Vol.2014-IOT-26 No.9
2014/6/28

大阪大学の仮想化基盤における Software Defined Storageの評価実験

柏崎 礼生[1,a)]　宮永 勢次[1,b)]　森原 一郎[1,c)]

概要：大阪大学では 2010 年にサーバ集約を目的として，VMware 社の VMware ESX 4.0 をハイパーバイザとして利用する仮想化基盤を構築し，現在 2014 年度にこの基盤の増強を行っている．この増強の目的の一つとして，VMware 社がリリースした Software Defined Storage 基盤である Virtual SAN (VSAN) の性能評価をすることが挙げられる．本稿では VSAN のインパクトについて解説し，その評価環境と一部の評価結果を示す．

キーワード：software defined storage, 仮想化基盤

An Evaluation of a Software Defined Storage on a Virtualization Infrastracture in Osaka University

Hiroki Kashiwazaki[1,a)]　Seiji Miyanaga[1,b)]　Ichirou Morihara[1,c)]

Abstract: A virtualized infrastracture in Osaka University started its service to aggreagte server machines inside the university in March 2010. Now, the infrastracture is being enhanced on 2014. One of the purpose of this enforcement is to evaluate perfomances of a software defined storage infrastracture “Virtual SAN” (VSAN) released by VMware Inc. In this paper, we show detail impacts of VSAN, an environment of evaluation and a result of evaluations.

Keywords: software defined storage, virtualization infrastracture

1. はじめに

組織が持つ情報は組織外部への公開が不可能な秘匿情報とそれ以外とに分類される．情報の保有を自組織以外に委託する場合，セキュリティなど組織が要求する情報のコントロール品質を実現する事が可能であるかどうかが評価基準となる．委託先が要求品質を実現可能である場合，委託費用と独自構築・運用維持費用を比較する事で委託の可否を判断できる．委託先が要求品質を実現可能でない場合，秘匿情報が漏洩するリスクと漏洩確率から漏洩による損害期待値を求め，その値と委託費用の合計値と独自構築・運用維持費用を比較することで委託の可否を判断できる．秘匿情報の重要度を複数に分類した場合も前述の算出方法で委託の可否を判断できる．秘匿情報を分類することや，漏洩リスク計算をできない場合，全ての情報資産を自組織で保有するという判断が下される．

経営の合理化を目指す組織は情報通信技術 (Information and Communication Technology: ICT) への投資を効率化しようとする．組織内に計算機が散在しており，しかもそれらが常時稼働している場合，計算機サーバ集約のための仮想化技術の利活用が積極的に行われ，自組織内で完結した仮想化基盤やプライベートクラウドコンピューティング環境の構築が選択される．仮想化ハイパーバイザが提供する仮想計算機 (Virtual Machine: VM) の仮想化オーバーヘッドが大きいと評価された時期においては，物理 CPU

[1] 大阪大学
Osaka University
[a)] reo@cmc.osaka-u.ac.jp
[b)] miyanaga-s@office.osaka-u.ac.jp
[c)] morihara@cmc.osaka-u.ac.jp

が提供する物理コア 1 つに対して仮想 CPU を 1 つ割り当てる設計が行われた．仮想化オーバーヘッドが十分に小さいと判断され，また仮想 CPU の利用率が低い VM が多い環境においては物理コア 1 つに対して複数の仮想 CPU を割り当てる設計も行われるようになった．

このような設計が行われると，仮想化ハイパーバイザが動作する仮想化ホスト上に割り当てられる VM の数は増大する．年々システムが扱うデータは単調増加するため，VM が発生させる IO 要求もまた増大する．そのため仮想化基盤やクラウドコンピューティングシステムが要求する単位時間あたりの IOPS もまた増大し，システムのパフォーマンスはストレージシステムの部分でボトルネックが発生しやすくなる [1]．CPU のコア数の増加やプロセスルールの微細化は現在順調に発展しており，メモリも容量あたりの単価は減少している．しかしストレージにおいて，ディスクの回転数は早い段階で頭打ちになり，プラッタあたりの容量の増大も鈍い．そのためストレージシステムとして性能を向上させるためには並列アクセスによる高速化が効果的である．それと同時に耐障害性を向上させるためにさらにディスクを増やす必要がある．

ディスクストレージの性能向上の鈍化に対して，より高いパフォーマンスのストレージが要求されるようになると，この問題解決として半導体素子メモリを用いたドライブが登場した．半導体素子メモリを用いた記憶装置はディスクより高いパフォーマンスを示すかわりに，単位記憶容量あたりの価格は高価である．半導体素子メモリのみで構築されたドライブは高いパフォーマンスを示すが，極めて高額となる．効率的な ICT 投資のためには，システムがストレージに要求するパフォーマンスと，ストレージが提供するパフォーマンスとが釣り合うことが求められる．これまでストレージベンダーは半導体素子メモリドライブが提供する領域をキャッシュとして利用し，ディスクアレイとのハイブリッドで構成されたストレージシステムを提供することで価格的にも性能的にも半導体素子メモリストレージとディスクアレイストレージの中間の製品を提供してきた．

2. VMware Virtual SAN (VSAN)

VMware 社 [*1] が 2014 年 3 月に General Availability をアナウンスしたソフトウェアである Virtual SAN (VSAN) は，VMware 社曰く Software-defined storage であるとされている [*2]．VMware 社 CTO の Richard McDougall は VSAN について計算機資源とストレージを converge したモデルであると述べている [*3] (図 1)．

ストレージベンダーの業界団体である The Storage Networking Industry Association (SNIA)[*4] は Software Defined Storage に関する White Paper をリリースしている [*5]．これによると Software defined storage という用語は Software Defined Network (SDN) に肖ったマーケッティング用のバズワードであるとしながらも，下記のような特徴 (属性) を持つとされている．

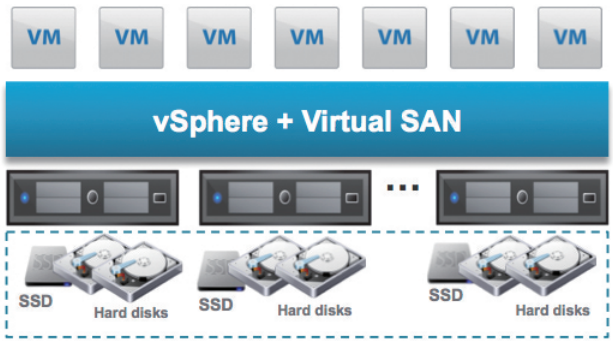

図 1　VSAN の模式図 (Office of the CTO の記事より)
Fig. 1　A diagram of VSAN

- May allow customers to “build it themselves,” providing their own commodity hardware to create a solution with the provided software.
- May work with either arbitrary hardware or may also enhance the existing functions of specialized hardware.
- May also enable the scale-out of storage (not just the scale up typical of big storage boxes).
- Nearly always includes the pooling of storage and other resources.
- May allow for the building of the storage and data services “solution” incrementally.
- Incorporates management automation.
- Includes a self service interface for users.
- Includes a form of service level management that allows for the tagging of metadata to drive the type of storage and data services applied. The granularity may be large to start, but is expected to move to a finer grained service level capability over time.
- Allows administrators to set policy for managing the storage and data services.
- Enables the dis-aggregation of storage and data services.

VMware 社が VSAN を Software defined storage であると位置づける理由については語られていないが，VSAN が

[*1] http://www.vmware.com
[*2] http://www.vmware.com/products/virtual-san/features.html
[*3] http://cto.vmware.com/the-dawn-of-vsan/
[*4] http://snia.org
[*5] http://snia.org/sites/default/files/SNIA\%20Software\%20Defined\%20Storage\%20White\%20Paper-\%20v1.0k-DRAFT.pdf

提供するソフトウェアにより，一般的な x86 サーバ上でスケールアウト可能なストレージシステムを構築することができることから Software defined storage と位置づけられていることが推測できる．

EMC 社 [*6] や NetApp 社 [*7] といったストレージベンダーが提供するストレージ製品は独自のハードウェアに独自の OS を搭載した製品であり，仮想化基盤やクラウドコンピューティング環境において CPU やメモリを提供する計算機リソース部とは区別されるストレージシステムが存在する．そのため計算機リソース部とストレージシステムのインターフェイスもまたボトルネックとなり得る．一方 VSAN は，計算機リソース部が持つ CPU，メモリ，そしてストレージ装置を使うため，構成としてシンプルになり，物理的な占有体積，消費する電力量の観点からも合理性に富む．Fusion-io[*8] などが提供する PCI Express に直接接続する半導体素子メモリドライブは，この潮流を一層加速していると言える．

しかしながら本稿執筆時点において，VSAN などの Software Defined Storage 製品が EMC 社や NetApp 社などが提供する独自ハードウェアによるストレージ製品より優れているかどうか判断するための材料となる定量的な評価は乏しい．VSAN は 3 台以上の仮想化ホストによる構成で動作し，各々の仮想化ホストは 6GB 以上のメモリを搭載していることが要求される．また各々のホストは 1Gbps 以上の帯域を持つネットワークで接続されることが要求される．また 3 台以上の仮想化ホストにおいて半導体素子メモリドライブが接続されていることが要求される．ドライブを接続するためのストレージコントローラは pass-through あるいは RAID0 をサポートすることが要求される (pass-through が推奨される)[*9] ．これらの要件はそれほど高額な機器を要求しないが，下記のような性能に関する疑問点が浮上する．

- 仮想化ホスト数 $N > 3$ の時に，半導体素子メモリドライブを具備する仮想化ホスト上に配置された VM から VSAN が提供するストレージプールに対する I/O を発生させた時と，半導体素子メモリドライブを具備しない仮想化ホスト上に配置された VM から I/O を発生させた時のパフォーマンス差．
- I/O アクセスパターンの違いによるパフォーマンスの差．
- 仮想化ホストの障害時におけるパフォーマンスの劣化度合い．

VSAN では仮想化ホストによりストレージプールが構築されるため，仮想化ホストのハイパーバイザソフトウェアのメンテナンスを考慮にいれる必要がある．VMware ESXi は再起動を必要とするアップデートの数は (例えば Microsoft Windows Server と比較すると) 少ないと主張されるが，EMC 社や NetApp 社などのストレージの製品において再起動を必要とするアップデートがどの程度の頻度で発生するかと比較することが合理的な比較であると考えられる．こういった性能上・運用上の問題点を解決するために，現在増強を進めている仮想化基盤を用いて VSAN の検証を行うこととした．

[*6] http://www.emc.com
[*7] http://www.netapp.com
[*8] http://www.fusionio.com
[*9] http://www.vmware.com/files/pdf/products/vsan/VMware-TMD-Virtual-SAN-Hardware-Guidance.pdf

3. 検証環境

3.1 仮想化基盤の拡張

大阪大学では ICT 投資効率化という大義名分のもとに，大学全体として業務フロー全体の最適化を行い，業務の効率化を目指すという目標が掲げられた．大阪大学程度の規模の総合大学では，部局ごとに独自の ICT 投資が行われ，事務業務フローも部局独自で構築されるケースがある．大学全体を俯瞰すると ICT 投資が分散し，業務改革も局所的な最適化に留まり，非効率な状態にあることもある．業務の全体最適化と ICT 投資の集約を実現する手段として，大阪大学は情報推進機構を設置し，仮想化技術を中心に据えたクラウド技術の活用に取り組んだ．本機構は，将来的には ICT リソースを外部にアウトソースする可能性も選択肢の 1 つとして考えながら，プライベートクラウド方式のプラットフォームシステムの構築を目指していた [3, 4].

大阪大学では上述の背景のもと，事務業務の効率化・改革の第一歩として事務系基幹システムを 2010 年に刷新した．事務系基幹システムの刷新の際には，今後の大学の様々なサービスを集約し実行可能なシステムの構築を目指し，仮想化技術を採用した共通基盤プラットフォームシステム「大阪大学キャンパスクラウド」(以下，キャンパスクラウド) を設計，構築した．このキャンパスクラウドは事務系基幹システムの計算機資源だけでなく，財務会計システムからも計算機資源の提供を受けて 2010 年からサービスを開始した．このキャンパスクラウドは hp のブレードサーバを 12 台利用し，物理コア数は合計 96，主記憶容量は合計 432GB で動作するものである．ここにさらに 2010 年度に教員基礎データベースシステムと呼ばれる計算機資源が追加され，2 つの業者が導入した 3 つのシステムによる 3 ブレードシャーシからなるキャンパスクラウドが出来上がった．この基盤上で 2014 年 3 月末時点で 32 システムによる 105VMs が稼働し，227 仮想 CPU を利用している．また本基盤の上で稼働するメールシステムは学内 38 部局，8,450 アカウントを提供している．

一方，これらの機器が設置されている大阪大学吹田キャンパスのサイバーメディアセンター本館では耐震・改修工事が 2014 年度に行われる事が決定し，スーパーコンピューター等を収納する地上 2 階建て，建物延べ面積 2,040m^2 の

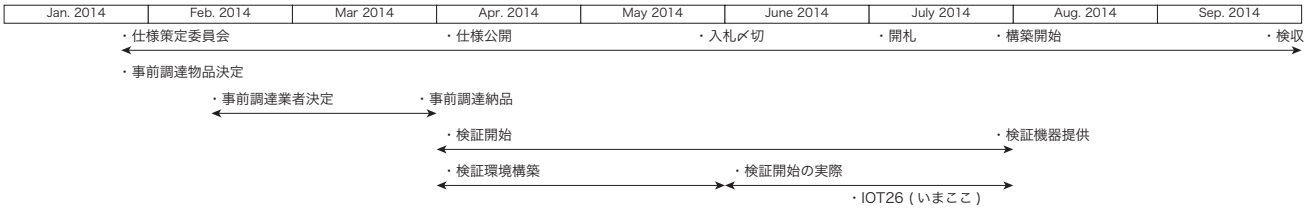

図 2　大阪大学の仮想化基盤増強の調達スケジュール
Fig. 2　A procurement schedule of an enforcement for the virtualization infrastracture in Osaka University.

IT コア棟が新設されることとなった．IT コア棟の竣工は 2014 年 9 月末を予定しており，その直後にサイバーメディアセンター本館の耐震・改修工事が着工することとなっている．サイバーメディアセンターの改修にあたって，全ての計算機資源は IT コア棟に移設される必要があるが，移設作業中，キャンパスクラウドのサービスが全て停止することが懸念された．移転作業は週末の休日と祝日を含めた 3 日間で完了することが可能と見積もられたが，メールシステムを 3 日間ダウンさせることにはユーザからの反発が想像することが難くない．それ以外にもソフトウェアのサポート上の問題点も指摘された．

2010 年のサービス構築時において仕様として要求されたハイパーバイザソフトウェアは VMware 社の VMware Infrastructure 3 である．このソフトウェアはバージョンアップを繰り返し，現在では VMware ESXi^{TM}5.5 となっているが，キャンパスクラウドは VMware ESX 4.0 で稼働し続けている．VMware 社ではハイパーバイザソフトウェアのメンテナンスアップデート，アップグレード，不具合とセキュリティの修正，および技術的な支援が提供される “General Support” の期限を製品ごとに設定しており，VMware ESX 4.0 の General Support は 2014 年 5 月 21 日で終了する．その後は 2016 年 5 月 21 日まで技術的な支援は提供されるが，その他のサポートは提供されなくなるため，ハイパーバイザソフトウェアに深刻な脆弱性が発見された場合は対処を行うことが極めて困難となる [2].

これらの問題を解決するために，基幹系プラットフォームの機器を拡張することとなった．この拡張のスケジュールについて解説を行う．

3.2　拡張のスケジュール

IT コア棟への移設に伴うキャンパスクラウドのサービス断時間を可能な限り短縮することを目的とした基幹系プラットフォームの機器拡張は，2014 年 1 月下旬に第 1 回の仕様策定委員会が開催され，同年 4 月上旬に仕様書が公開された (図 2)．同年 5 月下旬に入札が締め切られ，そして 7 月上旬に開札が行われる予定となっている．これに先んじて並行して 2014 年 1 月末に機器拡張の一部調達が始まり，同年 3 月末に機器が導入された．この事前に導入された一部機器を用いて，3 年後に予定されている本環境の増強の根拠評価を行うための検証環境を構築することが第 1 回の仕様策定委員会で決定された．

事前調達物品は下記の物品により構成されている．

- 仮想化サーバ Cisco UCS C240 (Xeon E5-2697v2 (2.7GHz) 2 基, 128GB メモリ, 2.5’ 300GB SAS HDD (10krpm) 3 基, 2.5’ 1TB SAS HDD (7.2krpm) 5 基, 2.5’ 400GB SAS SSD 1 基, VIC (10Gb SFP+ 2 ポート) 1 基) 5 台
- 統合管理スイッチ Cisco UCS 6248UP 1 台
- 10GbE 対応スイッチ Cisco Nexus 5548UP 1 台

また仮想化サーバ 5 台のうち 3 台には Fusion-io 社の ioDrive2 の OEM 製品である UCSC-F-FIO-785M を搭載しており，SAS 接続の SSD と PCI Express 接続の半導体素子メモリドライブを比較できる構成としている．仮想化サーバ 5 台のうち 3 台のみに搭載されているのは純粋に予算の問題と，年度内に検収を終えるために勘弁な調達フローを選択する必要があったことに起因する．全ての機器は 2014 年 3 月第 5 週末日に納品され検収を終えた (図 3).

図 3　導入された VMware VSAN 検証環境
Fig. 3　A verification environment for VMware VSAN

本来のスケジュールでは事前調達物品が納品された 3 月末からすみやかに検証環境を構築し，検証作業を開始するところだが，肝心の検証環境の構築や検証「される」機器は用意されたものの「する」機器や環境が揃うのに時間を

要してしまい既に二ヶ月が浪費されている．これはひとえに検証環境構築責任者の怠惰に起因するものであり，何らかの処罰が行われるのもやむなしである．

4. 検証

4.1 検証環境の論理構成

図 4 に vSAN 検証環境のネットワーク構成を示す．仮想化ホストである UCS C240 は統合管理スイッチ，Fabric Interconnect である UCS 6248UP を使わず，10GbE 対応スイッチである Nexus 5548UP と Twinax 接続している．GbE を使い Cisco Integrated Management Controller への接続用ネットワークと，ESXi サービスへのリモート接続用ネットワーク，および VM 用セグメント用のネットワークを提供している．検証環境の機材はセキュリティエリアに設置されているため，これら GbE 接続のネットワークは学内ネットワーク ODINS を経由してセキュリティエリア外の操作端末が設置されている区画まで VLAN により接続性を確保している．

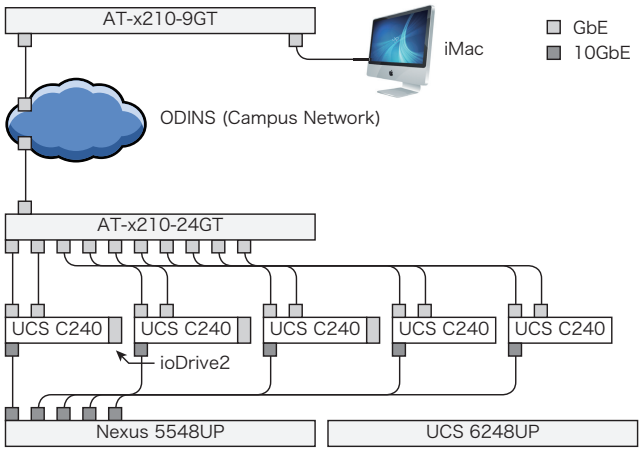

図 4 vSAN 検証環境のネットワーク構成
Fig. 4 A network diagram of verification environment for VMware VSAN

仮想化ホストである UCS C240 には ESXi 用の 4GB の USB フラッシュメモリが搭載されており，ハイパーバイザはこのメモリ上にインストールされる．VSAN を構築するためには VMware vCenter サーバが必要であるため，全ての仮想化ホストが備える 300GB のローカルストレージ上に VM を作成し，Windows Server をインストールして VMware vCenter サーバの環境を構築する．ESXi の仮想 NIC がこの VM と接続され，仮想スイッチにより UCS C240 の物理 NIC と接続されて各仮想化ホストを管理できる状態としている．VSAN クラスタを作成することで各仮想化ホストが VSAN をストレージプールとして利用することができる．このストレージプール上に VM を作成し，各種ベンチマークを実行する．

4.2 検証ツール

ストレージの評価のための検証ツールとして以下のツールを利用する．

- IOzone[*10]
- Bonnie++[*11]
- fio[*12]
- vdbench[*13]
- Oracle ORION[*14]
- Iometer[*15]

IOzone はファイルシステムベンチマークツールで，様々なアクセスパターンでの評価を行うことができる．Bonnie++もファイルシステムとディスクを対象としたベンチマークだが，ファイルサイズと作成するファイル数を指定することができる．fio はベンチマークソフトであり，負荷検証ソフトでもあり，ファイルだけでなくブロックデバイスも扱える点が特徴である．vdbench と Oracle ORION は Oracle 社によって開発されたディスク I/O ワークロードの計測ソフトウェアであり，ORION は特に Oracle のデータベースを利用することを前提としたディスク・ファイルシステム向けのベンチマークである．Iometer はもともと Intel 社が開発していた I/O サブシステムの計測ツールであるが，VMware 社が VSAN の評価のために公式に利用したツールでもある．

4.3 End User License Agreement (EULA)

VMware VSAN の評価を行う上で注意すべき点として，VMware End User License Agreement (EULA)[*16] でのベンチマーク条項が挙げられる．VMware EULA の 2.4 節では VMware Workstation や Fusion の場合はベンチマーク結果を VMware 社のベンチマーク担当 (benchmark@vmware.com) にメールするだけで良いが，それ以外の VMware 社製品のベンチマーク結果についてはベンチマーク手法，前提条件，ベンチマークのパラメータについてベンチマーク担当がチェックし，承認された場合にその結果を公開して良い，とある．この条項に同意できない場合は購入から 30 日以内にベンダーに返却しライセンス費用の払い戻しを受けるよう指示されている．条項を無視して使用したりベンチマーク結果を公開した場合にはカリフォルニア州法に従った訴訟を覚悟するべきである (VMware EULA 12.7 Governing Law. より)．また VMware 社製品については評価用ライセンスが用意されて

[*10] http://www.iozone.org
[*11] http://www.coker.com.au/bonnie++
[*12] http://freecode.com/projects/fio
[*13] http://sourceforge.net/projects/vdbench/
[*14] http://www.oracle.com/technetwork/jp/topics/index-096484-ja.html
[*15] http://www.iometer.org/
[*16] http://www.vmware.com/download/eula/universal_eula.html

おり，非生産的環境での指定された期間内での利用が許されている．評価のために製品版を購入する必要がない点にも留意されたい．

5. 課題とまとめ

大阪大学ではより早いサイクルで更改を繰り返す仮想化基盤の設計を行っている．これは 1980 年に American Society for Public Administration が出版する Public Administration Review 誌にピーター・ドラッカーが掲載した記事”The Deadly Sins in Public Administration” [5] で述べた「行政における大罪」を参考にしている．この「大罪」はキリスト教における七つの大罪 [6] に肖ったエスプリであるが，これをもとに大阪大学の仮想化基盤の増強計画について，以下の 6 つを基本設計指針としている [7].

(1) 卑俗な目標
(2) 仮想化基盤の実現を第一優先とする
(3) 規模は最低限のものを
(4) 分相応のシステムを
(5) 経験がなければこれから記録する
(6) 2.5 年ごとの見直し

大層な基本設計指針を掲げたものの，3.2 節で述べたように，まさに七つの大罪の一つに掲げられている「怠惰 (pigritia seu acedia)」により検証の進捗ははかばかしくない．この基本設計指針を掲げたこと自体が「高慢 (superbia)」だったという評価を下される誹りはまぬがれない．最低限の規模といいながらも Fusion-io の ioDrive2 の評価を行うのは設計者の「物欲 (avaritia)」を満たすためであったと解釈することも可能である．これは巨大予算によって潤沢な大規模計算機を構築することに成功した他研究機関に対する「嫉妬 (invidia)」によりこの設計が駆動されていることに起因する可能性は否定できない．あるいは設計者のままならぬ現実に対する「憤怒 (ira)」が設計思想の根底に存在することも考慮にいれるべきである．計算機資源に対するフェティシズムが反社会的な一線を超越すると「貪食 (gula)」と「肉欲 (luxuria)」として現れることが想定される．これは設計において最も高い優先順位で忌避すべきことと言える．

参考文献

[1] Jeffrey Shafer: I/O virtualization bottlenecks in cloud computing today, Proceedings of the 2nd conference on I/O virtualization (WIOV’10), pp.5–5 (2010).
[2] 柏崎礼生, 宮永勢次, 森原一郎: 大阪大学の仮想化基盤の増強と設計, 情報処理学会研究報告, Vol.2014-IOT-25, No.26, pp.1-6 (2014)
[3] 市川昊平, 江原康生, 長岡亨, 森原一郎: 大阪大学のキャンパスクラウドへの取り組み, 大学 ICT 推進協議会 2011 年度年次大会論文集, pp. 312-325 (2011).
[4] 宮永勢次, 市川昊平, 小林兼: 大阪大学のキャンパスクラウドシステムについて, 全国共同利用情報基盤センター研究開発論文集, No. 34, pp. 77-82 (2012).
[5] Peter F. Drucker: The Deadly Sins in Public Administration, Public Administration Review, Vol. 40, No. 2, pp. 103-106 (1980)
[6] Evagrius Ponticus: The Praktikos, Chapter 7-14 (345?-399?)
[7] 柏崎礼生, 森原一郎: 大阪大学における仮想化基盤の設計とその増強計画, インターネットと運用技術シンポジウム 2013 論文集, Vol. 2013, pp. 47-50 (2013)