The Usability of the Data-Cluster Based on the CEPH Platform in Real Network Environment

David Malanik

***Abstract*—**this paper is focused to the real implementation of the data cluster based on the CEPH [1] technology. The first part is focused to the solution with geographically separated nodes placed on the shared network infrastructure. Each node; store point is in different physical location and on different subnets. The second part of this paper shows the comparison of the separated node and non-separated nodes. The non-separated solution is located in one room with one isolated network infrastructure. Realized test show the dependency of number of concurrent clients connected to the cluster and cluster read/write bandwidth. These tests show the potential limits of the developed solution. The test compares the effect of shared infrastructure and geographical separation of data-cluster nodes.

***Keywords*—**Distributed file system, data backup, CEPH, RDB, HA data cluster, storage, performance

I. Introduction

THE problem based on the storage capacity and fail-proof run time, is one of the major problem in many companies. The first requirement is the scalability of the storage. The Storage must be capacity flexible as soon as possible. The second and probably more important is the question of high availability an error-proof solution for data storage. These two major problem grades in the critical infrastructure.

There is the request to have the data in secure locations; especially in the geographically separated location. The purpose is clear. If there is any problem in the main data center, the company does not need any blackouts. So it is strictly recommended to have separated power part of IT infrastructure (the blade servers, rack servers, etc.) and the storage (SAN/NAS).

The possible solution is the isolation (by different position) between servers and SAN storage. But for the critical infrastructure is necessary to have backup solution for each HW parts. So it is not only one SAN system, the system is projected with strategy N+1, or 2N+1.

The major limitation of the geographical separation is inside the stability, latency and bandwidth of the network connectivity. The second part of this paper looks inside the solution with non-separated nodes; this solution performs the low level of HA and security function; but shows the optimal state, which is possible to implement without network connectivity limitations.

This paper used different technology for the storage model. There is not the classical SAN system with FC or FCoE [4]. The designed system used classical server HW focused on storage capacity, the CPU and memory is not be a primary parts of structure. It is possible to use any servers with sufficient network connection and storage capacity. The purpose of this solution flowing from the possibility; if is possible to use classical (and in many cases, used servers older than 5 year) as the data storage.

David Malanik is with the the Faculty of Applied Informatics, Tomas Bata University in Zlín, Nad Stráněmi 4511, 760 05 Zlín, Czech Republic (e-mail: dmalanik@fai.utb.cz).

The work was performed with financial support of research project NPU I No. MSMT-7778/2014 by the Ministry of Education of the Czech Republic and also by the European Regional Development Fund under the Project CEBIA-Tech No. CZ.1.05/2.1.00/03.0089.

II. Used technologies

A. Storage cluster

The CEPH storage was used as a storage cluster technology. It is the object storage that provides seamless access to objects using native language bindings or radosgw, a REST interface that's compatible with applications written for S3 and Swift. The main technology for this solution is in CEPH Rados block storage device that provides access to block device images that are striped and replicated across the entire storage cluster. The CEPH also provides a POSIX-compliant network file system that aims for high performance, large data storage, and maximum compatibility with legacy applications [1].

B. Cluster node – HW/SW specification

The HW part of storage node is realized by two virtual machine with identic specification. The VM are hosted on FS-RX100 server with Proxmox VE hypervisor [1].

Specification of VM/(nodes):

CPU: 2x Intel ® Xeon ® X3320
RAM: 2GB
HDD: 50GB SATA-II
LAN: 1000BASE-T
OS: Debian 7.0.2 – 64-bit

C. Cluster client

The client for storage cluster is implement on physical server with Proxmox VE hypervisor. The hypervisor is for creating a VM for testing concurrent read/write operation to

the storage cluster. The physical computer simulates the server with many virtual machines using storage cluster. The data of all virtual machine is located on Rados block storage device provided by the cluster.

Specification of the cluster client:
CPU: Intel ® i7-3770 (4 physical cores, 8 logical)
RAM: 8GB
HDD: 250GB SATA-II
LAN: 1000BASE-T
OS: Proxmox VE 3.1

D. Network model

The network model is shown on figure below.

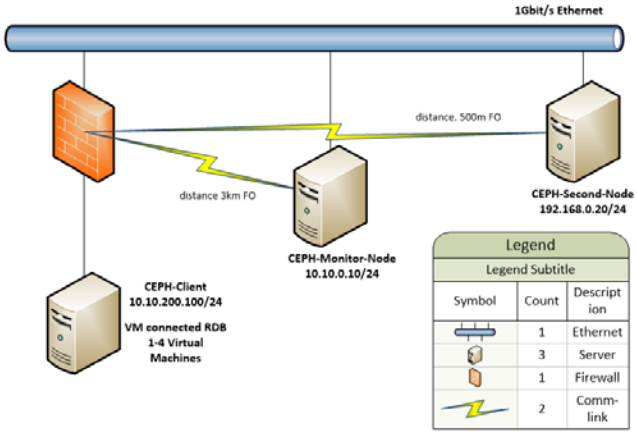

Fig. 1 Network schema

The nodes are located in different data center. The connection between nodes and CEPH-client is realized by 1 Gbit/s Ethernet. The connection between Firewall and nodes is realized by fiber optic, the rest connection between firewall and CEPH-client is on 1000BASE-T [5,6]. The network is not reserved only for this cluster. The lines is shared with other application.

E. VM – on CEPH client

The CEPH-Client server contains 4 virtual machine that simulates the real applications running on the one physical server. Each VM is connected to share 1000BASE-T Ethernet.

Specification of VM:
CPU: Intel ® i7-3770 (2 logical cores)
RAM: 1.5GB
HDD: 15GB SATA-II
LAN: 1000BASE-T
OS: Debian 7.0.2 – 64-bit

III. NETWORK BANDWIDTH TEST

The first part of CEPH cluster testing is in local network bandwidth testing. The network is not isolated from other application and the capacity is shared [9]. The first rand of tests become from LAN testing realized by the *iperf* Linux tool [7]. Test contains 50 measurements during whole day [3]. Each measure contains 4 part; the appropriate commands are shown below.

```
#iperf –c <IP> -t 10
#iperf –c <IP> -t 30
#iperf –c <IP> -t 60
#iperf –c <IP> -t 120
```

The –t parameter from command represents the time of bandwidth test in second.

Test procedure is realize between nodes:
CEPH-Client vs. CEPH-Monitor-Node
CEPH-Client vs. CEPH-Second-Node
CEPH-Monitor-Node vs. CEPH-Second-Node

The measure is not being realized only for data capacity of the network; the second monitored parameter was the stability of line. Results are shown below.

A. CEPH-Client vs. CEPH-Monitor-Node

The Table I shown the maximal, minimal and average value of network bandwidth examined by the testing procedure.

Table I CEPH-Client vs. CEPH-Monitor-Node

	Mbit/s			
	10s	30s	60s	120s
MIN.	180	172	181	188
MAX.	226	213	212	212
AVG.	198.9	201.5	202.1	201.5

The best stability of the network was with 30s time interval. The test duration is over 3 hour, and the stability graph is shown on Fig. 1.

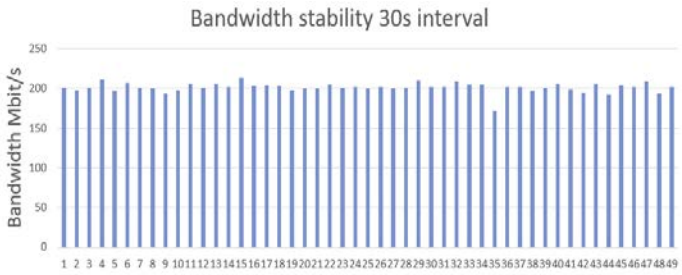

Fig. 2 Bandwidth stability - 30s interval

The opposite side of the stability log is represented by the 60s measure interval that is shown on Fig. 3.

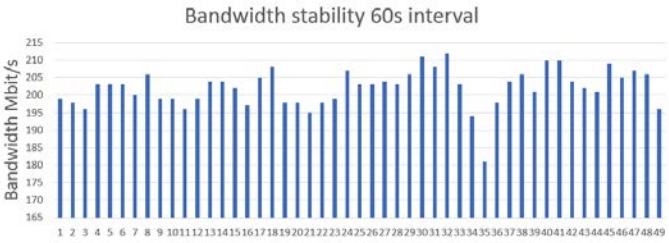

Fig. 3 Bandwidth stability - 60s interval

B. CEPH-Client vs. CEPH-Second-Node

The Table II shown the maximal, minimal and average value of network bandwidth examined by the testing procedure. The communications is routed by firewall and there is some

increase of network capacity; the firewall is shared with all infrastructures.

Table II CEPH-Client vs. CEPH-Second-Node

	Mbit/s			
	10s	30s	60s	120s
MIN.	498	510	512	511
MAX.	524	523	523	523
AVG.	514.4	515.1	515.5	515.3

The best stability of the network was with 120s time interval. The test duration is over 3 hour, and the stability graph is shown on Fig. 4.

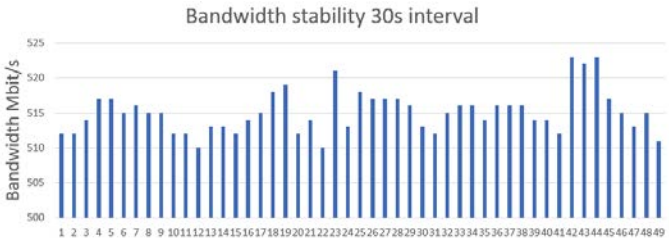

Fig. 4 Bandwidth stability - 120s interval

The opposite side of the stability log is represented by the 30s measure interval that is shown on Fig. 5.

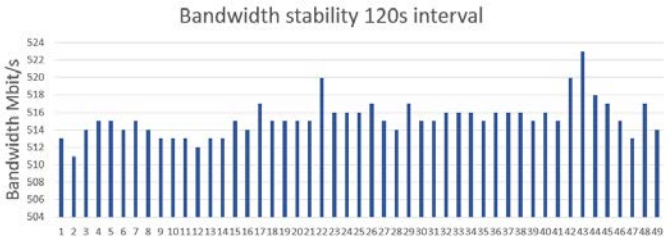

Fig. 5 Bandwidth stability - 30s interval

C. CEPH-Monitor-Node vs. CEPH-Second-Node

The Table III shown the maximal, minimal and average value of network bandwidth examined by the testing procedure.

Table III CEPH-Monitor-Node vs. CEPH-Second-Node

	Mbit/s			
	10s	30s	60s	120s
MIN.	560	571	558	571
MAX.	616	628	634	623
AVG.	583.2	590.9	589.8	590.9

The best stability of the network was with 30s time interval. The test duration is over 3 hour, and the stability graph is shown on Fig. 6.

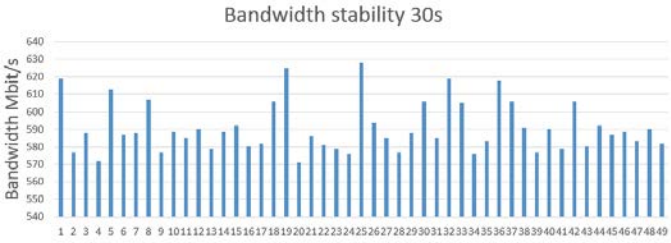

Fig. 6 Bandwidth stability - 30s interval

The opposite side of the stability log is represented by the 120s measure interval that is shown on Fig. 7.

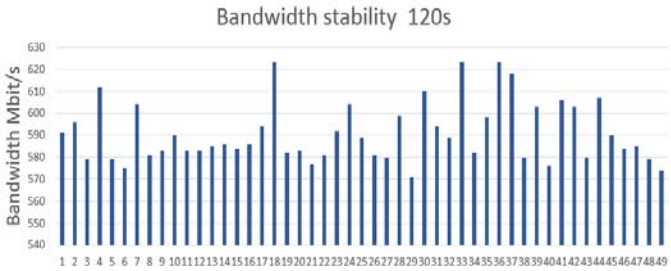

Fig. 7 Bandwidth stability - 120s interval

The result of test represents the actual bandwidth of all lines between active nodes in the cluster network. The speed is stable without any significant fluctuation. The minimal variety of speed is done with the common traffic inside the network.

The connection between CEPH-Client and both nodes is realized with firewall, which represented the bottleneck of this solution. The marginal different of speed (depended on firewall) is shown in the table below.

Table IV Bandwidth dependency on Firewall

Connection	AVG. speed	Firewall
CEPH-Client vs. CEPH-Monitor-Node	201 Mbit/s	YES
CEPH-Client vs. CEPH-Second-Node	**515 Mbit/s**	**NO**
CEPH-Monitor-Node vs. CEPH-Second-Node	**589 Mbit/s**	**NO**

IV. Cluster Nodes Local Storage Bandwidth

The next part of test is focused to real HDD speed of each cluster node. The test was realized by the Linux command *dd*; specifically by the various option of this command shown below [2].

```
#dd bs=4K count=2000 if=/dev/zero of=test conv=fdatasync
#dd bs=64K count=2000 if=/dev/zero of=test conv=fdatasync
#dd bs=256K count=2000 if=/dev/zero of=test conv=fdatasync
#dd bs=1M count=2000 if=/dev/zero of=test conv=fdatasync
```

The different size of block in set 4 KB, 64 KB, 256 KB and 1 MB represents the variability of saved data to the cluster. It

simulate the variability of file size copied to the cluster. The parameter count represents the repetition of each copy tests.

A. CEPH Monitor Node

The Table V shown the maximal, minimal and average value of disc bandwidth examined by the testing procedure. The test set contains 4 KB, 64 KB, 256 KB and 1 MB blocks.

Table V CEPH-Monitor-Node dd test

	MB/s			
	4KB	64KB	256KB	1M
MIN.	67.2	238.0	260.0	385.0
MAX.	173.0	408.0	427.0	492.0
AVG.	90.9	294.9	341.8	441.7

The best stability of the disk bandwidth was with 1 MB block. The test duration is over 1 hour, and the stability graph is shown on Fig. 8.

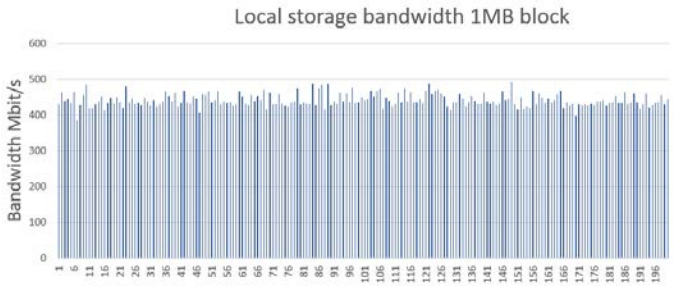

Fig. 8 Local storage bandwidth 1 MB block Monitor node

The opposite side of the stability log is represented by the 4 KB block size that is shown on Fig. 9.

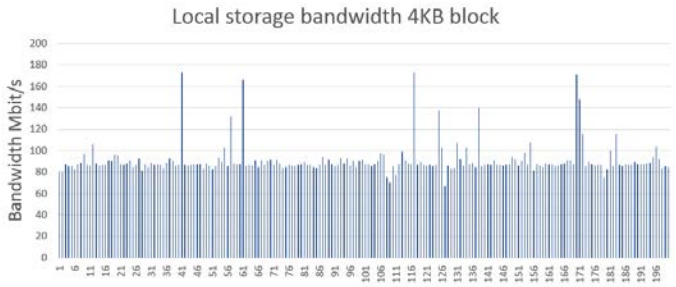

Fig. 9 Local storage bandwidth 4 KB block Monitor node

B. CEPH Second Node

The Table VI shown the maximal, minimal and average value of disc bandwidth examined by the testing procedure. The test set contains 4 KB, 64 KB, 256 KB and 1 MB blocks.

Table VI CEPH-Second-Node dd test

	MB/s			
	4KB	64KB	256KB	1M
MIN.	45.0	103.0	101.0	106.0
MAX.	163.0	173.0	136.0	135.0
AVG.	96.8	134.2	126.3	127.5

The best stability of the disk bandwidth was with 1 MB block. The test duration is over 1 hour, and the stability graph is shown on Fig. 10.

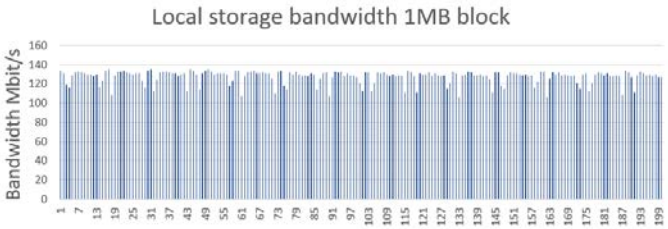

Fig. 10 Local storage bandwidth 1 MB block Second node

The opposite side of the stability log is represented by the 4 KB block size that is shown on Fig. 11.

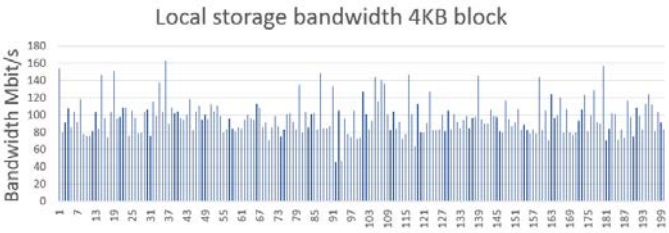

Fig. 11 Local storage bandwidth 4 KB block Second node

The maximal bandwidth of the local storage is higher than theoretical network bandwidth between nodes and the client server. The average local storage bandwidth was over 90MB/s. The best network bandwidth examined by test was approximately 600Mbit/s (approx. 75MB/s). That indicates the assumption; the local storage bandwidth does not be a bottleneck of designed solution. The local bandwidth is significantly higher than theoretical and real tested bandwidth of network.

The minimal tested network bandwidth was 172Mbit/s (approx. 21.5 MB/s).

V. CEPH STORAGE BANDWIDTH TEST

This part of paper is focused to real bandwidth in disk operation. The test is written with respect of real application [10].

The testing scheme is shown on Fig. 12. It is realized by one physical server with connection to the storage cluster [11]. There are 4 virtual machine hosted on the physical server. The VM data is stored in the CEPH storage cluster used by Rados block device. The local storage of server is not used. The local operation system has not any SWAP device.

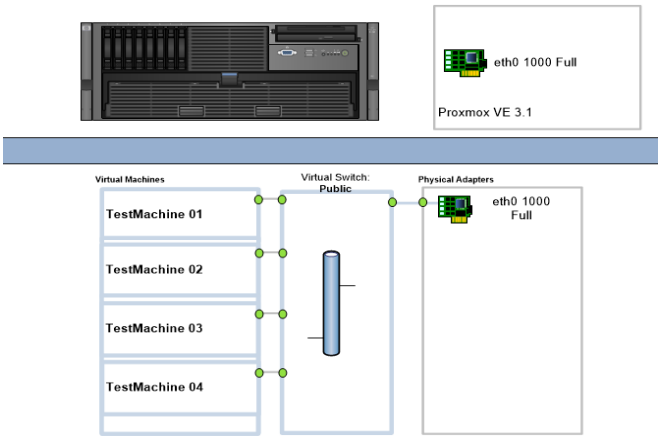

Fig. 12 VM schema

The test was realized by the Linux command *dd*; specifically by the various option of this command shown below [8].

```
#dd bs=4K count=2000 if=/dev/zero of=test conv=fdatasync
#dd bs=64K count=2000 if=/dev/zero of=test conv=fdatasync
#dd bs=256K count=2000 if=/dev/zero of=test conv=fdatasync
#dd bs=1M count=2000 if=/dev/zero of=test conv=fdatasync
```

The different size of block in set 4K, 64K, 256K and 1M represents the variability of saved data to the cluster. It simulates the variability of file size copied to the cluster [8]. The parameter count represents the repetition of each copy tests. The test set contains 200 repetitions of *dd* commands set.

The following parts reports about testing reports. These tests were parted to 4 subcategories. The first test is realized with one VM running on cluster. Next parts describe the test result with increased number of concurrent VM on one cluster. Tests were realized for 1-4 concurrent VM.

A. One VM on cluster

The first test was realized with one active virtual machine. The testing procedure was processed on real shared network infrastructure. The test lasted over 5 hour of continual bandwidth testing.

The Table VII shows minimal, maximal and average values of realized tests with variable block size for write/read operations. The values fluctuate **from 22.1 MB/s to 39.1 MB/s**. These values is significantly lower that the examined local and network storage bandwidth. The possible weakness is inside the communication over Rados block devices.

Table VII One VM dd test

	MB/s			
	4KB	64KB	256KB	1M
MIN.	9.3	20.6	25.2	27.1
MAX.	36.8	53.8	35.9	34.6
AVG.	22.1	39.1	28.6	30.6

The most stable bandwidth was reported by the 1 MB block size. The possible reason flowing from the latency of network interfaces. The impact of latency will be better with higher block size of data. The storage bandwidth was from 27.1 MB/s to 34.6 MB/s without any significant deviations. The stability test is shown on Fig. 13.

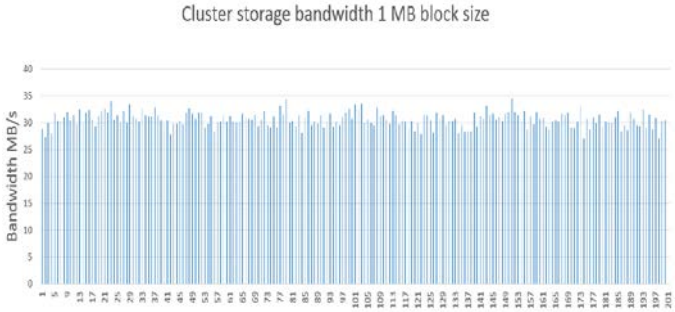

Fig. 13 Cluster storage bandwidth one VM 1 MB block size

The worst stable bandwidth was reported by the 4 KB block size. The storage bandwidth was from 9.3 MB/s to 36.8 MB/s with really significant deviations. The stability test is shown on Fig. 14. The network latency and service latency of writing to cluster consume more bandwidth with these small blocks.

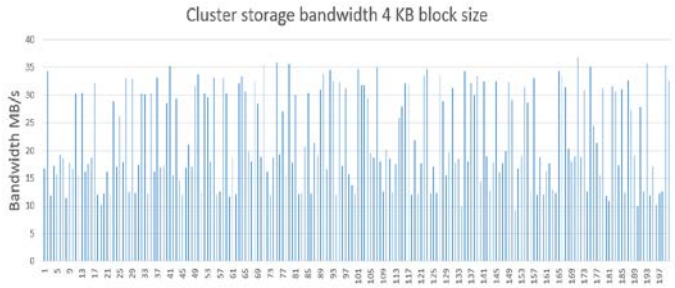

Fig. 14 Cluster storage bandwidth one VM 4 KB block size

The results from this scenario show the real bottleneck of this solution for only one virtual machine on the cluster. The storage bandwidth is lower than local storage bandwidth; this was predictable with only 1 Gbit/s network connection. But the storage bandwidth is lower that this network bandwidth (approx. 30 MB/s vs. 75 MB/s)[1]. But still it is quite higher than the minimal value (approx. 30 MB/s vs. 21.5 MB/s)[2].

B. Two concurrent VM on cluster

The second test was realized with two active virtual machines. The testing procedure was processed on real shared network infrastructure. The test lasted around 6 hour.

The Table VIII and Table IX show the comparison of bandwidth identified on each machine during the test. The bandwidth of both virtual machines are quite identical, there is no one significant difference. The cluster distributed the bandwidth to the two equivalent machines. In comparison with the one VM test; values of each machine is too close to the one VM solution. That shows; the one machine does not use 100% of distributed bandwidth.

[1] The maximal local network bandwidth was examined in chapter IV
[2] The minimal local network bandwidth was examined in chapter IV

Table VIII First VM dd test

	MB/s			
	4KB	64KB	256KB	1M
MIN.	3.7	15.9	17.8	23.5
MAX.	33.2	50.6	27.9	32.7
AVG.	15.9	32.5	21.8	25.7

Table IX Second VM dd test

	MB/s			
	4KB	64KB	256KB	1M
MIN.	3.7	15.8	17.5	23.7
MAX.	37.1	48.2	27.7	28.7
AVG.	17.4	32.9	21.9	25.9

The Table X shows minimal, maximal and average values of realized tests with variable block size for write/read operations. The values of the summary bandwidth of two concurrent machines fluctuate **from 33.4 MB/s to 65.5 MB/s**.

Table X Two concurrent VM summary dd test

	MB/s			
	4KB	64KB	256KB	1M
MIN.	9	40.7	36.8	48.0
MAX.	63.5	86.7	51.6	58.8
AVG.	33.4	65.5	43.7	51.7

The most stable bandwidth was reported by the 1 MB block size again. The possible reason flowing from the latency of network interfaces. The impact of latency will be better with higher block size of data. The storage bandwidth was from 48 MB/s to 58.8 MB/s without any significant deviations. The stability test is shown on Fig. 15.

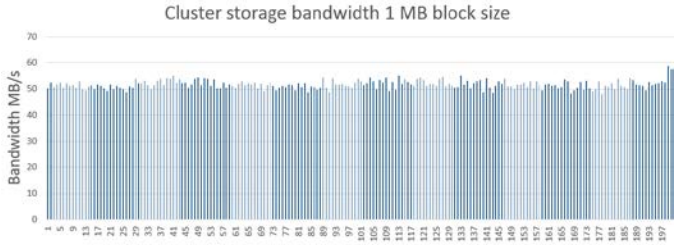

Fig. 15 Cluster storage bandwidth two VMs 1 MB block size

At the opposite side, the test with 4KB block size had the worst stability again. The storage bandwidth was from 9 MB/s to 63.5 MB/s with really significant deviations. The stability test is shown on Fig. 16.

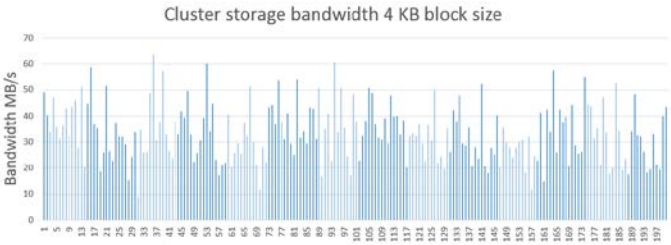

Fig. 16 Cluster storage bandwidth two VMs 4 KB block size

The solution of two concurrent virtual machines located on one cluster shows the better values of summary bandwidth than the solution with one isolated virtual machine. This assumes, that the cluster based on the designed infrastructure is able to serves the data to two concurrent machines without any problem. The bandwidth values is fluctuate from 33.4 MB/s to 65.5 MB/s which is quite interesting, because the test realized by the *iperf* command[3] shows that the network bandwidth between CEPH-client and CEPH-Monitor-Node is inside the interval 198.9 Mbit/s to 201.5 Mbit/s (24.9 – 25.2 MB/s).

C. Three concurrent VM on cluster

The third test was realized with three active virtual machines. The testing procedure was processed on real shared network infrastructure. The test lasted 8 hour and 30 minute.

Tables placed below (Table XI, Table XII, Table XIII) show the comparison of bandwidth identified on each machine during the test. The test shows the distribution of available bandwidth is equally to number of active clients. The measured values is practically same on each clients.

Table XI First VM dd test

	MB/s			
	4KB	64KB	256KB	1M
MIN.	1.4	12.9	10.9	13.5
MAX.	26.9	41.0	21.5	28.0
AVG.	7.9	23.3	15.5	19.0

Table XII Second VM dd test

	MB/s			
	4KB	64KB	256KB	1M
MIN.	1.4	11.9	9.3	13.6
MAX.	30.8	39.3	25.6	30.3
AVG.	9.0	23.2	15.7	18.9

Table XIII Third VM dd test

	MB/s			
	4KB	64KB	256KB	1M
MIN.	1.6	7.7	10.0	15.2
MAX.	35.7	41.0	21.3	23.7
AVG.	9.4	23.9	15.3	19.3

The Table XIV shows minimal, maximal and average values of realized tests with variable block size for write/read

[3] Tests described in chapter III

operations. The values of the summary bandwidth of three concurrent machines fluctuate **from 26.3 MB/s to 70.5 MB/s**.

Table XIV Three concurrent VM summary dd test

	MB/s			
	4KB	64KB	256KB	1M
MIN.	6.7	41.1	32.5	44.8
MAX.	72.8	103.5	65.7	78.2
AVG.	26.3	70.5	46.4	57.2

The most stable bandwidth was reported by the 1 MB block size again. The possible reason flowing from the latency of network interfaces. The impact of latency will be better with higher block size of data. The storage bandwidth was from 44.8 MB/s to 78.2 MB/s without any significant deviations. The stability test is shown on Fig. 17.

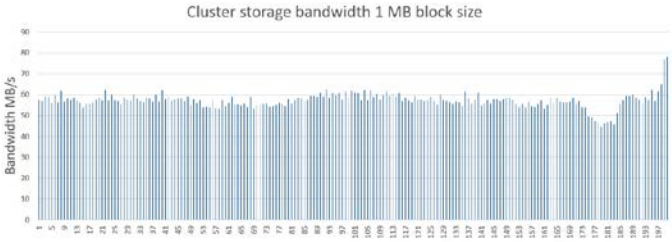

Fig. 17 Cluster storage bandwidth three VMs 1 MB block size

The opposite side is represented by the test with 4KB block size had the worst stability again. The storage bandwidth was from 6.7 MB/s to 72.8 MB/s with really significant deviations. The stability test is shown on Fig. 18.

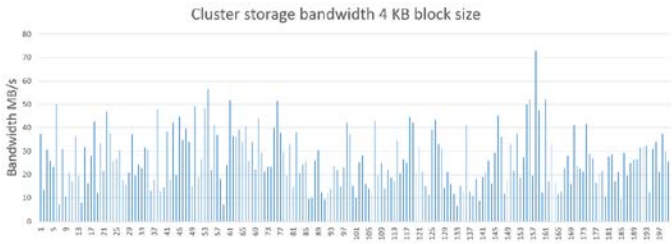

Fig. 18 Cluster storage bandwidth three VMs 4 KB block size

This solution shows additional increasing of bandwidth capacity. The maximal storage bandwidth is higher than in solution with one and two concurrent virtual machines.

The bandwidth values is fluctuate from 26.3 MB/s to 70.5 MB/s which is quite interesting, because the test realized by the *iperf* command[4] shows that the network bandwidth between CEPH-client and CEPH-Monitor-Node is inside the interval 198.9 Mbit/s to 201.5 Mbit/s (24.9 – 25.2 MB/s).

D. Four concurrent VM on cluster

The last test was realized with four active virtual machines. The testing procedure was processed on real shared network infrastructure. The test lasted approx. 11 hour.

Tables placed below (Table XV, Table XVI, Table XVII, Table XVIII) show the comparison of bandwidth identified on each machine during the test.

Table XV First VM dd test

	MB/s			
	4KB	64KB	256KB	1M
MIN.	0.76	7.7	8.2	12.2
MAX.	17.9	34.1	18.1	18.8
AVG.	4.4	17.6	11.9	15.0

Table XVI Second VM dd test

	MB/s			
	4KB	64KB	256KB	1M
MIN.	0.87	7.1	7.4	10.6
MAX.	27.1	44.5	25.7	31.3
AVG.	4.7	18.7	11.8	15.0

Table XVII Third VM dd test

	MB/s			
	4KB	64KB	256KB	1M
MIN.	0.78	7.2	6.7	12.2
MAX.	24.9	36.2	18.1	17.8
AVG.	4.8	17.7	12.2	14.9

Table XVIII Fourth VM dd test

	MB/s			
	4KB	64KB	256KB	1M
MIN.	1.0	7.3	5.4	11.6
MAX.	25.5	39.0	16.9	20.6
AVG.	5.1	17.9	11.7	15.1

The Table XIX shows minimal, maximal and average values of realized tests with variable block size for write/read operations. The values of the summary bandwidth of three concurrent machines fluctuate **from 18.9 MB/s to 71.9 MB/s**.

Table XIX Four concurrent VM summary dd test

	MB/s			
	4KB	64KB	256KB	1M
MIN.	5.1	48.0	40.3	55.0
MAX.	53.6	98.3	68.9	85.0
AVG.	18.9	71.9	47.7	60.1

The best stability of bandwidth was reported by the test with 1 MB block size. The examined bandwidth was from 55 MB/s to 85 MB/s without any significant deviations. The stability test is shown on Fig. 19.

[4] Tests described in chapter III

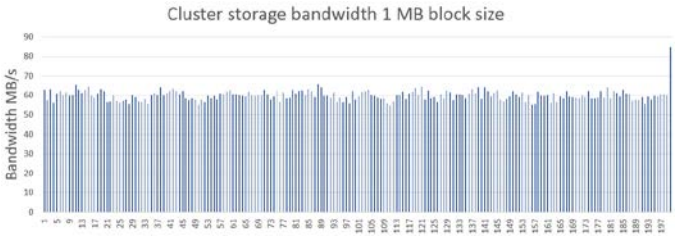

Fig. 19 Cluster storage bandwidth four VMs 1 MB block size

The worst stability of bandwidth was reported by the test with 4 KB block size. The examined bandwidth was from 5.1 MB/s to 53.6 MB/s with really significant deviations. The stability test is shown on Fig. 20.

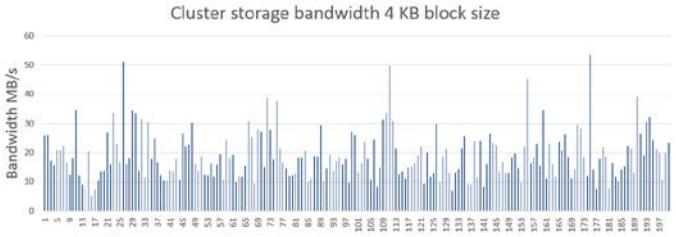

Fig. 20 Cluster storage bandwidth four VMs 4 KB block size

This scenario represents the limit of usage for one 1 Gbit/s card in the network infrastructure. There is a real decreasing of bandwidth with 4 KB block and the best performance is with 64 KB blocks. The line is quite instable with small blocks.

The bandwidth values is fluctuate from 18.5 MB/s to 71.5 MB/s which is quite interesting, because the test realized by the *iperf* command[5] shows that the network bandwidth between CEPH-client and CEPH-Monitor-Node is inside the interval 198.9 Mbit/s to 201.5 Mbit/s (24.9 – 25.2 MB/s).

VI. Usability of CEPH cluster on shared network infrastructure

Tests described in previous chapter show the potential problems with implementation with shared network infrastructure and geographical isolation with firewall device. The storage bandwidth was slower than the network and local storage bandwidth. The one virtual machine connected to storage cluster has less bandwidth than the local network. But if there is more concurrent virtual machines connected to the one storage cluster, the average bandwidth increasing too close to the maximal values of local storage/network bandwidth of each nodes.

Table XX Storage bandwidth comparison

VMs/ block size	MB/s			
	4 KB	64 KB	256 KB	1 MB
1	22.1	**39.1**	28.6	30.6
2	33.4	**65.5**	43.7	51.7
3	26.3	**70.5**	46.4	57.2
4	18.9	**71.9**	47.7	60.1

[5] Tests described in chapter III

The Table XX shows that the bandwidth with the smaller block 4 KB increasing only to 2 concurrent machines; more machines mean the increasing of bandwidth. Probably it is the limitation of network latency. The other block size shows that the average bandwidth increasing with more virtual machines. The highlighted columns represented the maximal average bandwidth for amount of concurrent machines.

VII. Comparison with cluster located on isolated network

This part of the paper examines the effect of the network connectivity. The first part of this paper show the result of many test realized onto wide network with geographically separated nodes. This part used different network model for experiment. The network is represented by the small local network on one switch. The network schema is shown on Fig. **21**.

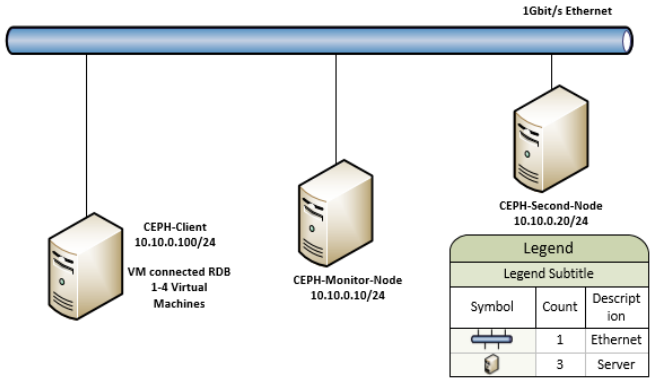

Fig. 21 Isolated network model

The configuration of CEPH nodes is the same than in the previous scenario with geographically separated nodes. The methodology of testing is same. The network bandwidth is examined by the iperf testing procedure described in the first part of this paper. The disk bandwidth is examined by the dd command. The next chapters report the comparison of the separated and non-separated location. Parts A – C report the difference in network speed and parts D - J report the difference in CEPH disk bandwidth.

A. CEPH-Client vs. CEPH-Monitor-Node

The Table XXI shown the maximal, minimal and average value of network bandwidth examined by the testing procedure in separated mode.

Table XXI CEPH-Client vs. CEPH-Monitor-Node Separated

	Mbit/s			
	10s	30s	60s	120s
MIN.	180	172	181	188
MAX.	226	213	212	212
AVG.	198.9	201.5	202.1	201.5

The Table XXII shown the maximal, minimal and average value of network bandwidth examined by the *iperf* testing procedure.

Table XXII CEPH-Client vs. CEPH-Monitor-Node

	Mbit/s			
	10s	30s	60s	120s
MIN.	948	940	950	949
MAX.	974	978	981	979
AVG.	959.9	960.9	964.1	965.4

The effect of the network model is significant. The isolated network was 5 times faster than the separated. The average bandwidth increases from **198.9-201.1 Mbit/s to 959.9-965.4 Mbit/s**. The effect of the firewall limitation is shown on this experiment briefly.

B. CEPH-Client vs. CEPH-Second-Node

The Table XXIII shown the maximal, minimal and average value of network bandwidth examined by the testing procedure in separated mode.

Table XXIII CEPH-Client vs. CEPH-Second-Node Separated

	Mbit/s			
	10s	30s	60s	120s
MIN.	498	510	512	511
MAX.	524	523	523	523
AVG.	514.4	515.1	515.5	515.3

The Table XXIV shown the maximal, minimal and average value of network bandwidth examined by the *iperf* testing procedure.

Table XXIV CEPH-Client vs. CEPH-second-Node

	Mbit/s			
	10s	30s	60s	120s
MIN.	943	940	955	952
MAX.	977	976	970	977
AVG.	960.5	957.5	963.1	966.3

The effect of the network model isn't significant as in the first comparison. There is increasing from **514.4-515.5 Mbit/s to 957.5-966.3 Mbit/s**. The bandwidth was approx. 2 times higher than in separated mode model. The lower increasing was done by the absence of the firewall between nodes in separated network model.

C. CEPH-Monitor-Node vs. CEPH-Second-Node

The **T**able **XXV** shown the maximal, minimal and average value of network bandwidth examined by the testing procedure in separated mode. The increase of the bandwidth wasn't so significant, because there wasn't a firewall between nodes.

Table XXV CEPH-Monitor-Node vs. CEPH-Second-Node Separated

	Mbit/s			
	10s	30s	60s	120s
MIN.	560	571	558	571
MAX.	616	628	634	623
AVG.	583.2	590.9	589.8	590.9

The Table XXVI shown the maximal, minimal and average value of network bandwidth examined by the *iperf* testing procedure.

Table XXVI CEPH-Monitor-Node vs. CEPH-Second-Node

	Mbit/s			
	10s	30s	60s	120s
MIN.	947	946	953	955
MAX.	968	976	977	974
AVG.	958.6	961.9	964.9	963.9

The effect of the network model isn't significant as in the first comparison. There is increasing from **583.2-590.9 Mbit/s to 958.6-964.9 Mbit/s**. The bandwidth was approx. 2 times higher than in separated mode model. The lower increasing was done by the absence of the firewall between nodes in separated network model. The fluctuation of measured values was also minimal.

D. Increasing of network bandwidth

The increasing of the network bandwidth is shown on Fig. 22. The best improve of performance was with smaller measure interval.

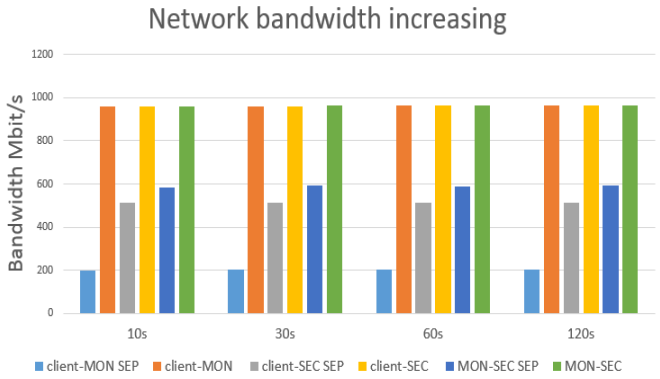

Fig. 22 Network bandwidth increasing comparison

The comparison of the values over different measurement interval shown that stability of bandwidth was significant better in scenario realized onto isolated network. But the second point that was flowing from this result was the stability is more fluctuate with smaller measurement interval.

E. One VM on cluster

The first real test of the real cluster disk bandwidth was realized with one VM on the client node; that VM perform the

continuous disk write process realized by the dd linux command[6].

The Table XXVII show the original values received from the test with separated nodes. The average write speed was **22.1-39.1 MB/s**. The best solution was with the 64 KB block size. The worst result was with 4 KB block size.

Table XXVII One VM dd test Separated

	MB/s			
	4KB	64KB	256KB	1M
MIN.	9.3	20.6	25.2	27.1
MAX.	36.8	53.8	35.9	34.6
AVG.	22.1	39.1	28.6	30.6

The Table XXVIII shows the measurement from the scenario with nodes on the small isolated LAN. The average disk write speed was **90.0-130.1 MB/s**. This value is quite similar that the maximal bandwidth of 1 Gbit/s network (approx. 125MB/s).

Table XXVIII One VM dd test

	MB/s			
	4KB	64KB	256KB	1M
MIN	41.1	106.2	98.7	108
MAX.	108	142.5	126	129.6
AVG.	90.0	130.1	108.8	119.4

The average bandwidth was **increased by 233%** (this approximation is flowing from comparison of values on 64 KB blocks).

F. Two VM on cluster

The Table XXIX shown the reported values of bandwidth from separated node scenario. The average bandwidth of two concurrently writing machines was **33.4 – 65.5 MB/s**. The fluctuation of values was done by different size of the block size. The smallest block size had a minimal bandwidth in all scenarios.

Table XXIX Two concurrent VM summary dd test Separated

	MB/s			
	4KB	64KB	256KB	1M
MIN.	9	40.7	36.8	48.0
MAX.	63.5	86.7	51.6	58.8
AVG.	33.4	65.5	43.7	51.7

The Table XXX shown the scenario with all nodes in one small isolated network. The average bandwidth was **57.2 – 118.5 MB/s**. The best bandwidth was with 64 KB block size. The block size 64 KB was examined as the best block size for higher performance in any cases.

Table XXX Two concurrent VM summary dd test

	MB/s			
	4KB	64KB	256KB	1M
MIN.	11.2	61.4	62.8	85
MAX.	112.4	163.4	120.4	107.4
AVG.	57.2	118.5	75.5	96.7

The average bandwidth was **increased by 81%** (this approximation is flowing from comparison of values on 64 KB blocks).

G. Three VM on cluster

The Table XXXI shown the reported values of bandwidth from separated node scenario. The average bandwidth of three concurrently writing machines was **26.3 – 70.5 MB/s**. The average bandwidth on 64 KB block size is quite similar than in scenario with 2 concurrent writing machines.

Table XXXI Three concurrent VM summary dd test Separated

	MB/s			
	4KB	64KB	256KB	1M
MIN.	6.7	41.1	32.5	44.8
MAX.	72.8	103.5	65.7	78.2
AVG.	26.3	70.5	46.4	57.2

The Table XXXII shown the scenario with all nodes in one small isolated network. The average bandwidth was **39.3 – 115.5 MB/s**. The best bandwidth was with 64 KB block size. There is some decreasing of bandwidth with 4 KB block size.

Table XXXII Three concurrent VM summary dd test

	MB/s			
	4KB	64KB	256KB	1M
MIN.	6.2	35.3	43	54.4
MAX.	110.2	195.8	120.2	144.4
AVG.	39.3	115.5	77.1	96.9

The average bandwidth was **increased by 64%** (this approximation is flowing from comparison of values on 64 KB blocks).

This report has shown the trend of lower speedup of the cluster data storage bandwidth addicted on the increasing of the number of concurrent machines. The high among of concurrent machines produce more overhead operation and perform smaller network communication frames onto local network.

H. Four VM on cluster

The Table XXXIII shown the reported values of bandwidth from separated node scenario. The average bandwidth of four concurrently writing machines was **18.9 – 71.9 MB/s**. The best performance was with 64 KB block size.

[6] Tests described in chapter III

Table XXXIII Four concurrent VM summary dd test Separated

	MB/s			
	4KB	64KB	256KB	1M
MIN.	5.1	48.0	40.3	55.0
MAX.	53.6	98.3	68.9	85.0
AVG.	18.9	71.9	47.7	60.1

The Table XXXIV shown the scenario with all nodes in one small isolated network. The average bandwidth was **28.4 – 108.2 MB/s**. The best bandwidth was with 64 KB block size.

Table XXXIV Four concurrent VM summary dd test

	MB/s			
	4KB	64KB	256KB	1M
MIN.	7.9	60.6	57.2	88
MAX.	216.4	163.4	115	133
AVG.	28.4	108.2	78.3	102.1

The average bandwidth was **increased by 50%** (this approximation is flowing from comparison of values on 64 KB blocks).

I. Storage bandwidth comparison

The next tables shown the performance differences between separated and isolated network solution. The first table represents the best performance for the separated solution with block size 64 KB. The bandwidth was **39.1 – 71.9 MB/s**.

Table XXXV Four concurrent VM summary dd test Separated

VMs/ block size	MB/s			
	4 KB	64 KB	256 KB	1M
1	22.1	**39.1**	28.6	30.6
2	33.4	**65.5**	43.7	51.7
3	26.3	**70.5**	46.4	57.2
4	18.9	**71.9**	47.7	60.1

The Table XXXVI reports the result of performance tests with isolated network model. The best performance was measured with 64 KB block size. The best bandwidth was **130.1 MB/s**.

Table XXXVI Four concurrent VM summary dd test

VMs/ block size	MB/s			
	4 KB	64 KB	256 KB	1M
1	90.0	**130.1**	108.8	119.4
2	57.2	**118.5**	75.5	96.7
3	39.3	**115.5**	77.1	96.9
4	28.4	**108.2**	78.3	102.1

J. Increasing of storage disk bandwidth

This experiment shows that the CEPH is more efficient with multiple accesses to the data cluster. The increasing of bandwidth is smaller while increasing the number of concurrent machines on the cluster. This result was predictable because increasing of concurrent machine brings the increasing of overhead operations. The cluster must provide more information about saving data. The second purpose of this asymmetric increasing is flowing from fragmentation of data. The test results showed the potential bottleneck in small block size.

The solution with 2 concurrent writing machines provides the average bandwidth with 4 KB block 57.2 MB/s. The scenario with 3 concurrent machines provided only 39.3 MB/s. This is the effect of the scenario operations with the small data blocks. The distributed file system must have bigger block size than 4 KB; as is flowing from these tests, the best solution will be 64KB block size.

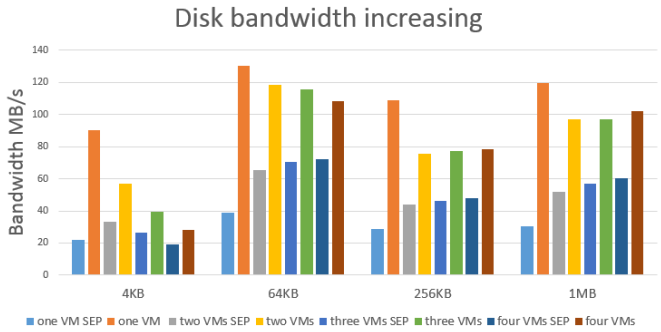

Fig. 23 Disk bandwidth increasing comparison

The next interesting result was in the part of comparison 64 KB block size in isolated scenario. The decreasing of bandwidth with the increasing number of virtual machines indicates that the bandwidth **fluctuate from 130.1 to 108.2 MB/s.** The same effect in separate scenario was significantly higher; the bandwidth **fluctuate from 71.9 to 39.1 MB/s**.

VIII. Recommendation of Implementation

The first recommendation is flowing from the tests of local storage infrastructure. The local storage bandwidth of each node must be significantly higher than the network connection to the node. The test of local storage bandwidth shows, the diametric different values of test with common SATA II drives in raid 1 and server connected to corporate storage over Ethernet (10 Gbit/s line SAN). The potential minimum value of local storage bandwidth will be over 100 MB/s; for the 1 Gbit/s network connection.

The second implementation recommendation is flowing from the local network stability and bandwidth. The sharing of network infrastructure (subnet, physical ports) with common network application is not a good idea. The common application might affect the storage bandwidth significantly. The firewall between cluster nodes is the bottleneck of many solutions. The cisco 6500 family firewall is between subnets in this scenario. It is powerful firewall, but it is shared with all network infrastructures. To minimize this issue, it is

recommended to isolate the storage network to single vlan over all network infrastructures. The nodes will be connected by the 10 Gbit/s ethernet with dedicated ports. The effect of completely isolated network only for the data cluster is described in the part VII of this paper. The stability of the network and network bandwidth play important role in the implementation of the cluster.

The last recommendation is flowing from these tests is about the CEPH-client (the server with VM stored on the CEPH storage). The model shown on Fig. 12 shows the main bottleneck of this solution. The main problem is in network connection. It is not possible to share one 1 Gbit/s network connection with physical machine and many virtual machines. This recommendation depends on the main purpose of the hosted virtual machine. For many websites with minimal storage requirement it will be without any problem, the storage bandwidth is sufficient for these solutions. But implementation with higher storage bandwidth requirements need network line with higher bandwidth (10 Gbit/s or multiple 1 Gbit/s lines).

IX. Conclusion

This paper contains test realized with the geographically separated nodes of the CEPH storage cluster. The main parts of infrastructure were tested in the chapter III-VI.

The main part of this report reflects the real usability of the data cluster stored in different location. The purpose of this isolation is flowing from the data security requirement. The data must be accessible when the part of infrastructure is down. The solution has specific issues described in this paper. The main issue is flowing from the network connectivity. The theoretical connection is realized by the 1 Gbit/s ethernet; but tests show, that the theoretical bandwidth of this solution is significantly higher than practical reports. These reports inspect the possible bottlenecks in firewall between network subnets and not stable latency. The next described issue is in sharing of this infrastructure with common network application.

The final part of this paper is dedicated to comparison of the separated solution with strictly local isolated solution; the local solution might represents the highest possible performance with these HW parts. But the isolated scenario isn't be implement in real infrastructure. There must be an entry point to the infrastructure and the network must be particularly shared with other devices.

The future work on this testing procedure will be in testing the cluster on hybrid network schema. The nodes will used the isolated network. The monitor node will be the entry point to the infrastructure. The bandwidth probably decrease; but the question will be; how much.

References

[1] Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI '06) November 6-8, 2006, Seattle, WA, USA. New York, N.Y, 2006. ISBN 19-319-7147-1.W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[2] NEMETH, Evi. UNIX and Linux system administration handbook. 4th ed. Upper Saddle River, NJ: Prentice Hall, c2011. ISBN 01-314-8005-7.

[3] CHANG, Bao Rong, Hsiu-Fen TSAI a Chi-Ming CHEN. Mathematical Problems in Engineering [online]. 2013, vol. 2013, s. 1-11 DOI: 10.1155/2013/947234.

[4] HIROFUCHI, Takahiro, Mauricio TSUGAWA, Hidemoto NAKADA, Tomohiro KUDOH. A WAN-Optimized Live Storage Migration Mechanism toward Virtual Machine Evacuation upon Severe Disasters. IEICE Transactions on Information and Systems [online]. 2013, E96.D, issue 12, s. 2663-2674. DOI: 10.1587/transinf.E96.D.2663.

[5] PETERSON, Larry L a Bruce S DAVIE. Computer networks: a systems approach. 5th ed. Burlington: Morgan Kaufmann, c1999, xxxi, 884 s. ISBN 978-0-12-385059-1.

[6] PETERSON, Larry L a Bruce S DAVIE. 2012 IEEE Globecom Workshops 3-7 December 2012, Anaheim, Ca, USA: a systems approach. 5th ed. Piscataway, N.J.: IEEE Computer Society, 2012, xxxi, 884 s. ISBN 978-146-7349-413.

[7] PALMER, Michael J. Guide to UNIX using Linux. 4th ed. Australia: Thomson/Course Technology, c2008, xx, 697 p. ISBN 14-188-3723-7.

[8] B. Djordjevic, V. Timcenko, Ext4 File System performance Analysis in Linux Environment, 11th WSEAS International Conference on APPLIED INFORMATICS AND COMMUNICATIONS (AIC '11), Florence, Italy, August 23,25, 2011. ISBN 978-1-61804-028-2

[9] M. LNENICKA, J. KOMARKOVA, E. MILKOVA. Performance Testing of Cloud Storage while Using Spatial Data. Latest trends in applied informatics and computing: proceedings of the 3rd international conference on applied informatics and computing theory (AICT'12) : Barcelona, Spain, October 17-19, 2012. 2012, p. 254-258. ISBN 978-1-61804-130-2.

[10] Medina V., Garcia J.M., Live replication of Virtual Machines, 10th WSEAS International Conference on Software Engineering, Parallel and Distributed Systems, SEPADS'11, Cambridge; United Kingdom, 2011, ISBN 978-960474277-6.

[11] Suba P., Horalek J., Hatas M., Comparison of technologies for software virtualization, 11th WSEAS International Conference on APPLIED INFORMATICS AND COMMUNICATIONS (AIC '11), Florence, Italy, August 23,25, 2011. ISBN 978-1-61804-028-2.

Ing. David Malanik, Ph.D.
Born in Zlin, Czech Republic, 1. March 1984
Bacelour degree in Information technology 2006, Tomas Bata University in Zlin. Master degree in Information technology 2008, Tomas Bata University in Zlin. Ph.D. Thesis based on the user identification provided by the neural network, 2011, Tomas Bata University in Zlin
Presently SENIOR LECTURER at Tomas Bata University in Zlin, Department of Informatics and Artificial Intelligence. Main specialization: computer security, computer viruses, penetration testing, artificial neural network, computer networks.