仮想サーバ環境における
iSCSIネットワークの利用について
～10GbEと1GbEの混在環境～

2013年1月28日　第一版

Japan Data Storage Forum
ストレージネットワーキング技術部会　iSCSI分科会

背景

近年、情報システムにおけるデータ量は増大の一途をたどっており、その対応としてストレージシステムの最適化、効率化は急務となっています。一方でデーターセンタを中心に普及している「ファイバチャネル」を使用したストレージ専用ネットワーク(SAN)の導入は、いまだコスト面や構築運用に必要とされる技術などを考慮すると中規模以上の情報システムなどに適していると言わざるをえません。このような状況から、小中規模システムを中心に近年導入が増加傾向にある「iSCSI」を使用したストレージ・ネットワークが注目されています。

iSCSIは従来のEthernet(LAN)を使用してストレージ装置とサーバなどを接続する技術ですが、その接続形態などからサーバ仮想化との親和性が高く、両技術を組み合わせてシステム導入されるケースも増えてきています。iSCSIの導入に関しては、『LAN経由で多ノードからストレージを使用した場合どのようなアクセス性能になるか』などがシステム構築のポイントとなりますが、現時点ではシステム設計のノウハウが書かれたガイドラインなど、参考となる資料は少ない状況と言えます。

目的

前述の背景より、JDSFとして一般的なiSCSIによるストレージ・ネットワーク構築の参考となることを期待し、本ドキュメントを公開します。

本ドキュメントにおける目的は以下になります。

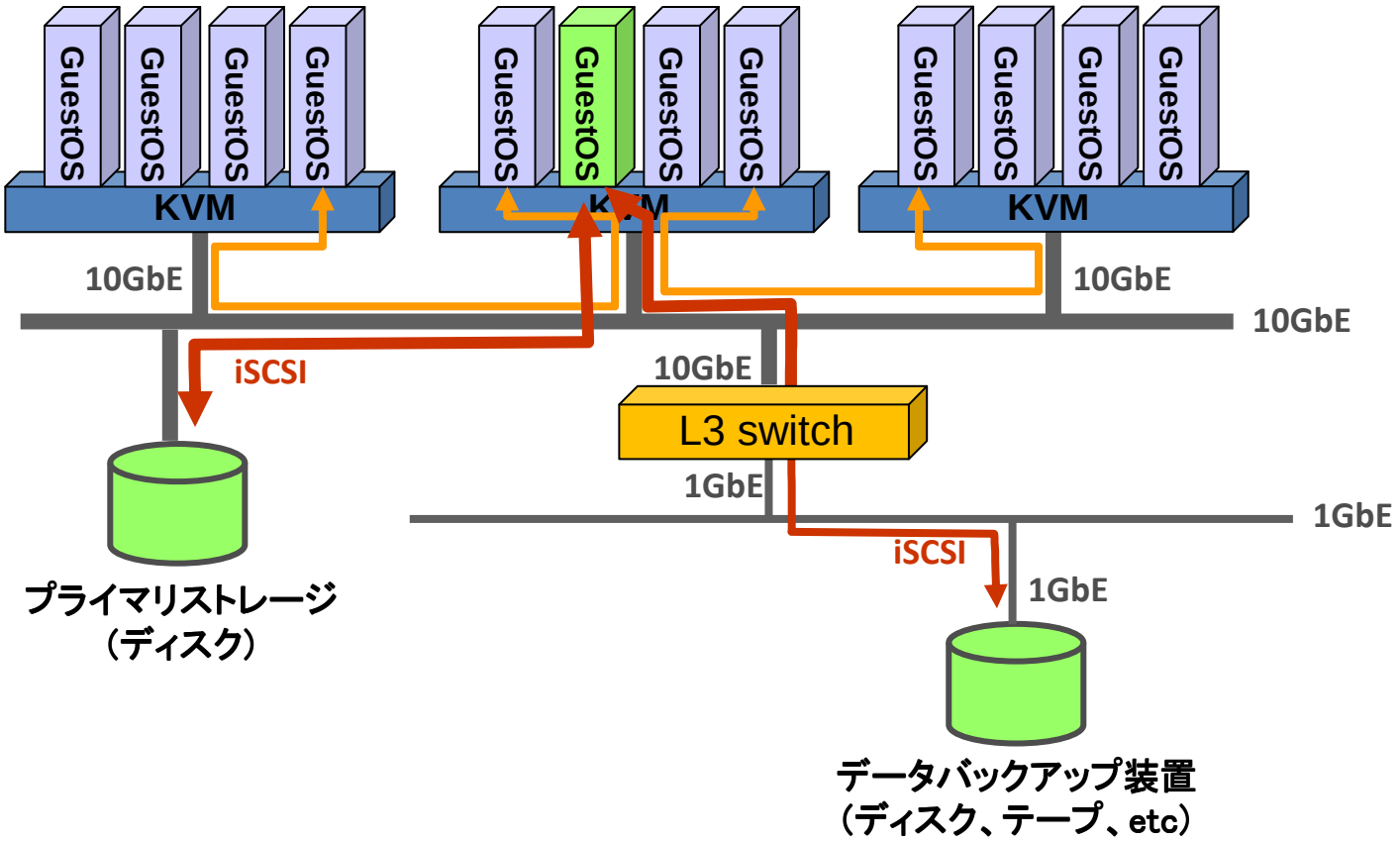
1) サーバ仮想化環境をiSCSI接続のストレージで構築すること
2) 仮想化システム上に構築されたOS上からストレージへのI/O性能の測定
3) iSCSI接続構成におけるネットワーク上の負荷状態の測定

この様な目的を持ち、仮想化環境におけるiSCSIストレージ・ネットワーク構成システムの構築と性能測定を実際に行い、その結果をまとめ、公開します。

前回は、iSCSIのベースとなるEthernetに1Gbit帯域を使用し、性能を測定しました。
今回は、Ethernetの帯域が送出側10Gbit、受信側1Gbitの歪な環境での性能とチューニング方法について市販バックアップソフトウェアでのI/Oを使って確認しました。

想定するシステム運用イメージ例

本検証を実施するにあたり、想定するユーザーシステムの運用イメージ例



検証用サーバスペック

	検証機＃1	検証機＃2
マザーボード	Intel Workstation Board S5520SC	Intel Server Board S1200BTL
CHIPSET	Intel 5520 Chipset Intel 82801JIR I/O Controller Hub (ICH10R)	Intel C204 Platform Controller Hub (PCH) chipset ServerEngines LLC Pilot III BMC controller
CPU	Intel Xeon Processor E5520 (8M Cache, 2.26 GHz, 5.86 GT/s Intel QPI)	Intel Xeon Processor E3-1240 (8M Cache, 3.30 GHz, 5GT/s DMI)
MEMORY	12GB (2GB 1333MHz DDR3 DIMM x6)	16GB (4GB 1333MHz DDR3 ECC CL9 DIMM x4)
LAN	オンボード(Broadcom 82575EB x2)	オンボード(Broadcom82574Lx1, 82579x1)
	Emulex OCe10102 x2 Broadcom x2	Emulex OCe10102 x2 Broadcom x2
HDD (OS)	Hitachi HDS721010CLA332 1TB (SATA2,7200rpm)	WD RE4 50 GB (SATA2,7200rpm)
(DATA)	Hitachi HDS721010CLA332 1TB x3(RAID0)	Intel SSD 520x3(RAID0)
RAID CARD	LSI Logic 1086	MegaRAID SAS 9260-4i
PCI SLOT	Slot1: 5V PCI 32 bit/33 MHz Slot2: PCI Express Gen1 x4 Slot3: PCI Express Gen1 x1 Slot4: PCI Express Gen2 x16 Slot5: PCI Express Gen2 x16	Slot1: 5V PCI 32 bit/33 MHz Slot2: PCI Express Gen2 x8 Slot3: PCI Express Gen2 x8 Slot4: PCI Express Gen2 x8 Slot5: PCI Express Gen2 x16

サーバソフト　バージョン情報

Hyper visor

	Test Server #1	Test Server #2
hostname	kvm1	kvm2
OS	CentOS release 6.3 (Final)	CentOS release 6.3 (Final)
kernel version	2.6.32-279.el6.x86_64	2.6.32-279.el6.x86_64
libvirt version	0.9.10	0.9.10

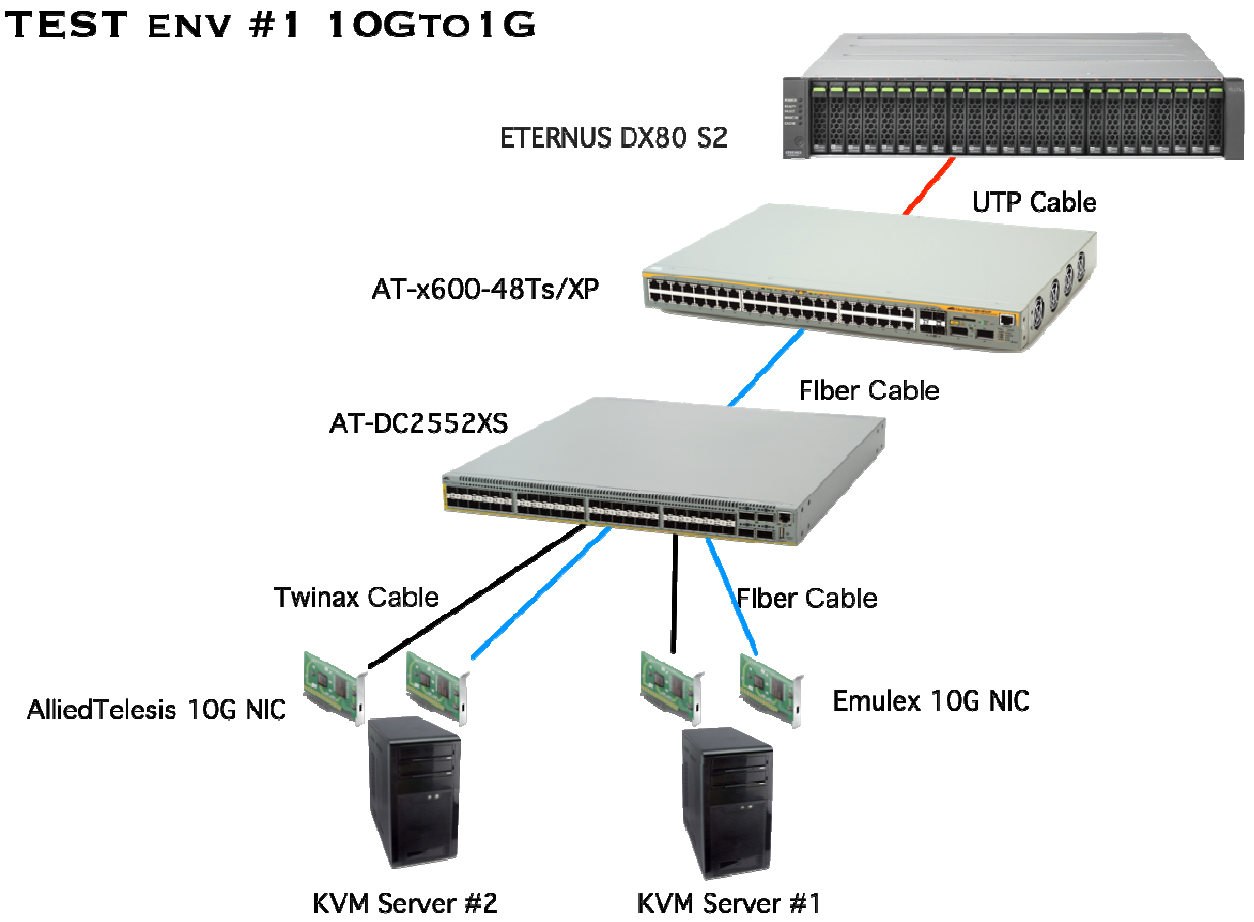
Guest Virtual Machine

Guest VM	vm1-1,vm1-2,vm1-3,vm1-4	vm2-1,vm2-2,vm2-3,vm2-4
OS	CentOS release 6.3 (Final)	
kernel version	2.6.32-279.el6.x86_64	
VNIC Model	e1000 or VirtIO	

検証用ストレージ、スイッチスペック

	AT-DC2552XS		AT-x600-48Ts/XP	
ポート	10/100/1000BASE-T SFP+ スロット(10G) QSFP+ スロット(40G)	× 44 × 48 × 4	10/100/1000BASE-T SFPスロット XFPスロット	× 44 × 4 × 2
スイッチング方式	ストア＆フォワード方式/ カットスルー方式		ストア＆フォワード方式	
最大パケット転送能力	952.38Mpps		136.9Mpps	
スイッチング・ファブリック	1,280Gbps		232Gbps	

検証環境の物理イメージ



検証内容

本検証では、10GbEネットワークに接続された仮想OS(サーバ)およびそのプライマリストレージである10GbE iSCSIストレージから、L3スイッチを通して1GbEネットワークに接続されたバックアップ用iSCSIストレージへバックアップを行う運用を想定した。

バックアップの種類には、大きくファイルバックアップとイメージバックアップがあるが、より高い転送パフォーマンスを期待して、今回はイメージバックアップを選択している。

検証はいくつかのパターンで実施したが、本書では比較のため以下2パターンの検証結果を取り上げる。

【検証パターン1】

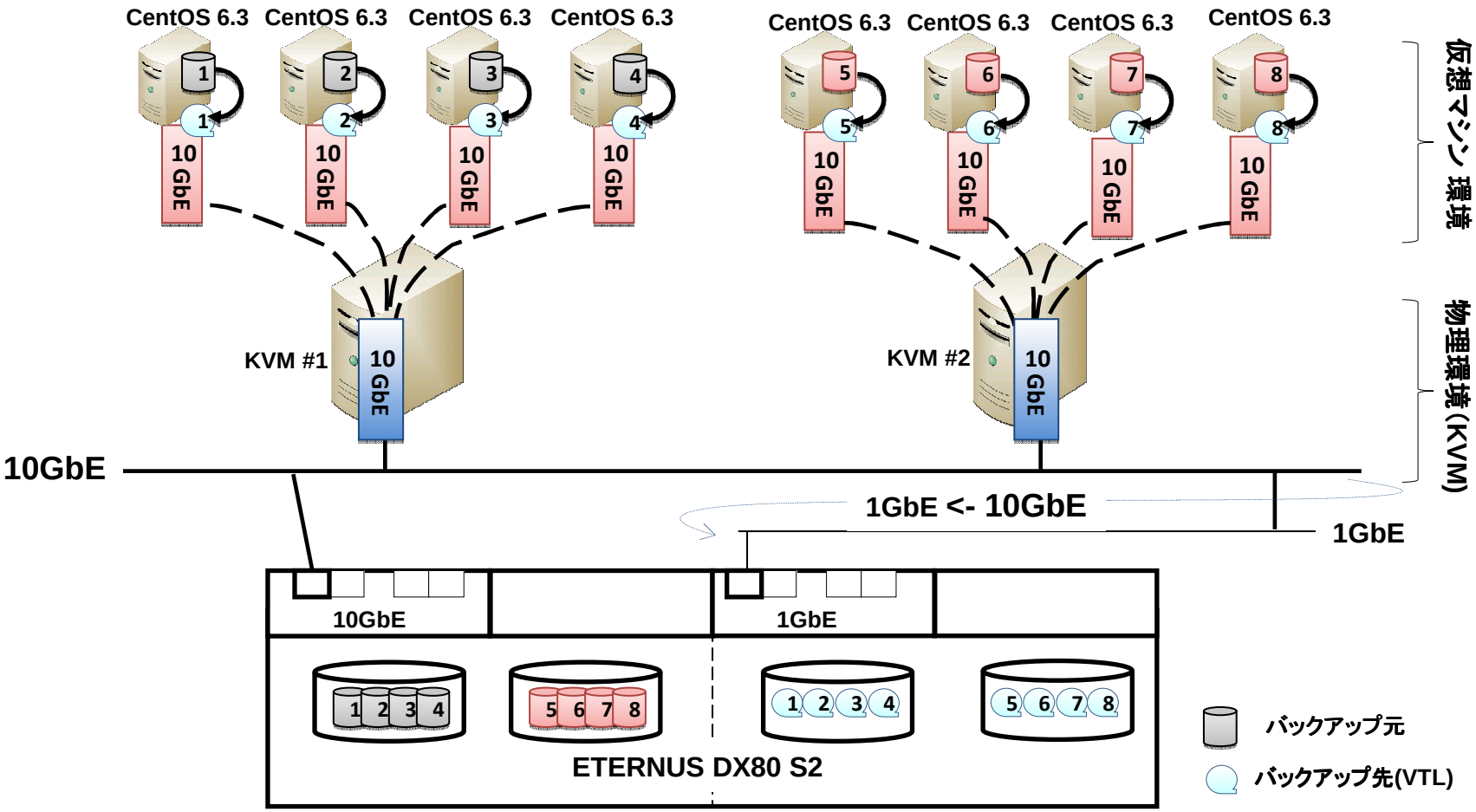
ゲストOSの仮想NICを10GbEに対応する「virtio」とし、1GbEインターフェースを持つiSCSIストレージへバックアップを行い、その処理時間からデータ転送速度を測定した。

【検証パターン2】

ゲストOSの仮想NICを1GbE用の「E1000」とし、検証パターン1と同じ処理を実施、同様に処理時間からデータ転送速度を測定した。

検証パターン1　10Gbpsのまま送出
（仮想10GbE ⇒ 1GbE）

■検証構成1（10GbE – 1GbE）

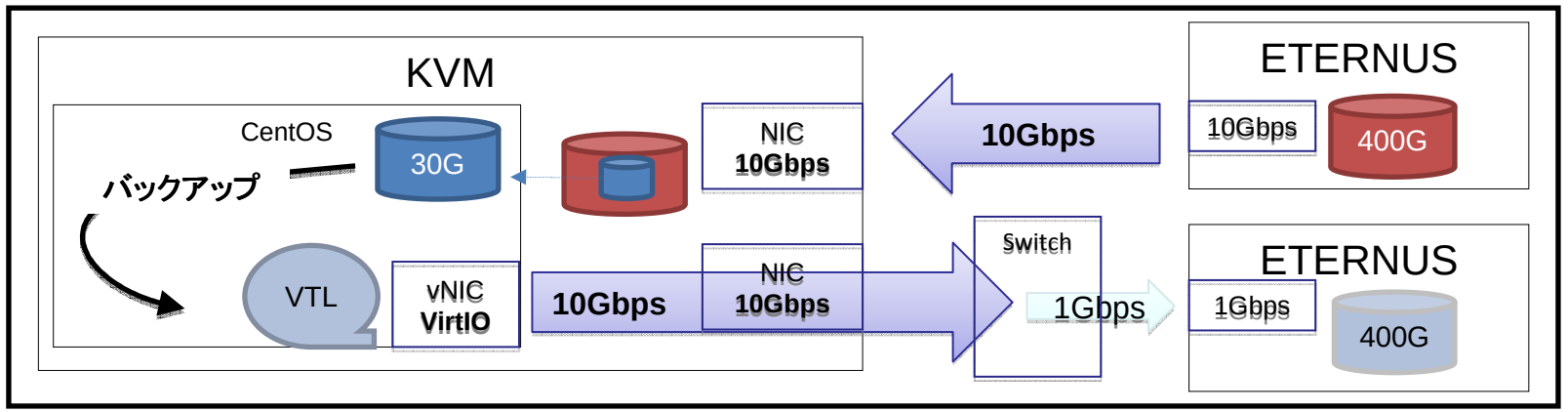


検証パターン1　10Gbsのまま送出
（仮想10GbE ⇒ 1GbE）

■10Gb -> 1Gb　4多重バックアップ実行結果

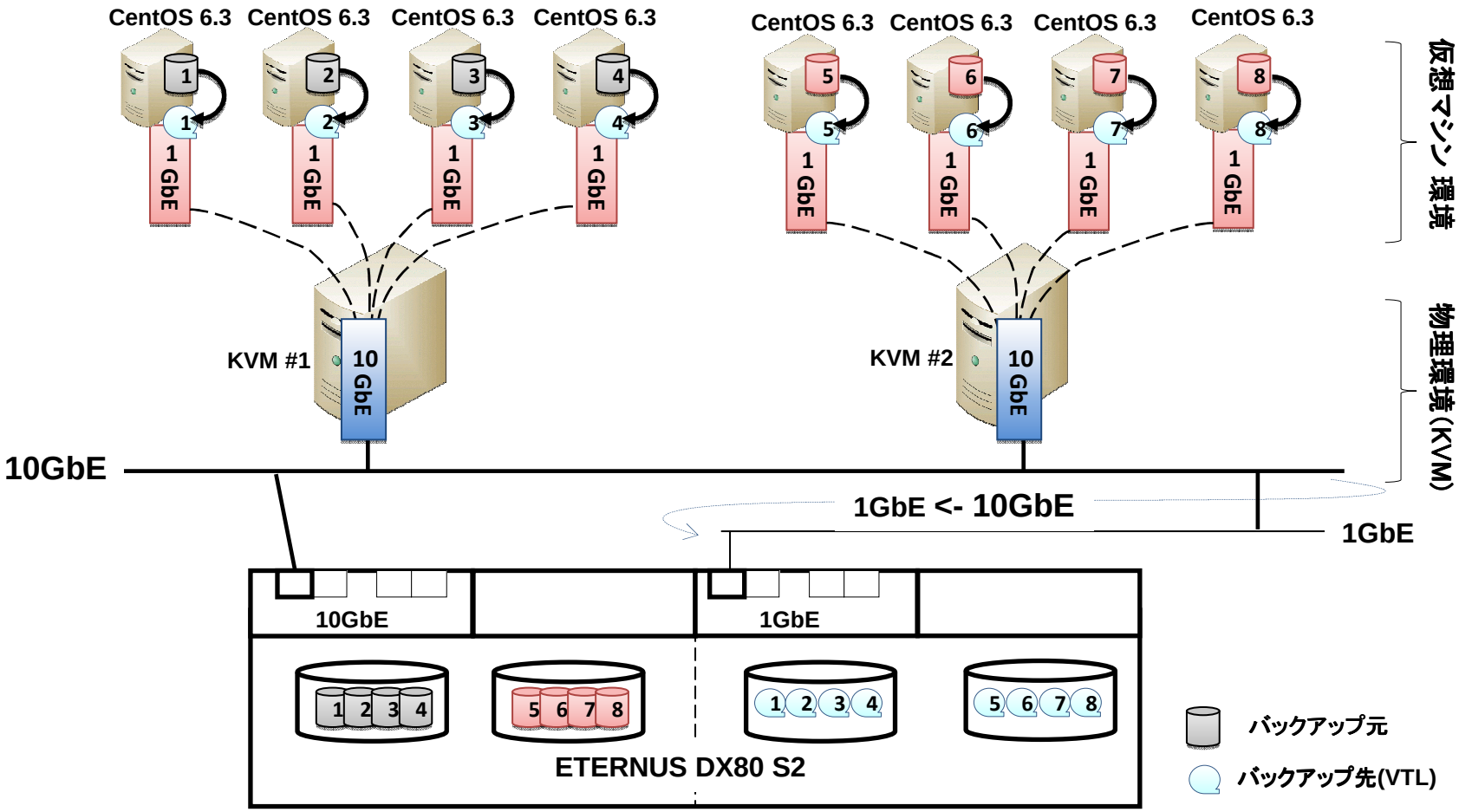
仮想マシン名	バックアップ容量	バックアップ実行時間	転送レート
vm2_1	30 GB	96m 28s	5.31 MB/sec
vm2_2	30 GB	96m 44s	5.29 MB/sec
vm2_3	30 GB	96m 33s	5.30 MB/sec
vm2_4	30 GB	95m 54s	5.34 MB/sec
			計 21.2 MB/sec

■10Gb -> 1Gb 実行環境



検証パターン2　1Gbに絞り送出
（仮想1GbE ⇒ 1GbE）

■検証構成2（1GbE – 1GbE）

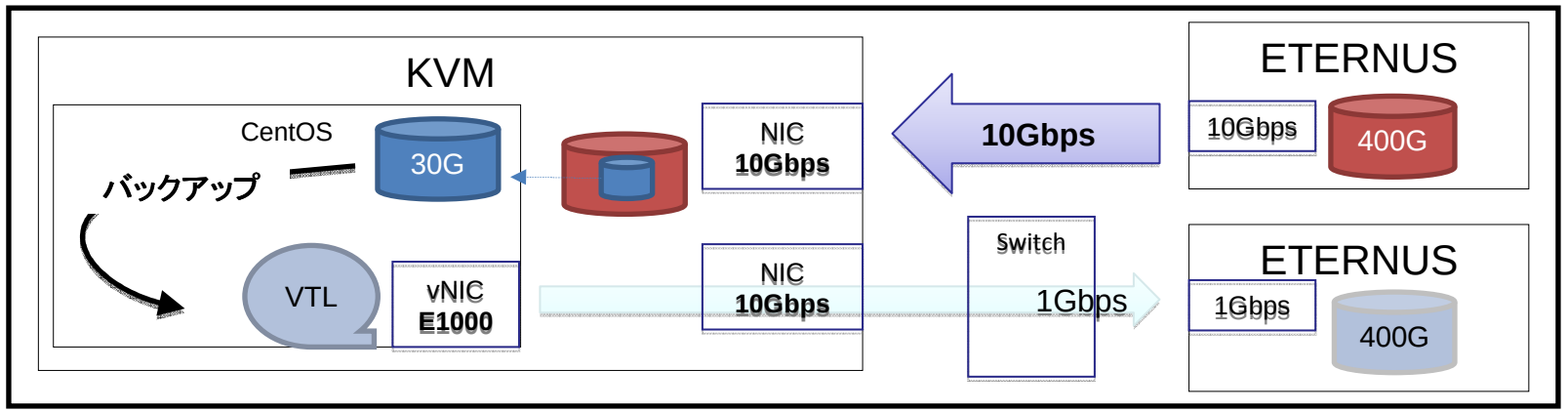


検証パターン2　1Gbに絞り送出
（仮想1GbE ⇒ 1GbE）

■10Gb -> 1Gb　4多重バックアップ実行結果

仮想マシン名	バックアップ容量	バックアップ実行時間	転送レート
vm2_1	30 GB	30m 55s	16.56 MB/sec
vm2_2	30 GB	19m 30s	26.28 MB/sec
vm2_3	30 GB	19m 12s	26.62 MB/sec
vm2_4	30 GB	19m 30s	26.26 MB/sec
			計 95.7 MB/sec

■10Gb -> 1Gb 実行環境



総括

	仮想NIC	トータルスループット ＝ 帯域利用率	バックアップ元	バックアップ先
検証パターン1	virtIO (準仮想化NIC)	21.2MB/sec = 21.2%	仮想ディスク （KVMにマウント）	iSCSIディスク （PCIパススルー）
検証パターン2	e1000 (Intel PRO/1000)	95.7MB/sec = 95.7%		

今回、10GbEのサーバから1GbEのストレージへデータを送出するという適正化されていない環境において、どのようなバックアップ性能が得られるかを目的に検証を行いました。

このようなネットワーク帯域に差異がある環境においては、帯域差の発生地点においてバッファあふれが発生し、大量のパケット破棄やそれに伴う再送処理による輻輳が発生することが一般に知られていますが、本検証においてもそれが確認されました。

LinuxゲストOSにおける仮想NICとしてはvirtIOドライバが最適なものとして広く使用されていますが、今回のような環境では1GbE NICのエミュレーションドライバを使用して仮想NICの利用帯域をあえて抑えることで輻輳が抑制され、virtIOドライバを使用した環境よりも良好なパフォーマンスが得られました。

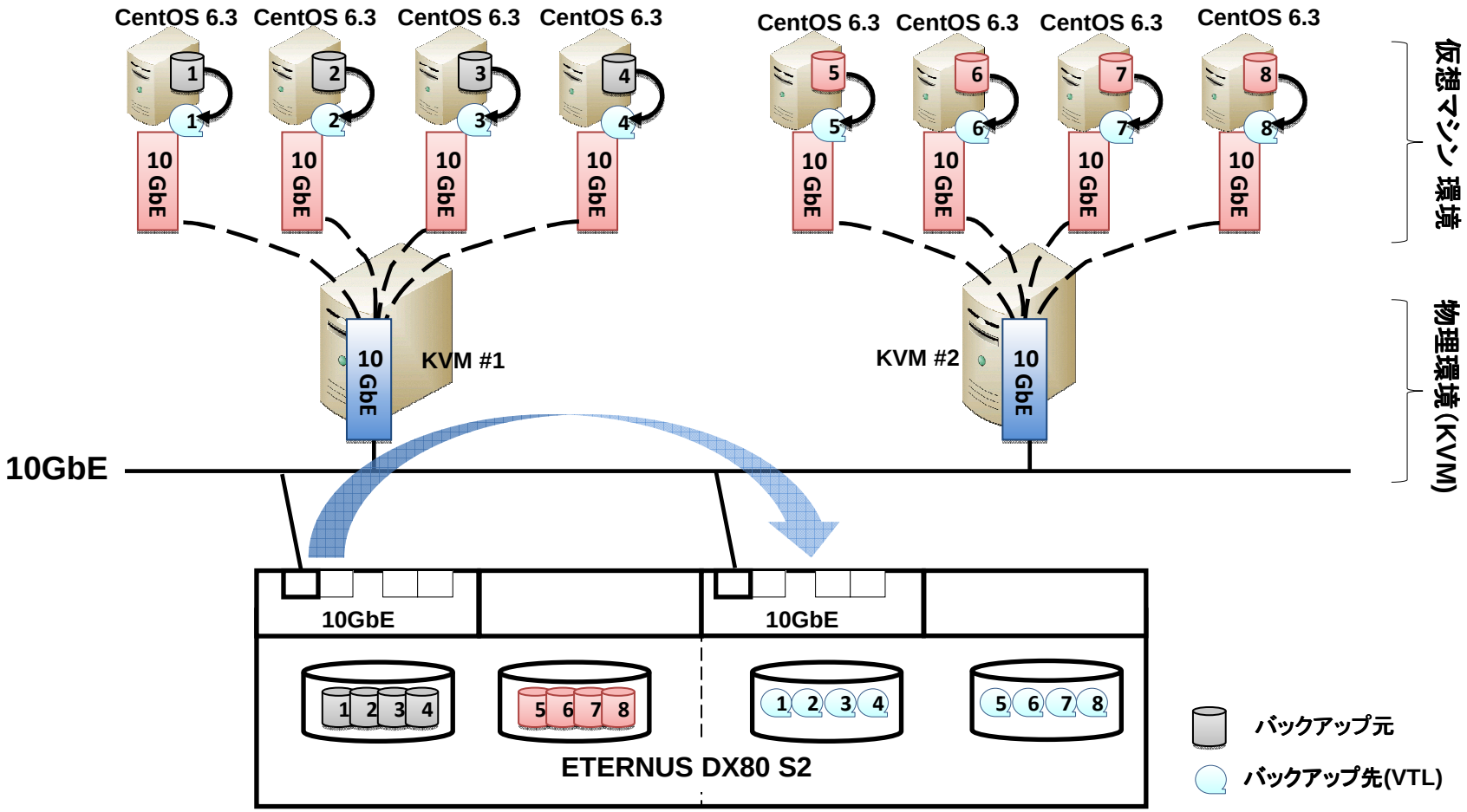
われわれの検証結果が、皆様方のシステム構築時の目安としてご活用いただけると幸いです。

Appendix

Appendix 1

参考：検証パターン3（10GbE ⇒ 10GbE）

■検証構成3（10GbE – 10GbE）

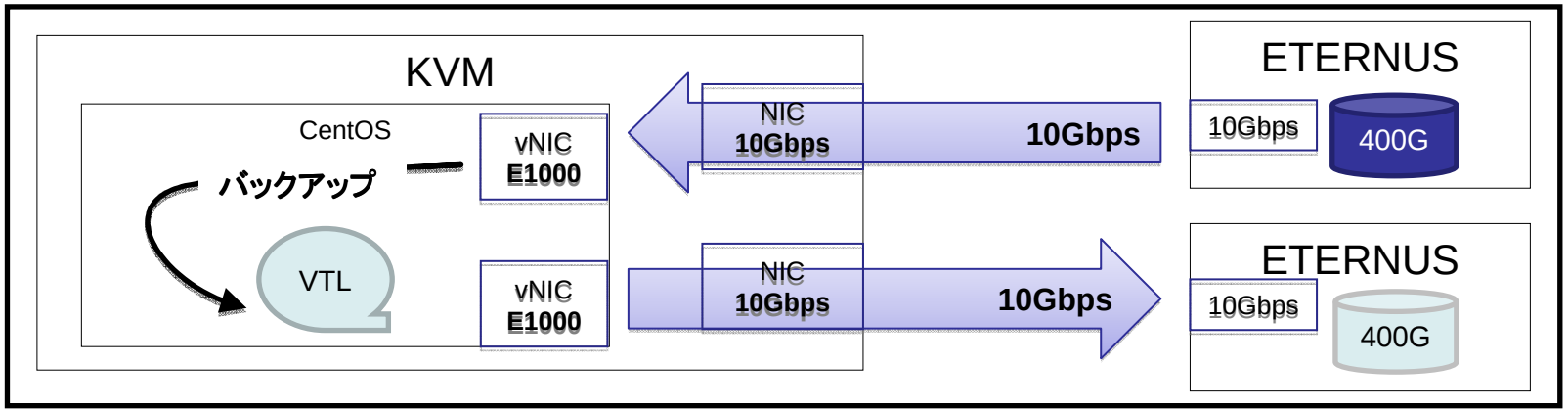


Appendix 1
参考：検証パターン3（10GbE ⇒ 10GbE）

○ 10G -> 10G 実行結果

仮想マシン名	バックアップ容量	バックアップ実行時間	転送レート
vm2_1	100.99 GB	26m 12s	65.87 MB/sec
vm2_2	101.00 GB	16m 35s	104.06 MB/sec
vm2_3	101.00 GB	17m 25s	99.06 MB/sec
vm2_4	100.99 GB	28m 56s	59.57 MB/sec
			計 328.56 MB/sec

○ 10G -> 10G 実行環境　　－ ゲストOSからのiSCSIイニシエータ接続 －



Appendix　2

参考：本環境における10GbEthernetの通信能力を測定

KVM Server間での対向の通信能力を測定

iperfによる測定　**9.3Gbps**

```
[root@kvm2 ~]# iperf -c 192.168.2.20   -P 2
------------------------------------------------------------
Client connecting to 192.168.2.20, TCP port 5001
TCP window size: 96.4 KByte (default)
------------------------------------------------------------
[  3] local 192.168.2.25 port 39715 connected with 192.168.2.20 port 5001
[  4] local 192.168.2.25 port 39714 connected with 192.168.2.20 port 5001
[ ID] Interval       Transfer     Bandwidth
[  3]  0.0-10.0 sec  5.39 GBytes  4.63 Gbits/sec
[ ID] Interval       Transfer     Bandwidth
[  4]  0.0-10.0 sec  5.45 GBytes  4.68 Gbits/sec
[SUM]  0.0-10.0 sec  10.8 GBytes  9.31 Gbits/sec
```

参加メンバー

アライドテレシス株式会社　延原 英棋 (iSCSI分科会会長)

富士通株式会社　齊藤 金弥 (SNT部会副部会長)

株式会社 日立製作所　須賀田 勉 (SNT部会副部会長)

株式会社　アーク・システムマネジメント　日吉 孝浩

株式会社　マーク・システムマネジメント　大串 将史

アライドテレシス株式会社　仁木 秀和

富士通株式会社　伊藤　佳治