would handle such calls. Given such processes are all treated locally, the ability to compare hundreds of millions of files dropped by as much as 95% when the copy job traversed a WAN connection with higer latency.

RTA (Remote Transfer Agent) identifies jobs by their job serial number that is issued at time of creation on the client side. Once the client establishes a connection to the RTA as the job is triggered, that job is then tight down to that specific client with that serial number. This is crucial as even if a malicious user gets a hold of that job serial number, they still would not be able to trick the RTA by sending data from another client.

A single RTA server can simultaneously serve multiple incoming jobs from the same client and others. It is estimated a single RTA agent can handle more than 5,000 jobs simultaneously given it as has the necessary compute resources and a robust storage subsystem.

Configuring the Remote TCP Copy function is a simple 3-step process:

- Configure your firewall to accommodate the TCP port (8008 by default)
- Install the Remote Transfer Agent (RTA) on the targeted cloud machine
- Quickly configure your network client source machine and RT agent

A typical configuration of Remote TCP Copy function is shown in figure2.

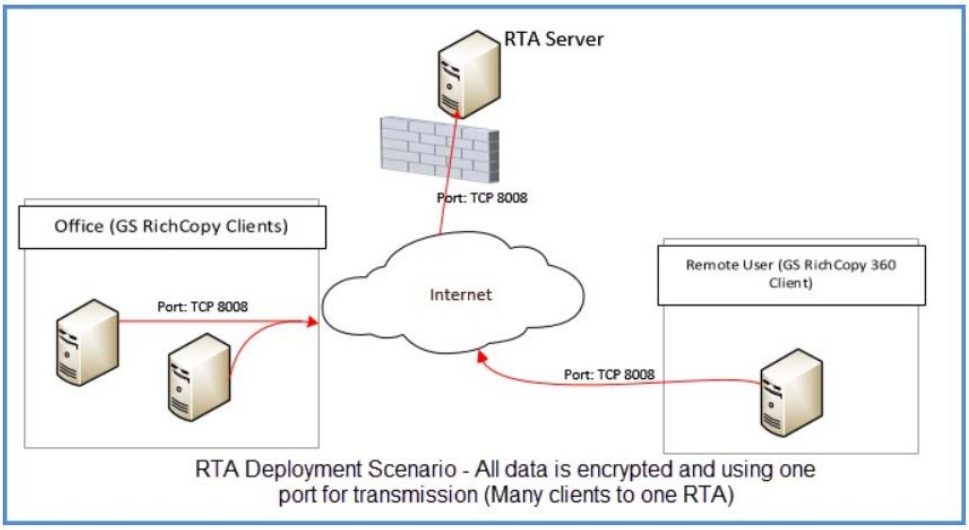

RTA Server
Port: TCP 8008
Office (GS RichCopy Clients)
Port: TCP 8008
Internet
Remote User (GS RichCopy 360 Client)
Port: TCP 8008
RTA Deployment Scenario - All data is encrypted and using one port for transmission (Many clients to one RTA)

Figure.2. Typical configuration of Remote TCP Copy function.

The two versions of GS RichCopy 360 provide a robust level of maximum performance during file copy operations.

We describe in the following sections the key features of the two solutions.

GS RichCopy 360 Performance

GS RichCopy 360 software offers the following advantages

1) The distribution of every thread onto the corresponding core allows the user monitoring what every thread is copying. This makes indentifying and resolving any problems during migration process easy. Existing multithreaded copiers: Microsoft Robocopy, PeerSync by Peer Software do not show their threads as all threads are managed by core 0 but copied via other cores and CPU contention is visible.

2) Can copy permissions, support copying files and folders that are longer than 260 characters long.
If the count of the characters of the path and the file name are higher than 260 characters long, windows by default fails. Most data copiers, including Rich Copy, are not able to copy it. Some copy programs can copy them under an enterprise grade can. The cost of the software in this case will be higher.

3) Can copy open files.

4) Can run as a service and according to schedule: The software runs in the background.

5) Can set the folder timestamp on the destination folders to reflect the same dates as the source timestamps. Typically, with most copying software applications, including Robocopy and RichCopy from Microsoft, the folders reflect the time stamp of the copying day, (Smith.R,2014).

6) All errors are aggregated and listed in one log.

Entreprise Version

GS RichCopy 360 Enterprise Backup and Replication Software is a powerful version extension of GS RichCopy 360 Standard. The latter software has been designed to meet today's entrprises requirements: automation, elasticity and security. The advantages of the advanced version are:

a) Compress files while in transit: Some files can compress as much as 95% which makes the rate of copying much higher. Our software uses different types of compression algorithm: lz4 and zstd. The two algorithms are the fastest in their inflating and deflating ratio (compress and decompress) and they provide great compression ratio, (Handte,F; Collet,Y; Terrell,N.2018).

GS RichCopy 360 Enterprise detects the file and based on defined types, chooses the better compression ratio. Not many applications offer compression as a client\agent on the source (compress before sending) and destination (decompress after receiving) is required. We quote PeerSync by Peer Software and DoubleTake by Carbonite. This is mainly critical copying over WAN, VPN, the internet, or slow unreliable links as those connections are not typically fast and very crucial to send as little traffic as possible to cut down on transfer time and not to impact other services looking to use those links.

b) Copying can be encrypted using Advanced Encryption Standard AES 128/256 bit encryption while in transit. This offers the possibility to copy across the internet and over untrusted networks. The National Institute of Standards and Technology (NIST) depicts three AES implementations: 128-bit, 192-bit, and 256-bit. Each type uses 128-bit blocks. The difference lies in the length of the key. As the longest, the 256-bit key provides the strongest level of encryption (Keshav KumarK. R. RamkumarAmanpreet Kaur, 2020). With a 256-bit key, a hacker needs to try 2256 different combinations to ensure the right one is included. This number is astronomically large, landing at 78 digits total. It is exponentially greater than the number of atoms in the observable universe. Government and financial institutions use 128- or 256-bit encryption for secret information (Franklin, R. 2020).

c) The software requires a single TCP port. This simplifies tasks dealing with firewalls. The TCP port 8008 is set by default and customer can change the port to any other port from within the GUI.
A normal copying operation between two Microsoft Windows machine use several ports some are deemed as senisitve ports to support SMB protocol requirements: ports 137-139 or TCP 445. These ports should not be exposed to untrusted networks due to vulrnabilities. Given SMB is the most common way to copy data between windows machines but yet requires sensitive ports to be open. It makes it difficult to do any data transfers if the traffic has to go across untrusted networks or the internet.

Using TCP Copy, data can be transferred using a different port which is less vulnrable, traffic can be encrypted using AES 128\256. This makes it not only easy to pass through untrusted networks, but also practical and easier to deploy due to less firewall restricitons. Coupled with the encryption option, transfers across the internet are also possible.

d) Existing copying software (Drive, Google teams, drop box) use SMB protocol to copy files across WAN connections or the internet (cloud.. etc). The protocol has a lot of issues with latency and WAN connections during copying and comparing files. TCP Copy performs the task in a different manner. The two agents, the first installed in source and the second in the destination compare files locally, then they communicate and report the changed files. This technique cuts down the comparison rate by as much as almost 90% at a high latency network.

e) The ablity to copy to several cloud providers including Azure, Office 365, AWS S3 and Google.

f) Reuced computational time: One of the main advantages of TCP copy is the ability to copy millions of files 80% faster than cited copy tools. TCP copy can lump up small files and send them as one chunk and have the agent spread them out. Copying 1,000,000 files , total the size of 1GB , from one server to another by a multi-thread tool over SMB lasts 8.5 hours (average time).. TCP Copy performs the same task in 45minutes (average time).

Guru Squad provide API access. The solution can be automated in case an entreprise decides to integrate their own systems to create copy jobs on the fly.

Migration of files from one server to another can be super expensive due to premium features. Only enterprises could afford them. GuruSquad created this new tool and made it sell in the range from $50 to $130 per license. Existing solutions in the market charge $2000 to $7000 for a license. This way it is affordable to businesses of all sizes.

CONCLUSIONS AND RECOMMENDATIONS

The aim of this paper is to represent a new solution to transfer files from one server to another with reduced computational time and cost. The files are copied on a multithreaded way that enables the user to monitor the data present on each thread during the copying process. The data transfer could be performed to run as a service and according to a defined schedule. The developed software copies open files and has priemum features copying copying NTFS permissions, support copying files and folders that are longer than 260 characters long. The destination folders timestamp reflect the same dates as the source timestamps. The enterprise version offers a copy method called TCP Copy that has the ability to copy millions of files 80% faster than existing copy tools.

REFERENCES

Abadi, D. (2009) Data Management in the Cloud: Limitations and Opportunities, IEEE Special Issue on Data Management on Cloud Computing Platforms (32):3-12

Abawajy, J., Deris,.M. (2014) Data Replication Approach with Consistency Guarantee for Data Grid, IEEE Transactions on Computers (63): 2975 – 2987.

https://www.ibm.com/cloud/learn/datamigration#:~:text=Data%20migration%20is%20the%20process,consolidating%20or%20decommissioning%20data%20center.

Franklin, R. (2020). AES vs. RSA Encryption: What Are the Differences? https://www.precisely.com/blog/data-security/aes-vs-rsa-encryption-differences

GuruSquad (2020) https://www.gurusquad.com

Handte, F, Collet Y., Terrell, N. (2018) 5 ways Facebook improved compression at scale with Zstandard, CORE Data, https://engineering.fb.com/core-data/zstandard/

IBM cloud education (2019)

Iqbal, A., Colomo-Palacios, R. (2019) Key Opportunities and Challenges of Data Migration in Cloud: Results from a Multivocal Literature Review, Procedia Computer Science 164.

Kumar, K., K. R. Ramkumar and Kaur, A. (2020) A lightweight AES algorithm implementation for encrypting voice messages using field programmable gate arrays, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.08.005

Lavelle, C., Konrad, A. (2007) FriendlyRoboCopy: A GUI to RoboCopy for computer forensic investigators, digital investigation 4.

Nahar, P., Joshi, A. , Saupp, A. (2012) Data Migration Using Active Cloud Engine, 2012 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)

Oracle (2011) Successful Data Migration, https://www.oracle.com/technetwork/middleware/oedq/successful-data-migration-wp-1555708.pdf

Smith, R. (2014) How Can I Copy Files and Preserve Date Timestamps?, https://petri.com/copy-files-preserve-timestamp

Son, S. H. (1987) Synchronization of replicated data in distributed systems, Information Systems (12): 191-202

Souibguia, M., Atiguib, F. Zammalia, S., Cherfib, S.& Ben Yahia, S. (2019) Data quality in ETL process: A preliminary study, Procedia Computer Science 159 :676–687

Taylor, C. (2019) Data Migration: The Strategy to Succeed, *https://www.enterprisestorageforum.com/storage-management/data-migration.html*

Xaviera, C., Moreira, F. (2013) Agile ETL, Procedia Technology 9: 381 – 387