Saint Petersburg State University of Information Technologies, Mechanics and Optics

Department of Telecommunication Systems

Implementing Green IT approach for transferring Big Data over Parallel Data Link

Co-funded by the Erasmus+ Programme of the European Union

Name : Stefanos Georgiou

Supervisors: Andrey Y. Shevel
Eric Rondeau

17 of June 2015

Presentation GuideLine

- Introduction
- Goal of Thesis
- The importance of this work
- Testbed implementation
- Cloud infrastructure
- OpenStack Virtual Machines
- Testing Utilities
- Scripts and their purpose
- Scenarios
- Results
- Conclusion
- Future Work

INTRODUCTION

- Basic concept of the this work is to transfer Big Data over parallel data links.
- Big Data is large or complex amount of dataset which can not be processed by traditional processing applications.
- "Triple V": Velocity, Volume, and Variety.
- Big Data is not a small field of studies, it consists by different aspects like: store, analyze, transfer, preserve, capture, visualize, and etc.

Goal of Thesis (1)

- Assist ITMO research team with their project (to transfer Big Data by using parallel data links with SDN Openflow approach) and apply Green IT methods for the data transfer system.

- Main task is to compare existing data transfer applications in case to verify which results the highest data transfer speed in which occasions and explain the reasons.

- Essential information had to be stored and preserved for future analysis.

Goal of Thesis (2)

- Create scripts which will allow users to repeat the exact same experiments.
- Gather big amount of information about the experiments and all the used parameters.
- Prepare a complete platform which can be used in the future for different data transfer applications for further research.
- Testing on virtual environment and real network.

The Importance of this Work

- Data transfer applications can transfer huge amounts of datasets over long period of time, that is a reason why Green IT approach have to be taken into account.

- It focus on Cloud computing which is an emerge aspect in Computer Science and lot of companies are using virtual environment.

- Many researches perform transferring using data transfer applications (utilities) but not many details where given. (system information, dataset, and etc)

TestBed Implementation (1)

- TestBed is a powerful platform to test our scientific theories (ITMO stor.naulinux.ru)

Hardware Type	CPU	Main Memory (RAM)	Hard Disk Drive	Operating System
Server 1 (SV)	2 x Intel 6 Core Xeon E5-2640v2 @2.5GHz	64 GB DDR3		Scientific Linux CE 6.5
Server 2 (SM)	2 x Intel 4 Core Xeon E5-2609v2 @2.4GHz	32 GB DDR3	4 x SATA 500GB raid 5 1000 GB	Scientific Linux CE 6.5
Storage Area Network			HP 16 HDD SAS – 450GB	

TestBed Implementation (2)

- Server on Petersburg Nuclear Physics Institute

Hardware Type	CPU	Main Memory (RAM)	Hard Disk Drive	Operating System
Server	16 x Intel(R) Xeon(R) E5-2650v2 @2.6GHz	99 GB DDR3	RAID6 100TB	Scientific Linux CE 6.5

Cloud Infrastructure

- OpenStack version Icehouse is running on the testbed servers.
- Free of charge utility.
- User friendly GUI to manage users Virtual Machines, the “Dashboard”.
- Offers the possibility to share among user the server resource and deploy number of different Virtual Machines.
- Uses less hardware.

OpenStack Virtual Machines

- Tests on Virtual Environment (tests on same server)

Instance Name	Location	VCPUs	RAM size	HDD size
Sender	ITMO	8	16 GB	160 GB
Receiver	ITMO	8	16 GB	160 GB

- Tests on Real Network (40km distance through public network)

Instance Name	Location	VCPUs	RAM size	HDD size
Sender	ITMO	8	16 GB	160 GB
Receiver	PNPI	8	16 GB	160 GB

All virtual machines were tuned from Energy Science Networks website – Linux Tuning

Testing Utilities

- Fast Data Transfer (FDT) – Written in Java, can be used as a server/client application.
- BBCP – Written in C, very easy usage, no need for server or daemon process.
- BBFTP – Written in C, based on client/server architecture, efficient for large files.
- GridFTP (Globus toolkit) – Written in both Java and C, share users computer resources.
- File Transfer Service (FTS3) – Written in C, C++ and bash, used as a web service, transfer scheduling.

Why we used these Utilities?

- Common features
  - Multi-streams transfer
  - User can change the TCP window size
  - Tune I/O buffer size
  - Encrypted authentication
  - Open-source data transfer applications

Why Using Script?

- Number of different scripts were created using BASH to execute the datasets transfer using different utilities. Can be found at https://github.com/itmo-infocom/BigData
- Brief description is given also how to use the scripts.
- Faster to executed multiple and different scenarios using the scripts instead of executing them one by one.
- Opportunity to capture multiple information about the transfer and the system parameters and store them for further analysis.

Scripts and they Purposes (1)

- Test-data: script responsible to generate a directory with binary files of random length. Reason for creating such is a script is to make data uncompressible. (fair testing for utilities)
- CopyData.[utility name]: scripts which initiates the dataset transfer and captures the essential information about the transfer. Creates files:
  - Abstract – Information about transfer execution time and parameters which are used
  - Log – All log information about the transfer
  - Traceroute – packets source and destination path
  - Sosreport – copy of /proc with all system information

SCRIPTS AND THEY PURPOSES (2)

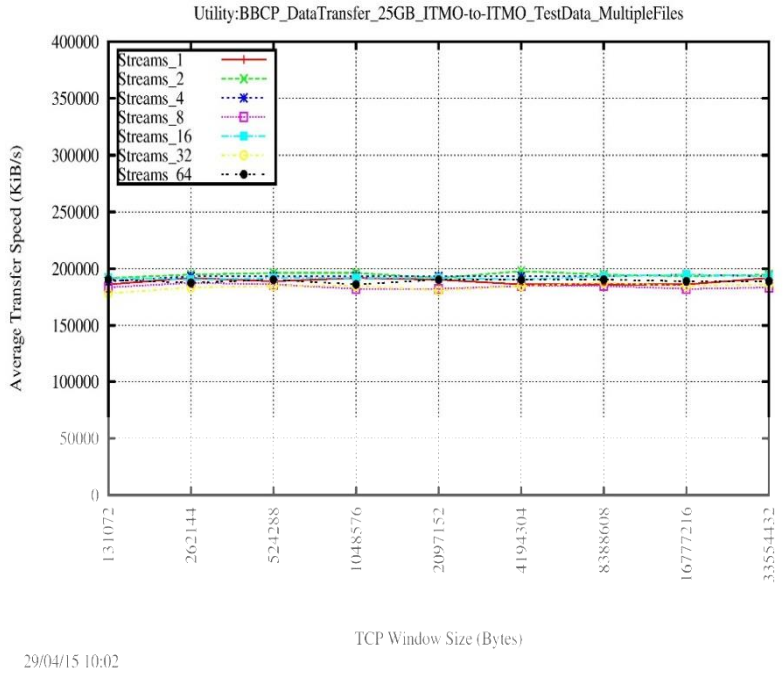
- ExecuteMultipleCopyData.[utility name]: script with purpose to lunch different scenarios. User can lunch this script using different number of parallel streams and TCP window size.
- PlotGraphs.[utility name]: this script will plot graphs using gnuplot and gives output a postscript file.
- set-passwordless-ssh : created to set password less ssh login to the remote host.

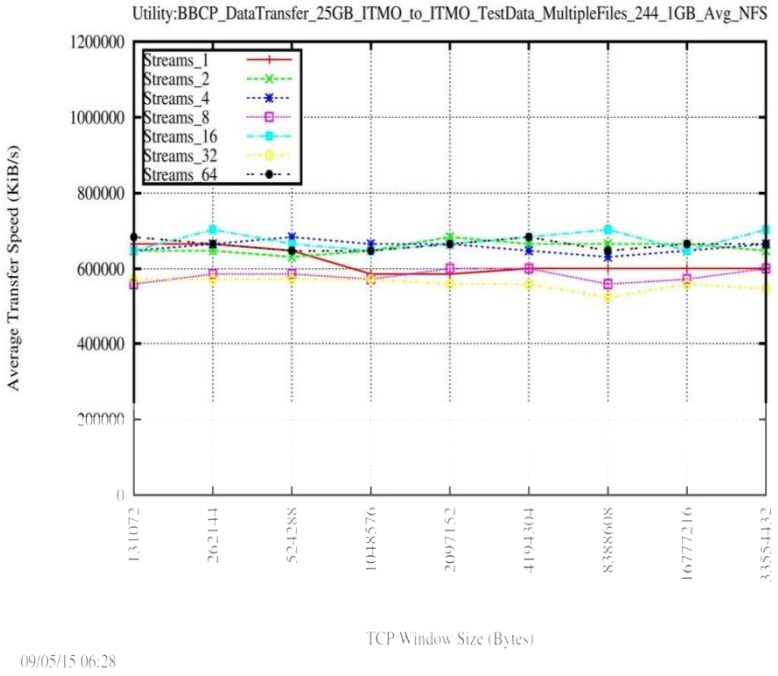
Tested Scenarios

- Virtual Environment (ITMO - ITMO)
  - Dataset Size: 25GB which 244 files of 100MB each
  - Parallel streams: 1,2,4,8,16,32, and 64
  - Window Size: 131, 262, 524, 1048, 2097, 4194, 8388, 16777, and 33554 Kilo Bytes
- Real Network (PNPI - ITMO)
  - Dataset Size: 25GB which 244 files of 100MB each
  - Parallel streams: 1,2,4,8,16,32, and 64
  - Window Size: 131, 262, 524, 1048, 2097, 4194, 8388, 16777, and 33554 Kilo Bytes

RESULTS

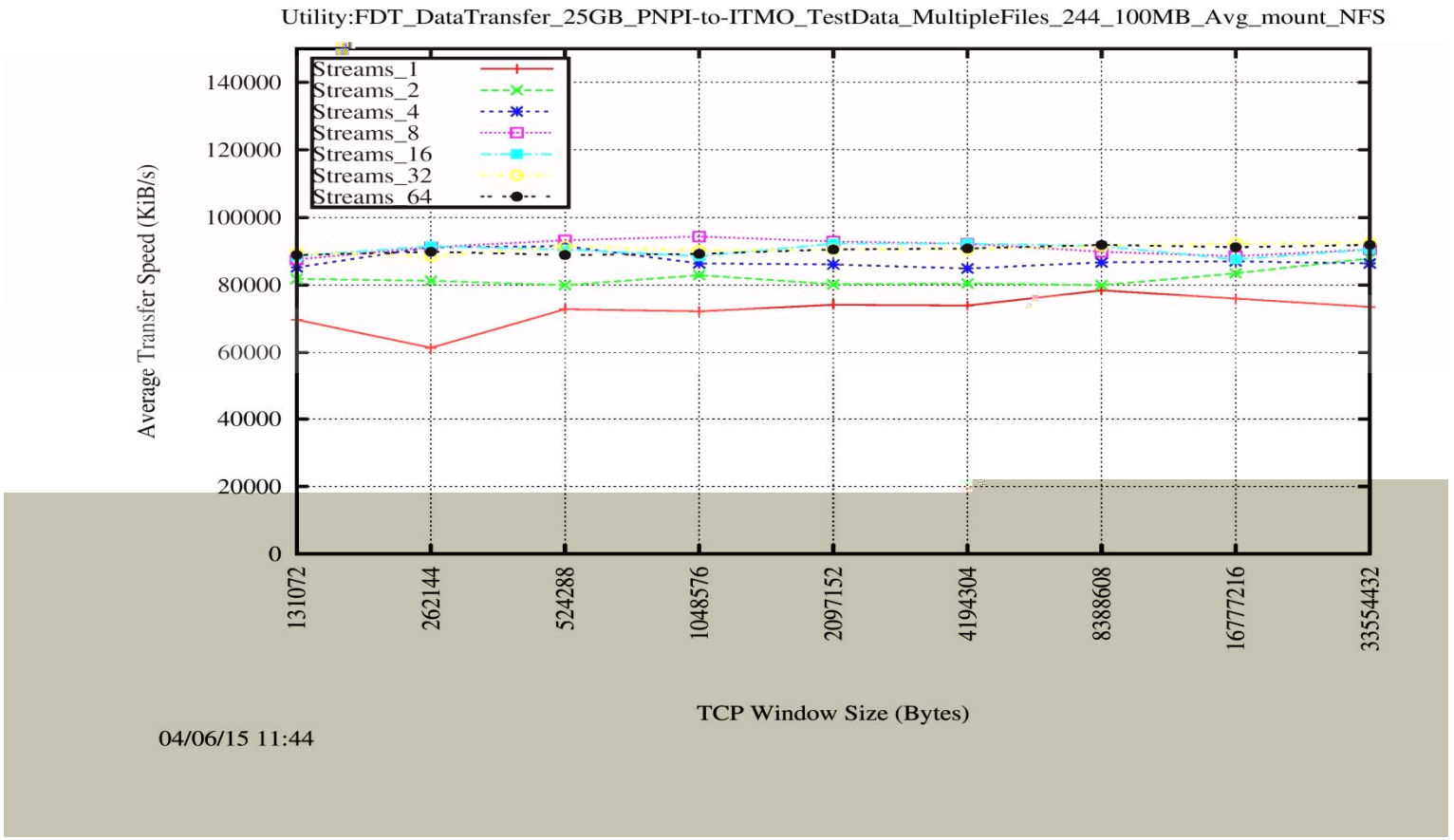
- Data located on Hard Disk Drive and mount NFS(BBCP)





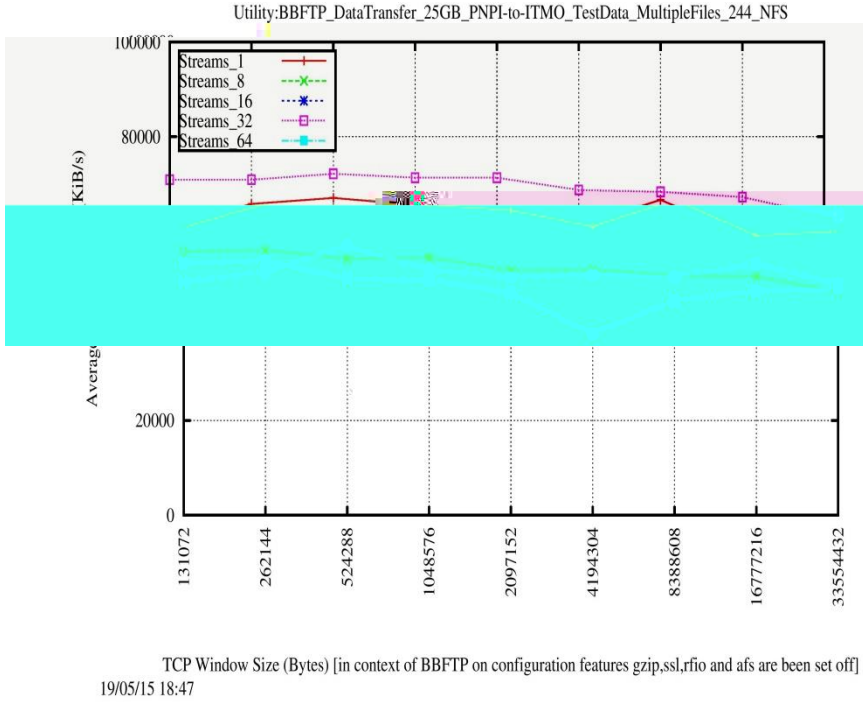
RESULTS(2)

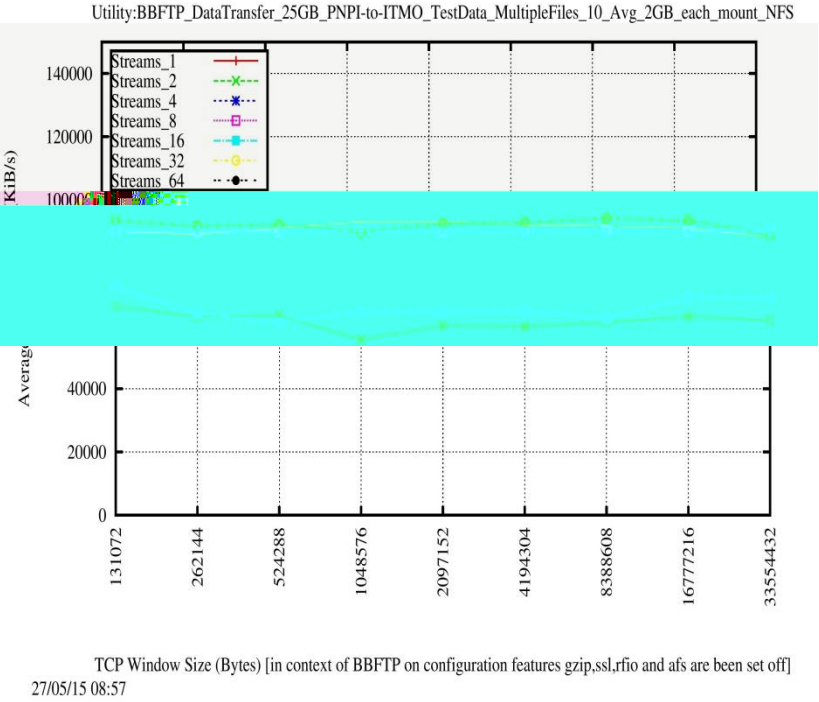
- Transfer from PNPI to ITMO (FDT)



RESULTS(3)

- Transfer from PNPI to ITMO (BBFTP)





Conclusion

- In virtual environment we can always achieve higher transfer rate because its network is software, on the other hand using real public network we are congestions and collisions.
- Running tests inside virtual environment before testing on real network was efficient in terms speed, and decisions could be taken before testing on real network.
- Transferring data from mount NFS instead of HHD helped to achieve higher transfer speed which makes our system more efficient.
- Using large amount of parallel streams and big window size it will only consume more resources and may not provide higher transfer speed and it will consume more energy.

Future Work

- Run tests with the remaining utilities
- Test large amounts of data
- Energy monitoring about is not possible for Virtual Machines.
- RAID storage system which will allow to parallel read/write to overcome the HHD limitation.

Questions