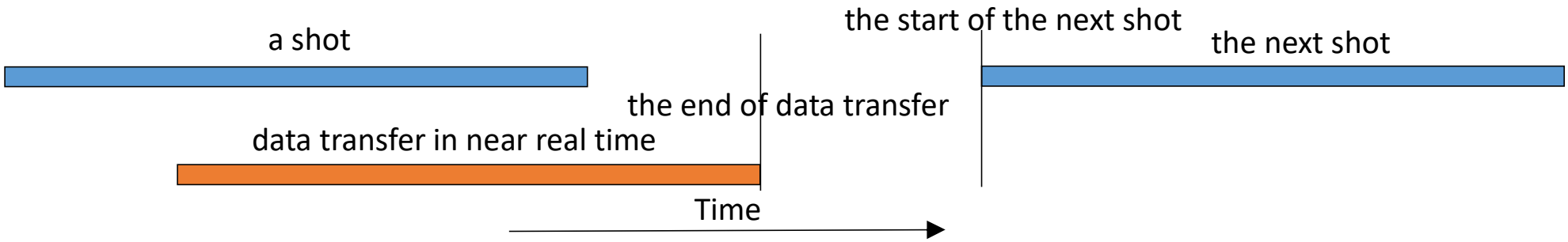
A data transfer method for physics data of experimental fusion reactors using virtual disks

K. Yamanaka[1], H. Nakanishi[2], S. Tokunaga[3] and S. Urushidani[1]

1 National Institute of Informatics (NII), Chiyoda, Tokyo, Japan
2 National Institute for Fusion Science (NIFS), Toki, Gifu, Japan
3 National Institutes for Quantum and Radiological Science and Technology (QST), Rokkasho, Aomori, Japan

Background and Objectives

- The Japanese fusion community is planning to transfer all ITER's physics data to the Remote Experiment Centre (REC) in Japan. Because experiments in ITER are crucially important for the roadmap toward DEMO.
- These data will be provided to domestic researchers along with a supercomputer so that researchers (who granted appropriate data access privileges) can freely analyze ITER's data and accelerate developments of DEMO.
- We aim to transfer data from ITER in near real time so that researchers can quickly access the latest data from the supercomputer.
- "Near Real Time" transfer: the transfer of one-shot data is completed before the start of the next shot.



ITER's physics data

- The generation rate of physics data in ITER is assumed as the table below.
- The physics data of experimental fusion reactors is a collection of files of various sizes from many sensors.
- It is difficult to read, write, and transfer many small files at high speed. This is called the Lots of Small Files (LoSF) Problem.

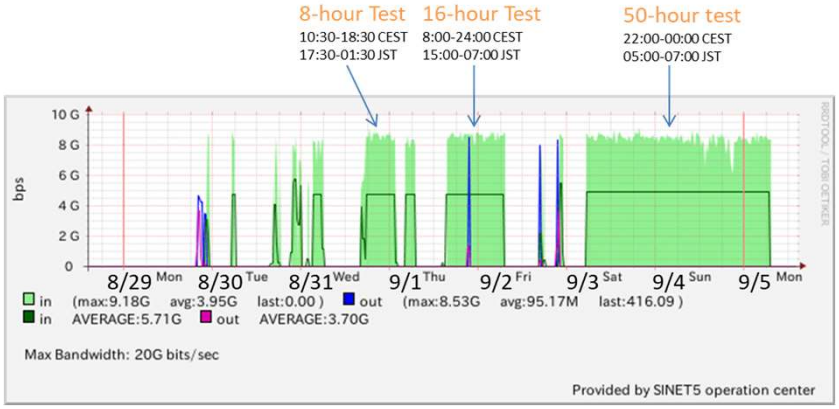
ITER phase	Total DAN† archive rate	Plasma duration Time	Data volume per shot
Initial	2 GB/s (16 Gbps)	400 - 500 s (6.7 - 8.3 min)	800 GB - 1 TB
Final	50 GB/s (400 Gbps)	1000 s (16.7 min)	50 TB

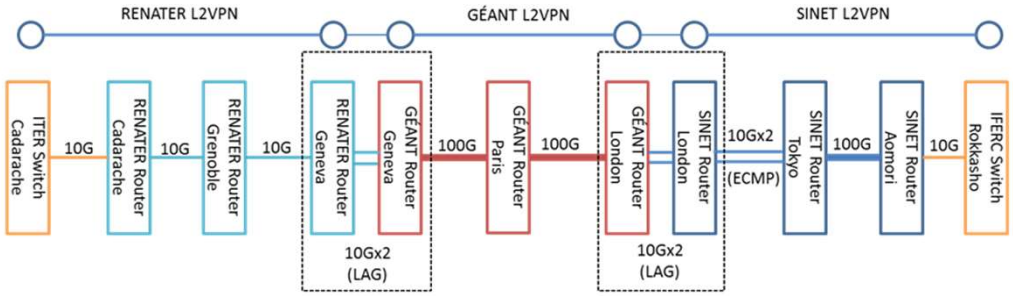
†DAN: Data Archive Network

Data to be sent to REC

MMCFTP (Massively Multi-Connection File Transfer protocol)

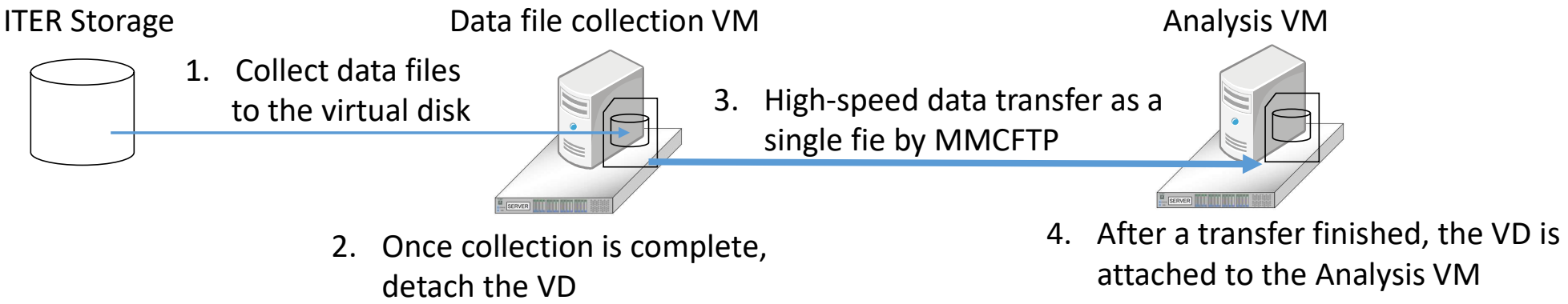
- MMCFTP is a high-speed file transfer tool (software) developed by NII.
- MMCFTP uses several thousands of TCP connections to sustain the specified target speed and controls the number of TCP connections dynamically according to network conditions.
- In 2016, Data transfer test from ITER to REC using MMCFTP executed. An 1TB tar file including LHD and JT60U data was transferred repeatedly at about 8 Gbps speed and about 20 min transfer time.





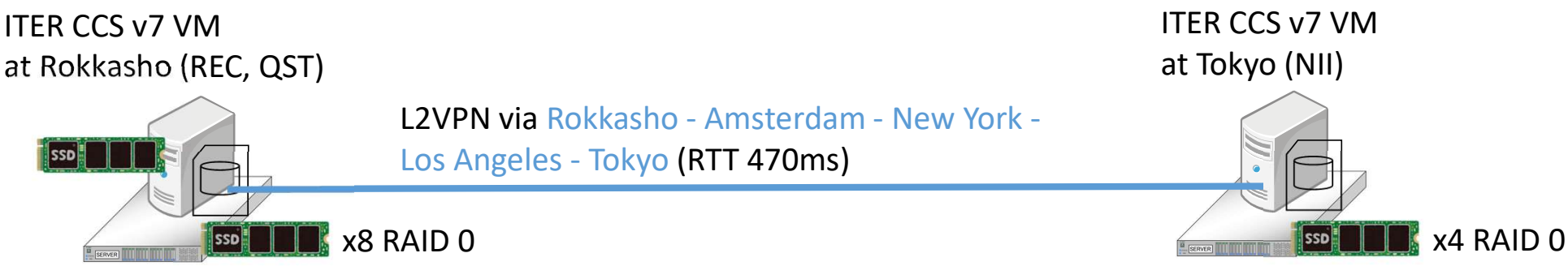
Data transfer method using Virtual Disk

- The conventional approach to the LoSF problem for MMCFTP was to archive many small files using TAR, transfer the archived file, and then unarchive it on the receiver side. The problem with this method is that archiving and unarchiving takes extra time. Virtual Disk method eliminates this overhead.
- A virtual disk (VD) is mainly used as storage for a virtual machine (VM) and is a file system on the VM. On the other hand, a VD is just a file on the VM host (Hypervisor).
- By storing physics data files into a VD on the VM and transferring the VD on the hypervisor, it is possible to transfer multiple files as a single file. By attaching the transferred VD to the analysis VM, analysis can be carried out as usual.



Evaluation test and environment

- There are some options about VD format (raw, sparse raw, qcow2). We investigated the effect of different virtual disk formats on data replication.
- As a substitute for ITER storage, we prepared an NVMe SSD containing LHD data, approximately 1.2TB (12,885 files), and attached it to the VM as an SSD device.
- Virtual disk images to be transferred was prepared on a software Raid device (MD). Storage of receiver also used a MD.
- Data transfers between REC and NII was carried out over a special L2VPN that passed through SINET's overseas nodes. The round-trip time was 470 ms, approximately double the RTT between ITER and REC (250 ms).



Test results

- This method consists of the following seven steps.
  1) File allocation for a virtual disk. 2) Format the VD. 3) Attach the VD to a data collection VM. 4) Data files copy from storage. 5) Detach the VD. 6) The VD file transfers to remote site. 7) Attach the VD to an analysis VM.
- We compare the differences in steps 1, 4, and 6 due to the different virtual disk formats because the other steps can be completed quickly.

	Dense RAW	Sparse RAW	Qcow2	Comment
1. File allocation	10 min 4 sec	< 1 sec	< 1 sec	disk size: 1.374 TB
4. File copy	9 min 54 sec	10 min 5 sec	10 min 50 sec	cp –r src dst
6. Data transfer File size Speed	2 min 16 sec 1.374 TB 80.6 Gbps	2 min 18 sec 1.374 TB 79.5 Gbps	2 min 3 sec 1.244 TB 80.3 Gbps	Since qcow2 allocates disk space as needed, the file size may be smaller than the disk size. We set the target speed at 80 Gbps, because the L2VPN consist of shared 100Gbps network lines.
Total	22 min 14 sec	12 min 23 sec	12 min 53 sec	

Consideration

- The key factor is the time required for “File copy.” In this experiment, we simply used ‘cp’, but this time may be shortened by using other tools. However, the current UDA seems to support no API for bulk data transfer. If we use these APIs for “File copy”, more time will be required.
- With the virtual disk method, because the VM host does not have access to the ITER system, ITER development teams do not need to support data transfer tools. Meanwhile, domestic agencies can freely choose their file transfer tool.
- To achieve true near real-time replication, other techniques are also required, such as data file segmentation. Segmentation involves splitting data from a single sensor into files at appropriate time intervals. Transfer can start from a file that has already been generated, so data transfer can start while the shot is still in progress.

Summary

- To address the LoSF problem, we proposed a data transfer method using virtual disks and evaluated it.
- This method enables high-speed transfer of ITER data, and speed of 80 Gbps is possible. 1 TB of data per shot of the ITER initial phase can be transferred in approximately 2 minutes.
- The transferred virtual disk can be attached to an analysis virtual machine and analysis work can begin immediately.
- A major factor in whether near real time transfer is possible is the time of “File copy”. The current UDA seems to support no API for bulk data transfer. I would like a method or APIs for bulk date transfer to be provided.