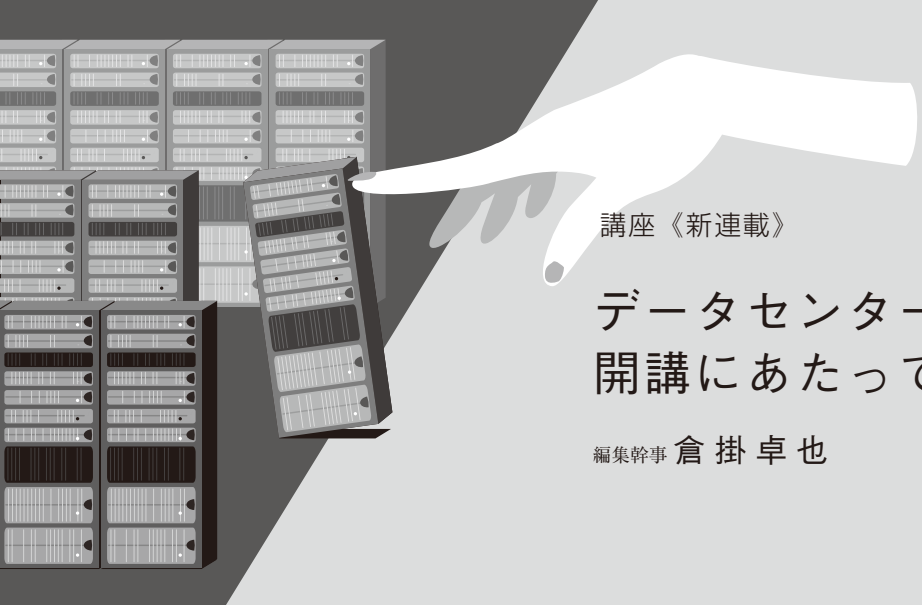


講座《新連載》

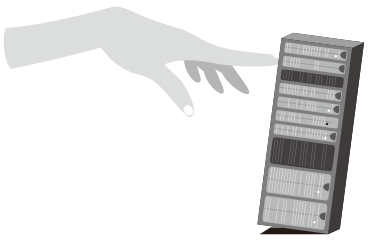
データセンターのインフラ技術〔全6回〕開講にあたって

編集幹事 倉掛卓也

Big TechやGAFAMなどと呼ばれているプラットフォーム事業者が生み出す技術トレンドは幅広い業界に影響を及ぼしています．放送システムについてもその影響を受け，将来は，「(放送業界専用ではない)汎用ハードウェアで動作するシステム」，「ソフトウェアにより実装されたシステム」，「仮想環境で動作するシステム」，「オープンアーキテクチャを採用したシステム」という方向に進んでいく可能性があります．そこで，放送技術者が汎用ハードウェアでシステムを組む際の背景知識となり得る「データセンターのインフラ技術」の今後の展望について，わかりやすく講座にまとめられれば有用ではないかと考え，本講座を企画しました．

本講座では，全体像としてデータセンターのトレンドを振り返るところからスタートし，省エネ技術などを含めたデータセンター設備の概要，サーバやネットワーク機器などのコモディティ化の動向，さらには，エッジコンピューティング，アクセラレーテッドコンピューティング，データセンタースケールコンピューティングについて，それぞれの分野の専門家に解説をお願いさせて頂きました．ご多忙な中，ご執筆を引き受けて頂いた著者の皆様には，この場を借りて深く感謝申し上げます．

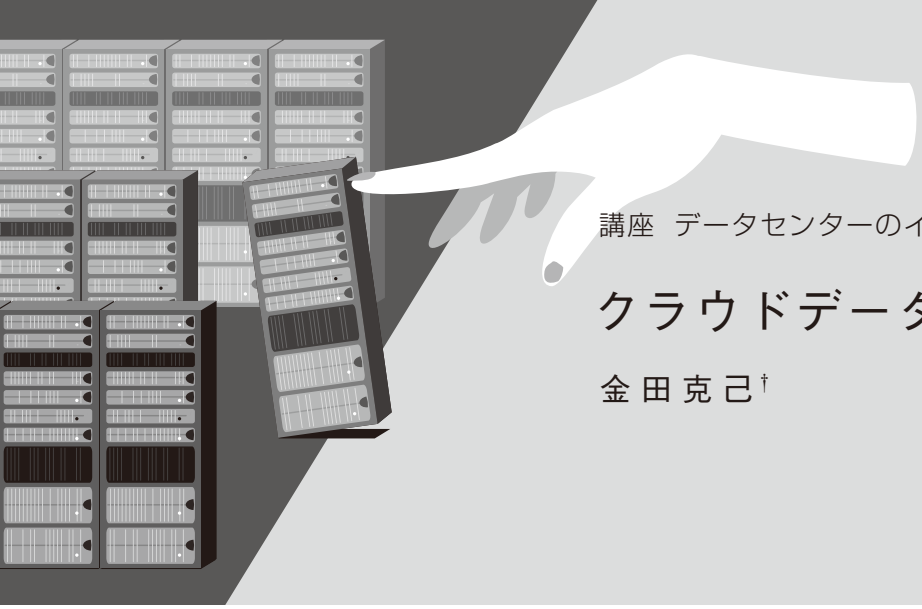
なお，本講座は，小島敏裕編集幹事と私倉掛が担当いたします．約1年間の講座となりますが，何卒よろしくお願いいたします．



予定目次

(第1回) クラウドデータセンターのトレンド	金田克己氏(インターネットイニシアティブ)
(第2回) データセンターの設備	中西草太氏(伊藤忠テクノソリューションズ)
(第3回) クラウドで使われる機材のコモディティ化	小泉利治氏(伊藤忠テクノソリューションズ)
(第4回) エッジコンピューティング	末永洋樹氏(インターネットイニシアティブ)
(第5回) アクセラレーテッドコンピューティング	佐々木邦暢氏(エヌビディア)
(第6回) データセンタースケールコンピューティング	野田　真氏(エヌビディア)

＊タイトルは今後修正の可能性があります．



講座　データセンターのインフラ技術〔第１回〕

クラウドデータセンターのトレンド

金田克己†

１．まえがき

この10年間で，アプリケーションの開発やITシステムを構築するときの実装先としてクラウドサービスを利用することが一般的になった．

2006年のアマゾンによるAmazon Elastic Computing Cloud（EC2）の提供開始から数えて今年で14年目となる．この期間にコンピューティングと通信に関わる技術は大きく進化しておりさらに発展しようとしている．

本稿では，クラウドサービスを提供するインフラストラクチャーを収めたデータセンターをクラウドデータセンターと定義し，このクラウドデータセンターの中身であるクラウドインフラストラクチャーについて解説する．

2章で，クラウドコンピューティングが成立した背景であるコンピュータの低価格化とインターネットの発展について説明し，3章ではインターネットユーザの増加に対応したシステムアーキテクチャについて解説する．続く4章ではシステムのサイロ化による問題とその解決にサーバ仮想化が提唱され，リソースを論理分割することでシステム・インフラをソフトェアで制御可能となったことを，5章ではクラウドサービスのうちIaaS（Infrastructure as a Service）と呼ばれているものを説明し，第6章ではクラウドインフラが抱えていた技術課題とそれらをどのように解決してきたかについて解説する．第7章では，クラウドコンピューティングの周辺で注目されている技術トレンドについて論じる．

２．クラウドコンピューティングに至る背景

クラウドコンピューティングが普及するに至る背景として，コンピュータの処理能力当たりのコストが年々下がり続けていること，インターネットの普及が挙げられる．

2.1　コンピュータの価格低下

1つ目の要因は，コンピュータの処理能力当たりのコストの低下である．1950年代の汎用コンピュータの販売開始から，その価格は年々下がるとともに，小型化が進んだ．

これまでのコンピュータの歴史を振り返ると，装置としてのコンピュータの主流は，台数ベースでみたときに，メインフレーム→ミニコンピュータ→RISC UNIX→パーソナルコンピュータ（PC）→スマートフォンと変遷している．

コストが下がり小型化が進むことで，国家機関だけが利用していたものを一般企業が利用するようになり，さらにはPCが個人の利用するものとして企業や家庭に入っていくようになり，今ではスマートフォンという形で個人が持ち歩くまでになった．

サーバに限ってみると，RISC UNIXからPCサーバへの置き換えが進んだ．このことで，それなりに安いサーバを沢山並べることが可能となった．

コンピュータの低価格化が進んだ結果として世界中にあるコンピュータの数は爆発的に増大した．

2.2　インターネットの普及

2つ目の要因として，インターネットの普及とそれに伴い組織内にあったコンピュータのネットワーク化が進んだことが上げられる．

図1に平成22年の情報通信白書から引用した日本国内のインターネットの利用者数と人口普及率の推移を示した．これを見ると，平成11（1999）年から平成16（2004）年にかけて普及率が急激に上昇していることが見て取れる．

この頃から，インターネットの利用人口が増加したことによりインターネットに向けて公開しているサービス（以下，インターネットサービス）へのアクセスが急激に増加したことが推定できる．

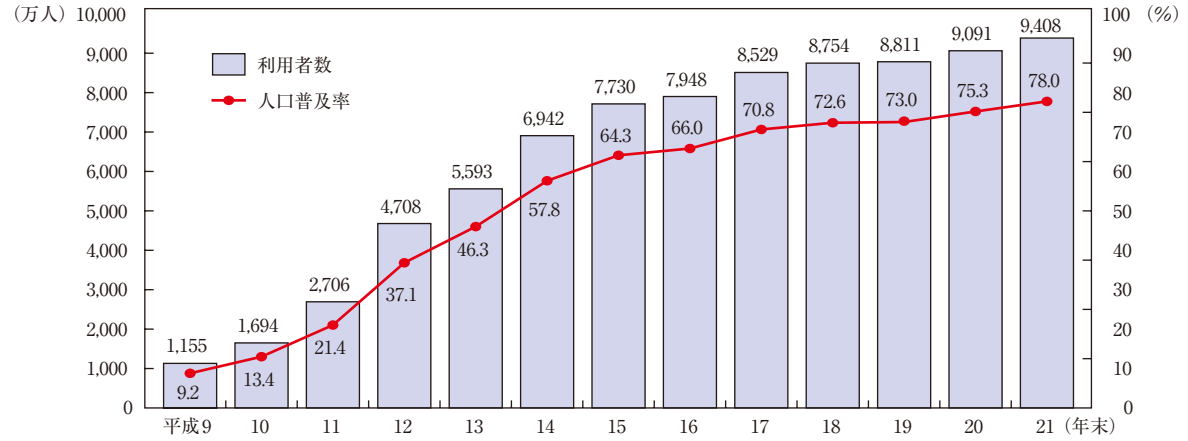
３．インターネットサービスを支えるシステムインフラストラクチャー

インターネットサービスのシステムインフラストラクチャーのアーキテクチャ（システムインフラ）も大きく進化した．

もともと，インターネットで使われている，サービス提供モデルはクライアントサーバコンピューティングモデルだった（図2）．現在インターネットで提供されているサービスも大部分はこのモデルで提供されている．

†株式会社インターネットイニシアティブ
"Infrastructure Technology for Datacenters (1); Trends of Cloud Datacenters" by Katsumi Kaneda (Internet Initiative Japan Inc., Tokyo)

映像情報メディア学会誌 Vol. 75, No. 2, pp. 220～225（2021）



①平成9～12年末までの数値は「通信白書（現情報通信白書）」から抜粋.

②インターネット利用者数（推計）は，6歳以上で，過去1年間に，インターネットを利用したことがある者を対象として行った通信利用動向調査の結果からの推計値．インターネット接続機器については，パソコン，携帯電話・PHS，携帯情報端末（PDA），ゲーム機等あらゆるものを含み（当該機器を所有しているか否かは問わない），利用目的等についても，個人的な利用，仕事上の利用，学校での利用等あらゆるものを含む.

③平成13年末以降のインターネット利用者数は，各年における6歳以上の推計人口（国勢調査結果及び生命表等を用いて推計）に通信利用動向調査で得られた6歳以上のインターネット利用率を乗じて算出.

④平成13年末以降の人口普及率（推計）は，③により推計したインターネット利用人口を国勢調査及び生命表を用いて推計した各年の6歳以上人口で除したもの.

⑤調査対象年齢については，平成11年末まで15～69歳，平成12年末は15～79歳，平成13年末以降は6歳以上.

図1　インターネットの利用者数および人口普及率の推移[1)]

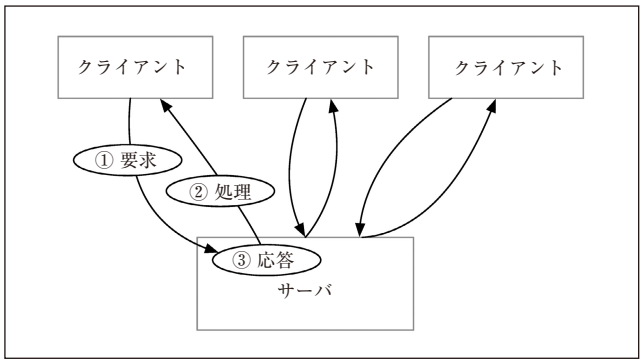


図2　クライアントサーバコンピューティングモデル

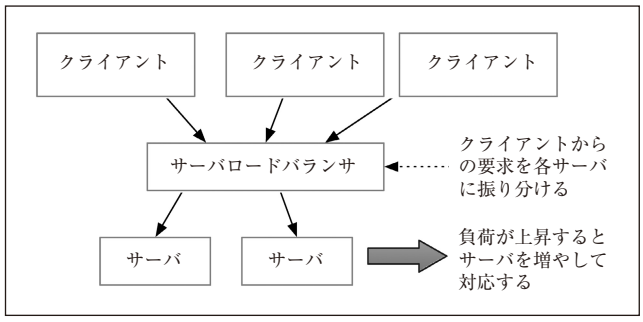


図3　スケールアウト型アーキテクチャ

表1　システムインフラに求められる要件

	組織内向け	インターネット向け
稼働率	組織の業務時間内で稼働していれば良い	世界中からのアクセスを想定すると24時間365日の稼働が求められる
利用者数	組織に所属する人に限られるため，大きく変動しない（特定多数）	インターネットのユーザ数がそのままユーザとなり得る（不特定多数）

3.1　システムインフラに求められる要件

では，ここでインターネットサービスを提供するシステムに求められる要件を，従来の組織内にサービスを提供するシステムと比較してみよう.

表1に示したように，インターネットの出現によってシステムインフラに求められる要件は大きく変化した.

3.2　冗長性とスケールアウトアーキテクチャ

2000年以前のインターネットが一般に利用できるようになった当初はユーザも少なく，インターネットサービスを提供するサーバも1台だけという状況が多かったように記憶している.

ところが，インターネットユーザが増えてくるとサーバ1台だけでは処理が追いつかなくなってきた．そこで，採用されたのがスケールアウト型のシステムアーキテクチャである（図3）.

スケールアウト型アーキテクチャは，負荷の増大に合わせてサーバを増加させればシステムとしての処理能力を増やすことができること，また，1台のサーバが停止したとしても，サーバロードバランサが停止したサーバを切り離すことでサービスを継続させることができるという利点がある.

クライアントからの要求を各サーバに振り分ける処理をするサーバロードバランサは2000年前後から採用され始めた．また，この頃から19インチラックに搭載可能なラックマウント型のサーバが採用され始めた.

さて，スケールアウト型のアーキテクチャの採用によりインターネットユーザからもたらされる膨大なリクエストを捌く準備が整った．一方で，24時間365日の稼働については予備の機器をホットスタンバイ状態でシステムに組み込んでおくHA（High Availability）構成を取り入れることで対応していた．

同時期にインターネット向けシステムに最適なアーキテクチャとしてwebサーバ，アプリケーションサーバ，リレーショナルデータベースサーバからなる3層アーキテクチャが提唱された．

図4に3層アーキテクチャの典型的なシステムインフラ構成を示す．

3層アーキテクチャはシステムの構成要素をプレゼンテーション層，ビジネスロジック層，データ層に分割して，それぞれの層にWebサーバ，アプリケーションサーバ，データベースサーバを配置したアーキテクチャである．

このアーキテクチャの利点は，各層は独立しているのでそれぞれの都合で変更を加えることが可能であること，Webサーバとアプリケーションサーバはスケールアウト可能であることである．当時は，データベースサーバはサーバのCPUを強化する，メモリーを追加する等のスケールアップが必要であった．現在はデータベースサーバもスケールアウト可能なものがいくつか存在する．

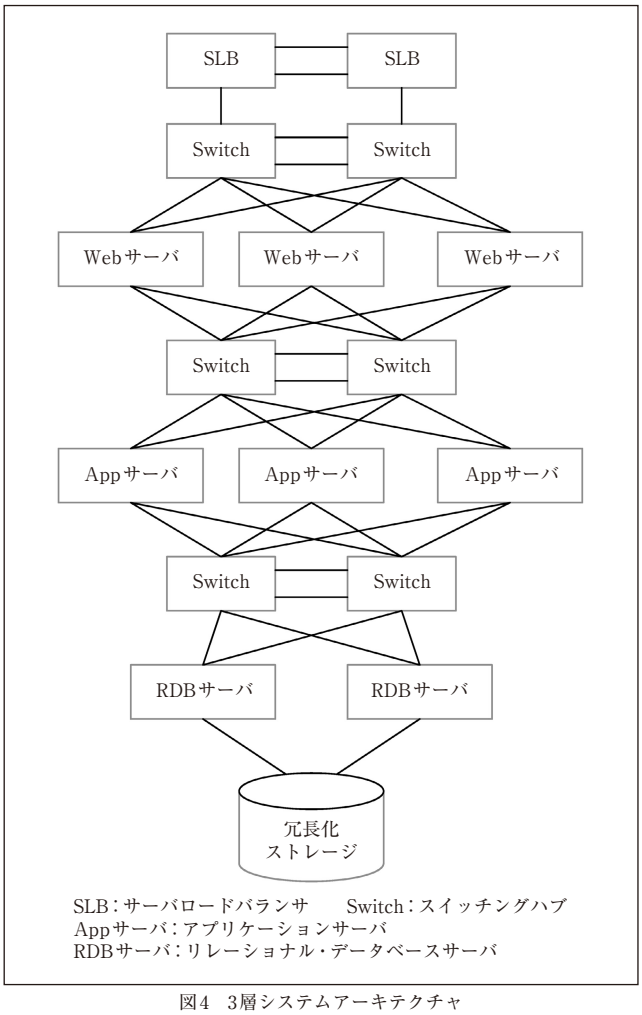


図4　3層システムアーキテクチャ
（インターネット接続用のルータ，ファイアウォールは省略）

4．サイロ化したシステムとシステムインフラ統合

さて，インターネットユーザの増加に伴って，インターネットサービスのシステムインフラは強化されていくことになるが，2005年ぐらいに今度はサイロ化したシステム（図5）が問題視されるようになった．

サイロ化したシステムの問題とは，ある組織内で目的毎にシステムを作った結果複数のシステムが乱立した状態となり，このとき各システムがその処理能力を想定される最大負荷に合わせて構築されるために，システムインフラ全体の能力が余剰となった状態のことである．一説には，複数システムを合わせたときにそのシステムインフラの30%程度しか使われていないということも言われていた．

システムのサイロ化が問題として取り上げられるのと同時期（2005年頃）にPCサーバを仮想化して利用することが実用可能になってきた．

サーバ仮想化とは，ハードウェアの機能，あるいは，ハイパーバイザと呼ばれるソフトウェアの機能を利用して，一台の物理コンピュータ内で複数の仮想マシン（VM）を動作させることである．このサーバ仮想化は，物理コンピュータの持つリソース（CPU，メモリー，HDD）を論理分割し，それらリソースをVMに割り当てることで実現される．

サーバ仮想化技術は昔からあったが，2000年ぐらいから徐々に普及してきたPCサーバで利用できるというのが画期的だった．例えば，VMware社VMware ESX，オープンソースソフトウェアのXen，KVM等がある．

図6にこのハイパーバイザを用いたサーバ仮想化の概念を示す．物理サーバをハイパーバイザで分割することで，複数のサーバを出現させることが可能となる．また，ネットワーク構成の都合からスイッチングハブを仮想化した仮想スイッチが用意される．より詳細な説明は紙幅の都合もあり割愛する．

PCサーバの高性能化と低価格化が後押しとなってサーバ仮想化は急速に広まった．

それでは，ここで図4に示した複雑なシステムをサーバ

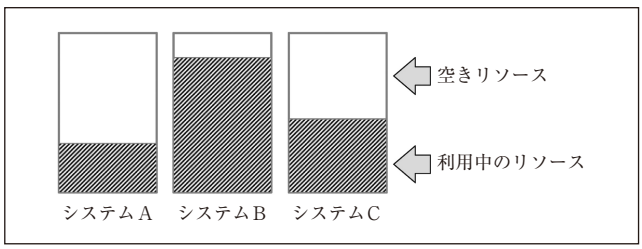
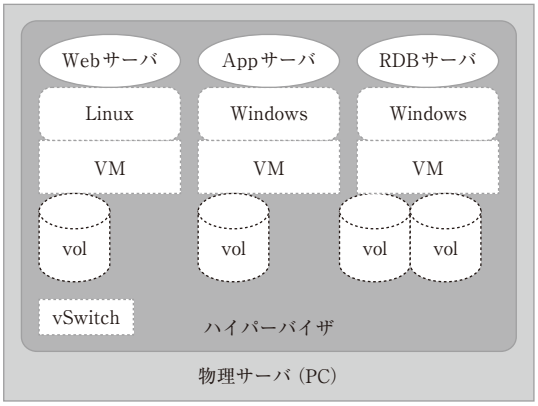


図5　サイロ化したシステム



VM：仮想マシン（Virtual Machine）
vol：仮想ストレージ（Volume）
vSwitch：仮想スイッチ（Virtual Switch）

図6　ハイパーバイザを用いたサーバ仮想化

仮想化技術を用いて集約してみよう．

図4のSLBから下のサーバとスイッチングハブ（図中ではSwitch）のみを対象とする．手順としては以下の通りとなる．

(1) 最初にサーバをVMに置き換える．
(2) 次に，スイッチングハブを仮想スイッチ（vSwitch）に置き換える．
(3) 最後に，冗長性を確保するために2台の物理サーバ上にこれらVMとvSwitchを配置する．

後は，物理構成時のネットワークトポロジーを保持するように，VMとvSwitchを論理的に接続することで完成となる．

その結果を図7に示す．物理スイッチの台数が6台から2台に，物理サーバの数が8台から2台に削減された．実際にはここまで極端に集約をすることはあまりないが，サーバ仮想化がシステムインフラの統合に有効であることをご理解頂けたのではないかと思う．

サーバ仮想化技術の出現によって，サイロ化したシステムのインフラを集約することが技術的には可能となった．ここで重要なのは，適切に設計されたシステムインフラがあり，その構成要素がソフトウェアで制御可能であれば，物理的に手を加えることなくサーバ，ストレージ，スイッチングハブに適切な設定を行うことで，論理的な構造を変更できるようになったことである．

これで，クラウドを構成するための基本的な技術は揃った．次章では初期のクラウドについて解説する．

5．IaaS（Infrastructure as a Service）

ここでは，クラウドコンピューティングのうちIaaSについて解説する．

最初に，クラウドが備えるべき要件を説明するために，アメリカ国立標準技術研究所（NIST）が2011年に発表した「NISTによるクラウドコンピューティングの定義」（NIST SP800-145）[2]，IPAによる日本語訳[3]から，クラウドが備える基本的な特徴の一部を引用する．

(1) オンデマンドセルフサービス（On-demand Self-service）
・ユーザは，各サービスの提供者と直接やりとりすることなく，必要に応じ，自動的に，サーバの稼働時間やネットワークストレージのようなコンピューティング能力を一方的に設定できる．

(2) スピーディな拡張性（Rapid elasticity）
・コンピューティング能力は，伸縮自在に，場合によっては自動で割当ておよび提供が可能で，需要に応じて即座にスケールアウト/スケールインできる．ユーザにとっては，多くの場合，割当てのために利用可能な能力は無尽蔵で，いつでもどんな量でも調達可能のように見える．

では，図8を用いてIaaSの仕組みの概略を示した上で，上記の基本的な特徴をどのように実現するかを説明しよう．

IaaSの構成はおよそこのようになる．

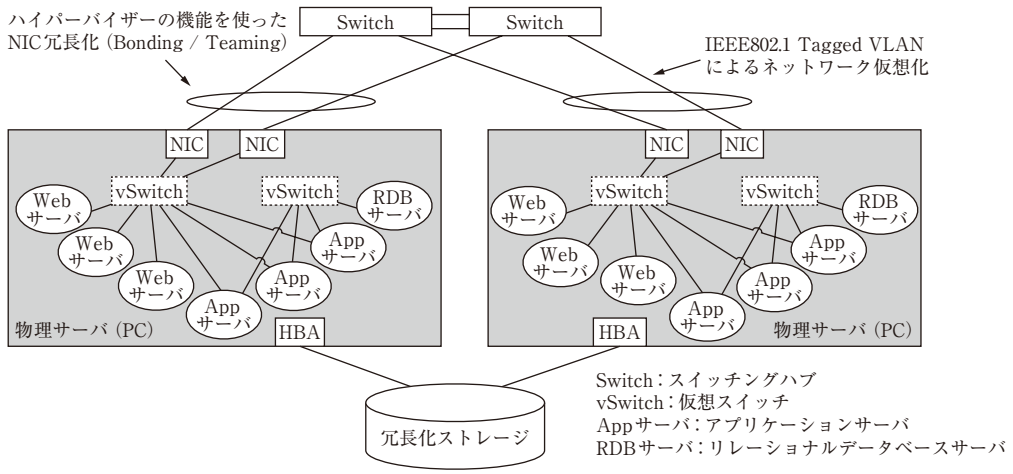


図7　サーバ仮想化技術を用いたインフラの集約

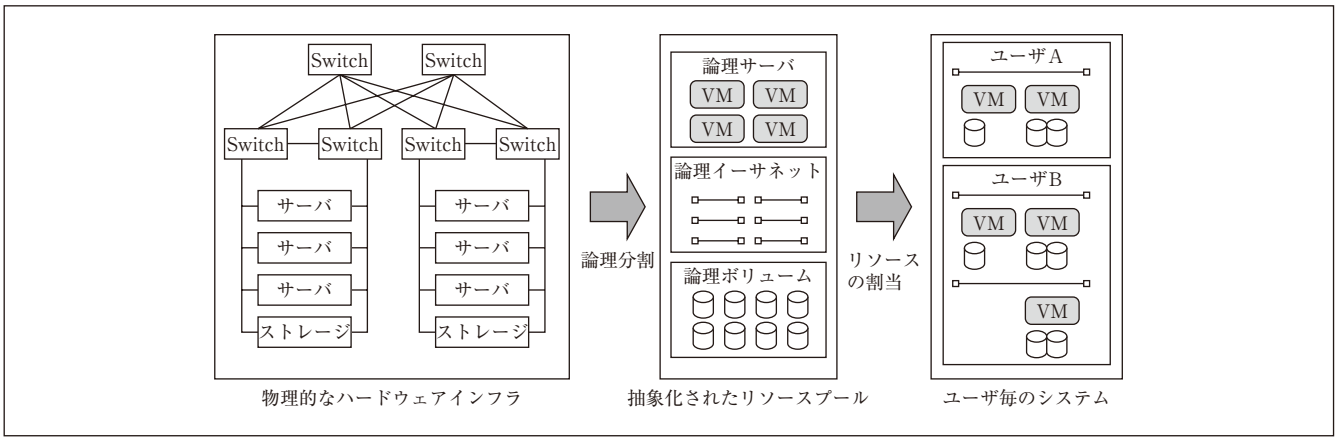


図8　IaaSの仕組み

(1) まず，サーバ，ストレージ，スイッチングハブを適切に接続し，仮想化技術を用いて論理的に分割する．
(2) 次いで，VM，論理的なイーサネット，仮想ボリューム（ストレージ）からなるリソースプールを作成する．
(3) 最後にユーザからの要求に応じてリソースプールから必要なリソースを取り出し，ユーザの要求したシステムを構成する．

ここで前述したクラウドが備えるべき特徴から必要となる要素を考えてみる．オンデマンド・セルフサービスを実現させるためには，契約の申込み，契約したリソースの配置などをユーザ自身で実現できるようなGUI，または，API（アプリケーションインタフェース）が必要である．また，物理的なハードウェアインフラの論理分割，契約に基づいたリソースの割当などを人手で実施していてはスピーディな割当を実現するのは非常に困難であるのでソフトウェアにより実行する必要がある．特に，クラウド環境では数千台から数万台のサーバを制御する必要があるのでソフトウェアによる制御が必須となる．

これらクラウドを制御するためのソフトウェアをクラウドオーケストレータと呼ぶことにする．

6．クラウドコンピューティングにおける技術の進展

ここからは，Big TechやGAFAM等と呼ばれている事業者が保有している大規模クラウドについて，2010年ぐらいからの技術の進展について解説する．

最初に注目を浴びたのは，電力コストが安い地域に大規模データセンターを建設する動きである．BigTechは機器のランニングコスト削減のために自社で大規模なデータセンター(クラウドデータセンター)を建設するようになった．

連載2回目に，このようなクラウドデータセンターの解説がある．

また，調達する機器のコストを削減するために従来のサーバメーカではなく，ODM（Original Design Manufacturer，相手先ブランドによる設計製造）ベンダに対して大規模な発注をするようになった．BigTechによるODM機器の採用については連載3回目で解説がある．

なぜ，BigTechがこのように行動したかを推測すると，システムインフラのランニングコストは，機器費用，電気代（機器に供給するものと空調に要するもの），運用人件費それぞれが1/3ずつを占めるという話があり，運用人件費についてはソフトウェア化を進めてコンピュータに置き換えることで自社の技術で手当はしたので，残る2つについて手を打ったと考えられる．

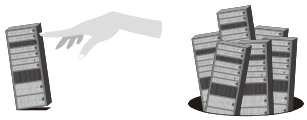
続いて，これも2010年ぐらいの話となるがBigTechは自社のサービスに合わせてクラウドを構成する機器に手を加え始める．

これは，当時市販されていた機器では，クラウドオーケストレータを使ったソフトウェア制御が困難であったこと，物理機器を論理分割したときの粒度が充分でなかったことに起因すると推測される．

サーバについては，ソフトウェアで制御することは容易であり論理分割数も搭載したCPUとメモリーに依存するので，ODMから調達したコモディティサーバを使いオープンソースのハイパーバイザーに手を加えて利用していたようである．Amazon AWSがXenに独自の改良を加えてサービスを提供していたということが言われていた．

ストレージについては公開されている情報がほとんどないため推測となるが，当時の市販ストレージ装置はユーザが外部からソフトウェアで制御することはほとんど考慮されておらず，クラウドの特性上少数の大規模なストレージ装置を導入したとしても，大量のVMに提供するには論理分割数が不充分であり，アクセスの集中による負荷に対応できなかったという事実を考え合わせると，分散型の自社開発ストレージを採用していたであろうと推測される．

ネットワーク機器についても同じ様な状況で，機器の制御は管理者がCLI（Command Line Interface）を用いて実施する前提で作られていたため外部からソフトウェアで制御するのは非常に困難であり，論理分割数についても当時一般的に使われていたIEEE 802.1Q VLANは4096個までと



いう制約があり大規模クラウドの要求を到底満たせるものではなかった．

リソース同士を接続するネットワークは特に重要で，リソースの在庫があり提供可能であったとしてもネットワークの在庫が払底してしまうとリソースを提供することができなくなる

このことから，BigTechは以下のような選択をしたと考えられる．

(1) ネットワーク機器で構成される物理ネットワークは単なる土管として考え，論理分割はサーバ側のネットワーク仮想化で行うこと．

(2) 安価なイーサネットスイッチを調達して，それを制御するOS（ネットワークOS）を自社で開発すること．

(1)については，VXLAN（Virtual eXtensible Local Area Network）やNVGRE（Network Virtualization using Generic Routing Encapsulation）といったネットワーク仮想化プロトコルを用いて実装したであろうと考える．(2)については，マーチャントシリコンと言われる大量生産しているスイッチASIC（Application Specific Integrated Circuit）を用いたイーサネットスイッチ（ホワイトボックススイッチとも呼ばれる）に自社開発のネットワークOSを搭載して利用していたと推測する．例えば，マイクロソフト社の提供するクラウドサービスであるAzureで利用しているSONiC（Software for Open Networking in the Cloud）というネットワークOSについてはオープンソース化されており，最近注目されるようになった．

さらに進んで，ネットワークの高速化により顕著になった，サーバ仮想化に用いるハイパーバイザーのオーバヘッドを嫌って，I/O処理をオフロードするのにASICやFPGAを用いるようになる[4]．このようにCPUで処理していた処理をハードウェアで実行することをハードウェアオフロードと呼ぶ．こちらは連載5回で解説される．

最終的には，BigTechは自社のサービスで利用するためにコンピュータを自社開発する．例えば，2017年に発表されたGoogleのTPU（Tensor Processing Unit）や2018年に発表されたAmazon AWSのGraviton[5]が有名である．

7．進化し続けるクラウド，集約と分散

クラウドデータセンターにコンピュータを集約する形で成長してきたクラウドサービスだが，集約とはまた別の流れがある．一つはハイブリッドクラウドと呼ばれる，組織がもつ自分たちのクラウドとクラウドサービスを併用する利用形態，もう一つは，地理的にデータ処理を必要とする場所の近くにクラウドを置くエッジクラウドという概念が提唱されている．エッジクラウド，あるいは，エッジコンピューティングについては連載第4回で解説される．

さらに，近年ML（Machine Learning）/DL（Deep Learning）の重要性が高まり，それを効率的に行うためにGPU（Graphics Processing Unit）の利用が盛んになされている．最近では1サーバでは処理しきれないために複数のサーバを連携させて処理を行わせるデータセンタースケールコンピューティングが提唱されている．

このデータセンタースケールコンピューティングについては連載6回で説明がある．

8．むすび

2000年前後に起こったインターネットの発展とそれを受けて2010年前後に始まったクラウドコンピューティングの進展について解説してきたが，その進化速度には改めて驚かされる．

1950年代に商用コンピュータが出現した当時はとても高価なものだった．それがコンピュータの低価格化，小型化が進んだことで個人が所有できるまでになった．

コンピュータは世界のありとあらゆるところで使われるようになり，インターネットの出現によってそれらコンピュータはネットワークで繋がった状態となった．ここまでは，コンピュータが分散して広がっていく過程と捉えることができる．

このような背景の中でクラウドサービスが世の中に出現し急激に成長した．今では，クラウドサービスは世界中のコンピュータで実行されていたアプリケーションを強烈な勢いでクラウドデータセンターに吸い上げている．このことは，世界中のコンピュータがクラウドに集約されて行く過程と言えよう．

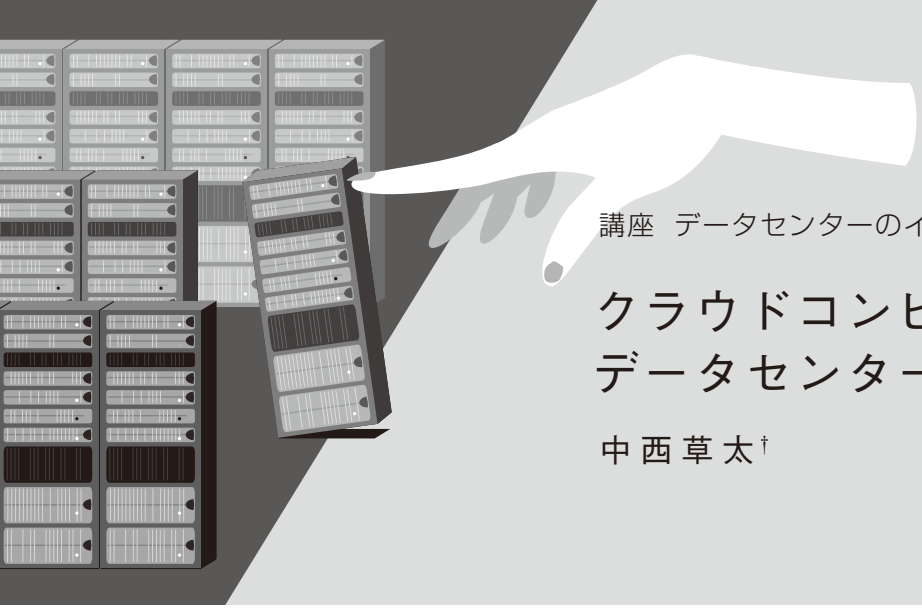
この先，クラウドコンピューティングはさらに発展していくだろうが，今のクラウドデータセンターへの大規模集約がさらに進むのか，コンピュータの歴史をなぞるように小型化して分散していくのか興味は尽きない．

（2020年12月8日受付）

〔文　献〕

1) 総務省：“平成22年版情報通信白書”，https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h22/html/me411100.html
2) Peter Mell, Tim Grance: "The NIST Definition of Cloud Computing" (2011)
3) IPA: "NISTによるクラウドコンピューティングの定義" (2011)
4) Daniel Firestone et al: "Azure Accelerated Networking: SmartNICs in the Public Cloud" (2018), https://www.microsoft.com/en-us/research/uploads/prod/2018/03/Azure_SmartNIC_NSDI_2018.pdf
5) Amazon：“AWS Gravitonを搭載したArmベースのインスタンスのAmazon EKSサポートの一般提供を開始”（2020），https://aws.amazon.com/jp/about-aws/whats-new/2020/08/amazon-eks-support-for-arm-based-instances-powered-by-aws-graviton-now-generally-available/

金田　克己（かねだ　かつみ）　1991年，長岡技術科学大学工学部電気・電子システム学科卒業．1993年，同大学院修士課程修了．2010，（株）インターネットイニシアティブ入社．以来，クラウドインフラストラクチャーの研究・開発に従事．現在，同社先行開発室長．



講座　データセンターのインフラ技術〔第２回〕

クラウドコンピューティングを支えるデータセンターのファシリティ

中西草太†

１．まえがき

いまや放送業界のみならず，ほぼすべての産業分野において，ICT/ディジタル技術の進展は必須であり，クラウドコンピューティングとその需要は増え続ける一方である．

その潮流の中で，「DX：ディジタルトランスフォーメーション*1」というキーワードが現在の産業界を席巻している．ディジタルトランスフォーメーションという新たな概念の普及は，今日における技術や組織の概念に対する変革を要求し，新たな産業・事業の創生に向けた活動を活発化させている．そして，その活発化された活動がクラウドコンピューティングの更なる拡大を加速度的に押し上げている．

ジークムント・バウマンはその著『リキッド・モダニティ』の中で，「新しく出現した時間の瞬間性は人間の共生形態を激しく変化させた．とりわけ，人間の集団事象とのかかわり方，そして，ある事象を集団的なことがらにする方法の変化はいちじるしかった」[2]と述べている．ディジタルトランスフォーメーションのような既存の概念や知識，技術の範疇を超えるものは，概ね大きな不安の念を持って迎えられるものであるが，同時にディジタルトランスフォーメーションの到来は新たな挑戦でもある．ディジタルトランスフォーメーションは，産業における新たな知を創造する一翼を担うであろう．

これまで企業では，自社内に置いたサーバを自社のネットワーク回線を通して利用し，個々の業務アプリケーションを運用してきた．しかし，インターネットの普及とIT技術の進化，PCサーバの高性能化/低価格化とあいまって，この10年では，大規模なデータセンター上にあるクラウドサービスを利用してシステムを構築/運用することが一般的となってきた．政府も『クラウド・バイ・デフォルト原則』[3]を提唱し，基盤システムのクラウドへの切り替えを宣言するなど，クラウドサービスはすでに社会基盤になくてはならない存在となっている．

データセンターは，クラウドコンピューティングの需要の急拡大を実現するためにはかかせない，重要かつ必須な戦略的施設であり，サービス提供拠点である．データセンターにおけるサーバ機器等の高性能化/高密度化が進む一方で，効率的かつ堅牢なデータセンター施設・設備の運用とサービスの提供が今後ますます重要となってくる．

データセンターのインフラ技術第２回では，ミッション・クリティカルなシステムを支えるデータセンターのファシリティについて，その概要と技術について述べる．

２．データセンターとは

今日，ITシステム基盤は，各企業，社会にとっても非常に重要なものとなっている．個人的な活動から，企業，政府のさまざまな活動にいたるまで，情報の取り扱いの重要度は増大し，それが滞ることはすなわちそれが支える活動自体の停止を招く．一方で，情報システムの安定的な稼働環境を個々のシステムごとに用意し，維持管理することは，費用/スキル/労力/体制/場所といった諸々の観点に照らして，得策ではないことが多い．したがって，それを包含し，まとまった形で構築/準備し，運用していくことが現実的なソリューションとなるため，これらを1つの（あるいはまとめたいくつかの）建物に収容し実現する．それをデータセンター（図1）と呼ぶ[4]．

データセンターは，社会的インフラとして，ミッションクリティカルなシステムを安定して安全に稼働し続ける土台である．システムは止まらないのが当たり前であり，データセンターも止まらないことがその最上位使命である．万が一システムが停止すると，あらゆる方面で経済的損失が発生するだけでなく，情報収集やコミュニケーションの方法にも大きな打撃を与えることになる．

例として，2019年8月23日，Amazon Web Services（AWS）東京リージョンで大規模障害が発生し，広範囲にわたって影響を及ぼしたことは記憶に新しい．AWS東京

＊1　DX（ディジタルトランスフォーメーション）：企業がビジネス環境の激しい変化に対応し，データとディジタル技術を活用して，顧客や社会のニーズを基に，製品やサービス，ビジネスモデルを変革するとともに，業務そのものや，組織，プロセス，企業文化・風土を変革し，競争上の優位性を確立すること[1]．

† 伊藤忠テクノソリューションズ株式会社
"Infrastructure Technology for Datacenters (2); Data Center Facilities Supporting Cloud Computing" by Souta Nakanishi (ITOCHU Techno-Solutions Corporation, Kanagawa)

映像情報メディア学会誌 Vol. 75, No. 3, pp. 370～376（2021）

図1　データセンター内部イメージ

リージョン内のデータセンターの空調設備において障害が発生し，それによって一部のサーバの温度が許容限度を超えた結果，サーバの電源が停止し始め，同一データセンター内のAmazon Elastic Compute Cloud（EC2）サービスなどのパフォーマンスが劣化した．復旧するまでの間，大手ショッピングサイトやキャッシュレスサービスサイトなどがサービスを停止するなど，さまざまなサービスに影響を与えた．

3．安心・安全を支えるデータセンターの立地・建物

データセンターの安心・安全・安定稼働を支える主な構成要素として，以下4つの要素があげられよう．

(1) 防災・防犯：災害リスクが低い立地であること．耐災害性に優れ，高セキュリティであること．
(2) 安定した電源：無停電であること．ノイズが入らないきれいな電源が提供されること．
(3) 良好で正常な空気環境：適切な温湿度を保持できること．
(4) 安定した通信インフラ：常にネットワークがつながっていること．

周知のとおり，日本は災害大国であり，地震を代表とする各種災害の影響をいかに最小限に抑えるかはデータセンターにとって大切な要素である．また，ビジネスを支える機密性の高いデータを扱うため，防犯についても高レベルでの対応が求められる．

3.1　データセンターの立地

立地選定と建物の選定はデータセンターにとって最重要事項のひとつである．データセンターを取り巻く自然起因のリスクは，立地選定と建物でほとんど回避可能だからである．また，人工的災害の各リスクも回避しなくてはならない（表1）．例えば，硫化水素はIT機器に影響を及ぼすため，火災などのリスクを含めて，工場の近くでの立地は避けるべきである．大使館や官公庁もテロの標的になる可能性があるため，その近くも避けることが望ましい．

表1　自然災害/人工的災害リスクの一例

自然災害のリスク	人工的災害のリスク
地震，活断層/塩害/津波/液状化現象/落雷/台風/洪水，河川の氾濫/森林火災	飛行経路/トンネル，湖，埋立地/無線電波，電磁波/工業汚染/大使館，官公庁，発電所，テレビ局，駅・空港の近く

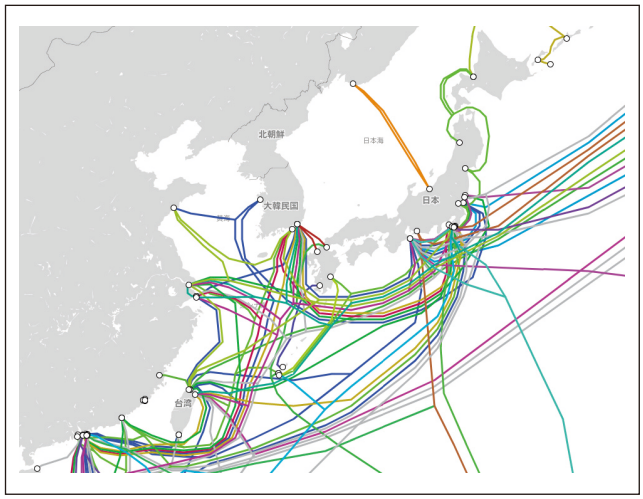


図2　日本近海における主な海底ケーブルルート[5)]

高い可用性/事業継続性が求められるデータセンターにとって，立地選定におけるデューデリジェンス*2を行うにあたり，前述のような耐災害性を検討することは必須である．クラウドコンピューティングを支える大規模なデータセンターはそれに加え，膨大な量のサーバ資源を収容し稼働させるための，広大なスペースと大容量電力の確保が必要となる．また，電力会社から66,000 Vなどの特別高圧の電力ラインをデータセンターに引き込むにあたっての経路/コスト/引き込みやすさ（納期）も重要な選定要素である．

さらに言えば，クラウドサービスの根幹を担うデータセンターにとって，レイテンシー（通信の遅延）を目標以内に収められる立地であることは，必須の要件となる．IX（インターネットエクスチェンジ）からの距離はもとより，昨今では，海底ケーブルの陸揚げ地点を考慮し，用地選定することも多い（図2）．

4．データセンターの生命線となる電源設備

データセンターのおよそすべての設備は，電気がないと稼働できない．現在のIT機器が電気のON/OFFで情報を処理する原理である以上，停電はすべての情報の「0クリア」を意味するため，通常運用のみならず，設備の点検や更新時，広域停電時や大災害といった場合においても，無停電を実現する必要がある．IT機器への給電は，一瞬でも停止が許されないという点に，データセンターの本質がある．一方で，電気事故は人命に直結するため，あらゆる操作や

＊2　Due Diligence（デューデリジェンス）：投資を行うにあたって，投資対象となる建設用地の価値やリスク等を調査・査定すること．

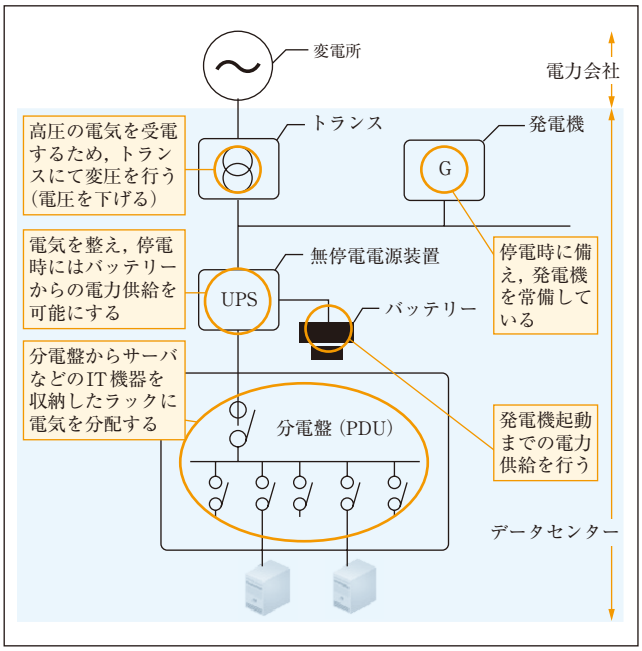


図3　データセンターの電気設備構成要素

作業は安全確保が前提となる．すなわち，「何かを行う」=「停電させる」ことを意味する．この正反対のまったく矛盾した両方の要件を満たすための冗長化や二重化などの設計/構築および運用手順の確立と教育訓練，確実な実施，さらには検証と改善の継続が必須である．それらが伴って初めて，データセンターが機能する．

4.1　データセンターの一般的な電気設備構成要素

データセンターにおける，電気の安定供給を支える一般的な設備の構成要素を図3に示す．

4.2　受電方式の種類

データセンターの電気は，基本的には電力会社から66,000 Vあるいは22,000 Vなどの特別高圧，ないしは6,600 Vの高圧にて受電している．受電する方式にはいくつかの方式があるが，ここでは，代表的な3つの方式について概略を述べる．

(1) 本線予備線受電方式（図4）：本線予備線受電方式は常時は2回線で受電し，そのうち1回線（本線）を常用している．本線故障時またはメンテナンス時に本線側遮断器を開放し，予備線側遮断器を投入することにより，短時間での停電で受電が再開できるという長所がある．

(2) ループ受電方式（図5）：ループ受電方式は常時2回線で受電するため，1回線が故障してもその回線を遮断することにより，もう片方の1回線から受電を継続できるという長所がある．ただし，各需要家間や変電所-需要家間の区間ごとに故障を監視する必要があるため，保護継電器が複雑になる．また，需要家の遮断器などはループの電流を流すため大型になる．ループ受電方式は，22～66 kVの特別電圧で採用されている．

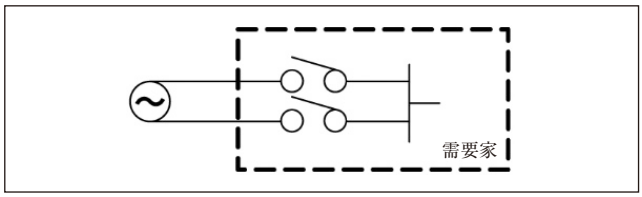


図4　本線予備線受電方式

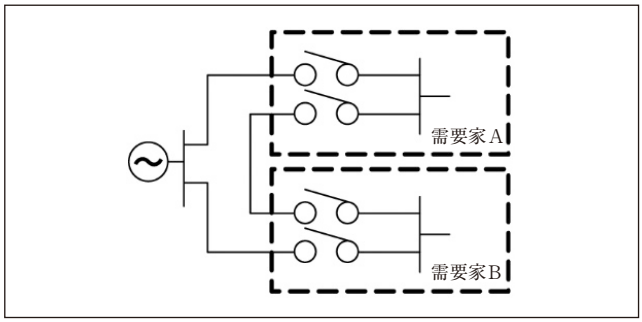


図5　ループ受電方式

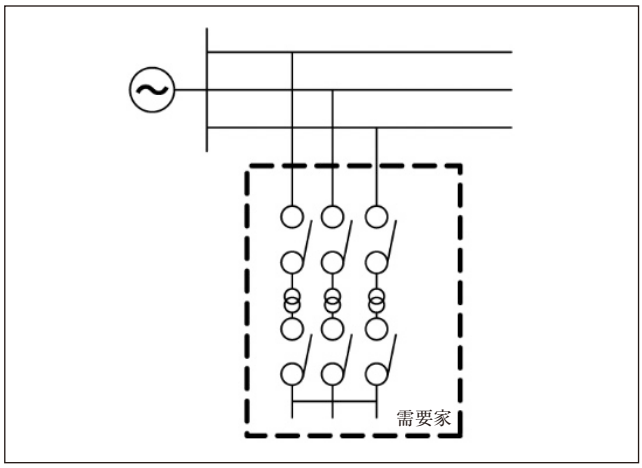


図6　スポットネットワーク受電方式

(3) スポットネットワーク受電方式（図6）：スポットネットワーク受電方式は常時3回線から受電する．保守時には，1回線ごとに停止させるため，需要家の停電は発生しない．また配電事故時には，変圧器二次側の遮断器が自動的に遮断され，復電時には自動的に投入される．主に大都市部で採用されている．

4.3　発電機の方式

データセンターでは，電力会社からの電力供給が停止した場合でも電源をIT機器に供給できるよう，発電機を常設している．

2018年9月6日の北海道胆振東部地震の影響により北海道電力からの電力供給を失ったさくらインターネット社の石狩データセンターが，約60時間もの間，自データセンターの発電機より電源を供給し続けたことをニュースで知られた方も多いのではないだろうか．

データセンターでは，ガスタービン発電機およびディーゼル発電機が採用されることが多い．それぞれ一長一短があ

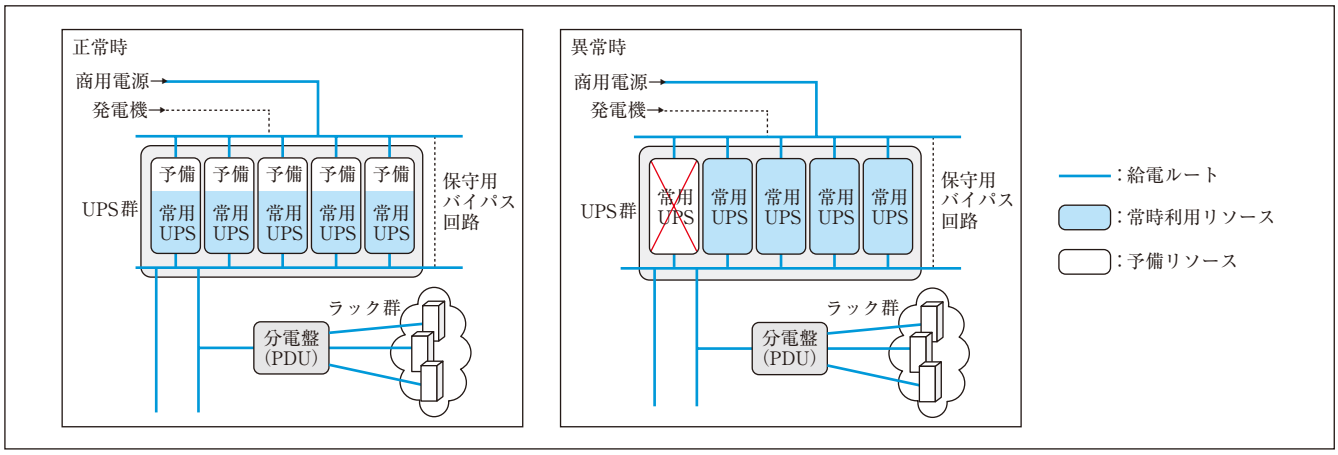


図8　並列冗長方式UPS構成

表2　ガスタービン発電機とディーゼル発電機の比較

発電機	ガスタービン	ディーゼル
外　観		
単機容量	4,000～5,000 kVA	2,000～4,000 kVA
騒　音	対策しやすい	対策しにくい
据付面積	小さい	大きい
起動信頼性	高い	普通
燃料消費量	多	小
冷やし方	空冷	冷水
物品価格	高	低
メンテナンス難易度	高（専門的）	低（汎用的）

図7　UPSとバッテリー

り，設備環境に応じた発電機を採用する必要がある（表2）．大型のデータセンターではガスタービン発電機を採用することが多いが，昨今の超大型のクラウド・データセンターにおいては，発電機の大型化によるスケールメリットよりも，備蓄用の燃料タンクが小さくできる点や，メンテナンスが容易である点などから，ディーゼル発電機を多数備える方式が採用される事例も散見される．

4.4　無停電電源装置（UPS）の方式

無停電電源装置（UPS: Uninterruptible Power Supply）とは，通常の電力供給がストップした場合においても，接続されているIT機器などに安定した電気を供給し続ける電源装置である（図7）．ただし，電力供給できる時間は数分～数十分であることが一般的であり，長時間の停電への備えとしては，発電機と組み合わせた措置を講じる必要がある．データセンターでは，通常，発電機が稼働し安定した電気を供給できるようになるまでの時間（通常3分～4分ほど），UPSにて給電する．

データセンターで設置するUPSは，サーバラック内に設置するような5～10 kW レベルの小型のUPSではなく，非常に規模が大きいものとなる．ここでは，代表的なUPSの接続方式と停電時の流れについて記載する．

(1) 並列冗長方式（図8）：電力負荷に対し，UPS1台分の余裕をもって設置されている方式（N+1構成の場合）．常用UPS1台に障害があっても，残りのUPSでカバーし，問題なく稼働できる．設備構成がわかりやすく，メンテナンスが容易である．

(2) 待機冗長方式（図9）：2つのUPS（またはUPS群）があり，通常は常用UPSのみを利用し，予備UPS（または共通予備UPS群）を待機させている方式．常用UPSに障害があった場合でも，無停電で予備UPSにて給電可能となる．なお，大規模なクラウド・データセンターにおいては，待機冗長方式をさらに発展させたブロックリダンダント方式*3を採用しているデータセンターも多い．

5．高密度なサーバ群に対応する空調設備

システムを構成する各IT機器は，CPUの指示で動いている．CPUは処理を行うことで熱が発生するが，これが極限になるとIT機器のダウンにつながるのである．IT機器の稼働を止めないためにはその熱を冷却する必要があり，データセンターにとって，その空調環境は非常に重要である．

データセンターには各種のIT機器が設置され，それらの求める動作環境は同じではない．したがって，空調設計にあたっては共通仕様として最も一般的なASHRAE*4 TC9.9サーマルガイドライン（Thermal Guidelines for Data

＊3　ブロックリダンダント方式：共通予備UPSを複数台の並列冗長方式とし，予備給電能力を高めたUPSシステム構成のこと．

＊4　ASHRAE（American Society of Heating, Refrigerating and Air-Conditioning Engineers）：アメリカ暖房冷房空調学会の略称．

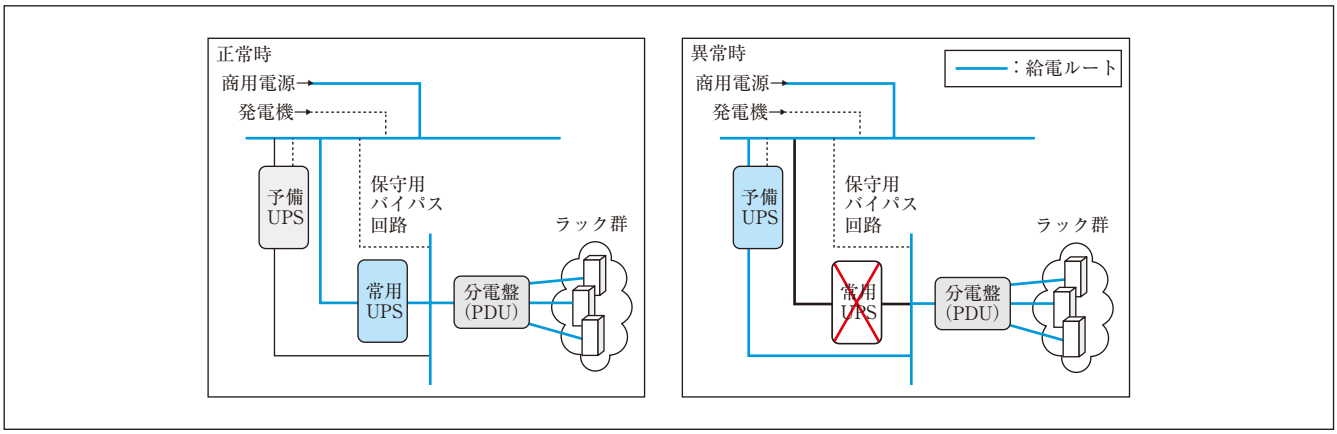


図9　待機冗長方式UPS構成

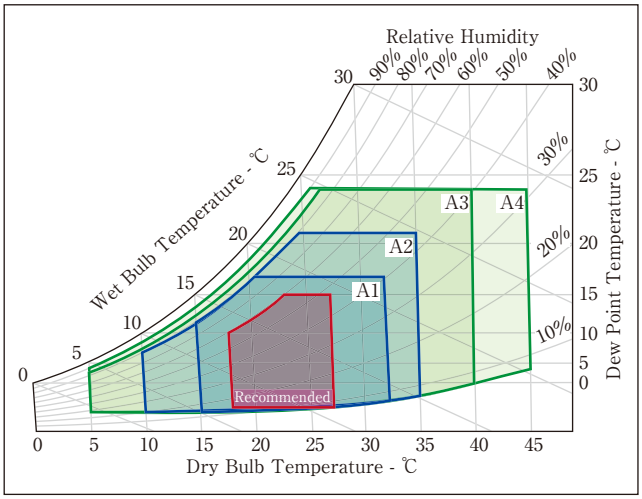


図10　ASHRAE温湿度ガイドライン（2015年）[6]

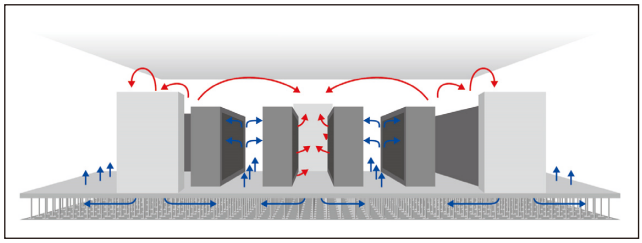


図11　一般的なデータセンターにおける空気の流れ

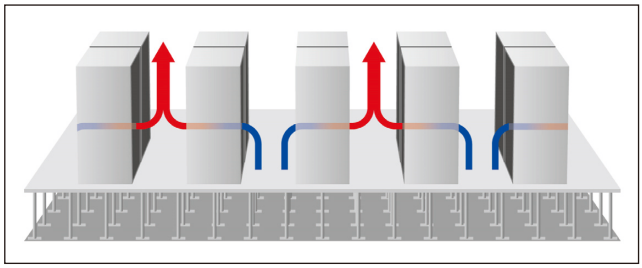


図12　ホットアイルとコールドアイル

Processing Environments 4th Ed.）で示される推奨温湿度が採用されることが多い（図10）．ASHRAEのガイドラインをもとに，データセンターでは一般的に，IT機器の吸込み温度範囲が18～27℃となるよう，設計される．

5.1　データセンター内の空気の流れ

データセンター用の空調機の多くは，ダウンブロー型が採用されている．このタイプの空調機は，上部から暖気を吸い込み，空気を冷却して，床下に冷気を送り込む．ラックを冷却するために，冷気を床下から穴あきパネルを通して空気を押し上げる．ラック内のサーバは，前面から冷気を吸い込み，ラックの後方から暖気を排出する．排出された暖気は，天井裏などを伝わって，または天井に沿って空調機に戻っていく．空調機はその暖気を冷やし，また送り込むというサイクルを繰り返している（図11）．

5.2　暖気と冷気の分離

データセンターの冷却ポイントは，サーバにいかにして冷気を当てるかではなく，サーバから出る暖気をいかに効率的に排出できるかにある．最も重要な点は，機器から排出された高温の風が回り込んで，再びサーバに吸い込まれないよう，冷気が通る道である「コールドアイル（Cold Aisle）」と，IT機器で熱せられた暖気が通る道「ホットアイル（Hot Aisle）」を分離することにある（図12）．

データセンターがまだ「電算室」と呼ばれていたような大型汎用機/ホストコンピュータの時代においては，筐体から排出される単位面積あたりの熱量はそれほど多くなく，電算室内において，空気の流れや排熱はあまり考慮されていなかった．むしろ，障害時において筐体の状態を示すランプが一目で確認できるなどといったオペレーション上の理由から，図13のように筐体やラックを同一方向に配置すること（フロントトゥバックレイアウト）が多かった．

しかしながら，ホストからオープン化の波が進み，サーバの時代が到来するにつれ，1ラックあたりの排熱量は増大していった．元来のフロントトゥバックレイアウトでは，前方にあるラックに搭載されたサーバの排熱を，その後方列にあるサーバが直に吸気してしまうことになる．今日のデータセンターでは，暖気は暖気同士，冷気は冷気同士となるよう，ラック列を対面させて立架するレイアウト（バックトゥバックレイアウト：図14）が一般的である．

サーバから排出された高温の風が回り込み，サーバに吸い込まれ再循環してしまうことを総じて「ショートサー

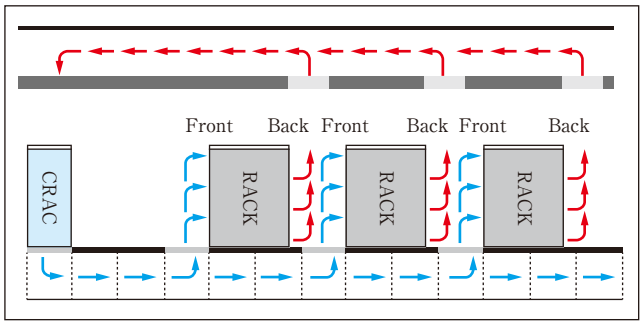


図13　フロントトゥバックレイアウト

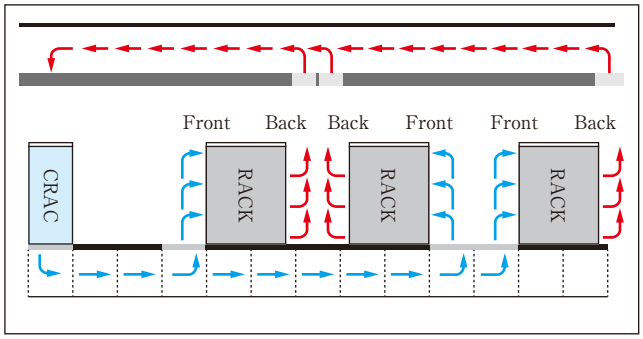


図14　バックトゥバックレイアウト

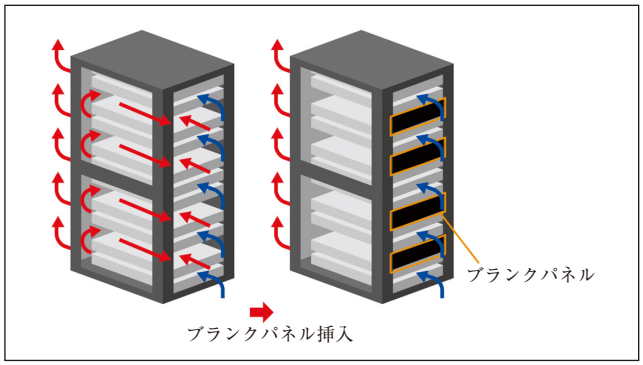


図15　サーバの隙間に生じるショートサーキットとブランクパネルのイメージ

キット」と呼ぶ．ラック内に設置するサーバとサーバの隙間がある場合，特に昨今のサーバは吸気ファンの力が強いこともありその隙間からサーバ自身が排出した暖気をそのまま吸い込んでしまうという問題も発生する．そのため，ブランクパネルなどを利用し，その隙間を埋めることも冷却管理上重要である（図15）．

5.3　高密度化/高集積化するサーバと新たな冷却の試み

これまで，一般的なデータセンターにおける空調/エアフローについて述べてきた．しかしながら，爆発的に増加するデータ量を処理するために，近年サーバはますます高性能になり，年々小型化している．クラウドコンピューティングを支える大型データセンターにおいては，IT機器を非常に高い密度で設置すること，いわゆる「高密度化」/「高集積化」が求められており，それによりデータセンターの単位面積あたりの発熱量も増加し続けている．

1999年以前のホストコンピューティングの時代では1筐体あたり1～2kW程度の発熱量であったが，2000年代，オープンコンピューティングの時代では平均2～3kW，2021年現在，クラウドコンピューティングの時代においては，1ラック当たり6～8kWが水準となってきている．さらに来るAIの時代においては10kW以上が水準となるであろうことが予想されており，現時点においても，局所的には20～30kWの冷却能力を保持することを標榜するクラウド・データセンターも増えてきた．

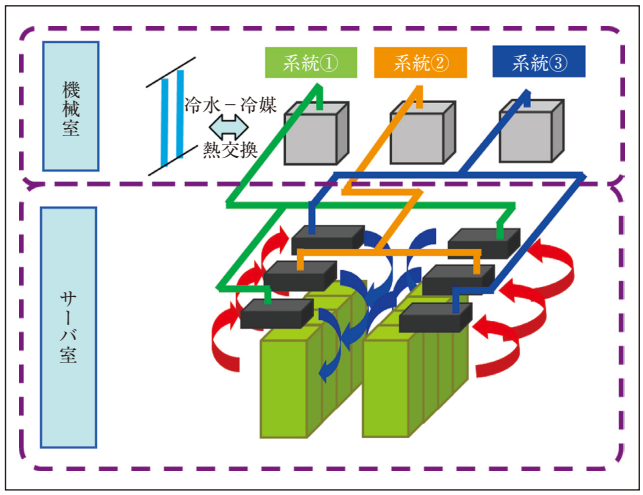


図16　天吊り型空調方式イメージ

求められる「高密度化」に対応するため，今日のデータセンター業界では，空調技術に対するさまざまな新技術の導入が試みられている．以下，主だった取り組みを紹介する．

(1) 天吊り空調方式（図16）：従来のサーバ室の端などに床置型ダウンブロー型空調機を設置し，サーバ室全体をまんべんなく冷やすのではなく，サーバラック上部に天吊で局所空調システムを設置することでサーバラックを必要最小限の循環風量で冷却する方式である．サーバラックからの発熱を直近で冷却できるため，いわゆる熱だまりを防ぐことができ，局所的には15～30kW/ラックの排熱も処理することが可能となった．天吊り空調方式は，データセンター冷却方式の選定基準の一つとしてITU（International Telecommunication Union）*5の国際標準化の勧告本文にも採用された[7]．

(2) 壁吹き出し空調方式：空調機械室とサーバ室との間にあるチャンバ室に空調機からの冷気を供給し，チャンバ室の間仕切壁の開口部から直接サーバ室内に冷気を供給する方式である．床吹出し方式で必要であった二重床（フリーアクセス）は壁吹出し空調方式では不要となるため建築コスト削減につながり，また通常の床吹出しに比べ吸気ダクトや床下ケーブルの排除などにより空気圧力の損失が少ないため，

＊5　ITU（International Telecommunication Union）：国際電気通信連合「高集積のサーバラック（例えば5～8KW/ラック以上）を収容するデータセンターでは，「天吊り型局所冷却」を選定すべきである．」

空調機の送風動力が低減可能であるとされている．

(3) 液浸冷却システム：IT機器を筐体ごと，例えば，空気の1,200倍の比熱容量をもつ冷却油などで満たした特殊なラックに浸けて冷却する液浸冷却も実用化が進んでいる．液浸冷却は，スーパーコンピュータなどの高発熱なIT機器に利用され始めており，ディープラーニングなどの普及に伴いGPGPUサーバ，FPGAサーバなどの高発熱機器に対する冷却方法として，データセンターでの普及が始まっている．

(4) 自然環境を利用した冷却環境の構築：空調環境の効率化は，高集積化の要望に応えるだけが目的ではなく，省エネルギー化，エコロジーの観点からも重要である．データセンターの消費電力の空調設備に占める割合はIT負荷についで高く[8]，空調設備の消費電力を削減することは重要である．

近年，特に寒冷地において，直接的または間接的に外気冷熱を取り入れる設備，いわゆる外気空調を実装しているデータセンターが増えている．また，雪を利用した空調システムも実用化が進んでいる．寒冷地においては，雪は廃棄物であり，やっかいものであるが，冬季に廃雪を保管し，夏季のデータセンターに利用しようとするホワイトデータセンター構想に対し，2020年環境省が支援することが決定された[9]．

寒冷地環境を利活用したデータセンターという観点でいえば，ノルウェーがその世界的な鏑矢となっている．寒冷な気候を生かしたデータセンターはもとより，山をくり抜きトンネルを利用して建築した洞窟データセンターや，データセンターから出る廃熱を利用し，生産エネルギーが消費エネルギーを上回る新たな街の建設など，ユニークかつ大胆なデータセンター/プロジェクトも数多く存在する．ノルウェーは国家戦略として「データセンター国家」を目指しており，MicrosoftやGoogleなどがノルウェーにデータセンターを建設している．

さらには，海底にデータセンターを建築する実験も進んできている．2018年，Microsoftはスコットランドの海底約36 mにデータセンターを沈めた．データセンターは864台のサーバで構成されており，ストレージの総容量は27.6ペタバイトにも及ぶものであった．海底に沈めたデータセンターは，2年後の2020年に引き上げられた．Microsoftはプロジェクトが成功したと発表し，「海中にデータセンターを沈める」というアイデアが優れたものであると実証された[10]．

6．むすび

現代社会は，あらゆる局面においてITの提供するサービスを基盤としており，その安定的/継続的稼働および発展が，社会そのものの安定と安心，進歩に直結する．クラウドコンピューティング，AIやIoT，5Gといった革新的な技術が実用化を迎え，社会の在り様にまで影響を及ぼしつつあり，ディジタルトランスフォーメンションという大きなうねりは，さらなるイノベーションをもたらすであろうことはもはや疑う余地もない．

今回は，クラウドコンピューティングを支えるデータセンターの構成要素のうち，主に電気設備および空調設備の概説を述べるにとどまったが，データセンターはその運用に際し，建築，設備（電気，空調，通信，防火・防災，衛生）のみならず，ITとしての要素（サーバ，ネットワーク，セキュリティ，オペレーション），さらには，事業継続，プロパティマネジメント，財務，保険，法令等々，ほぼ全方位的な知識が必要であり，かつ多額な費用を扱うビジネスでもある．

ITの形態そのものが生き物であり，時間とともに変化するデータセンターは，その変化をも迅速に把握し，稼働環境を対応させていかなければならない．「絶対にデータセンターを止めない」という信念のもと，データセンターに携わる人々は，社会基盤を支えている自負をもって，最適化を継続的に行う運用を確実に実施し続けている．

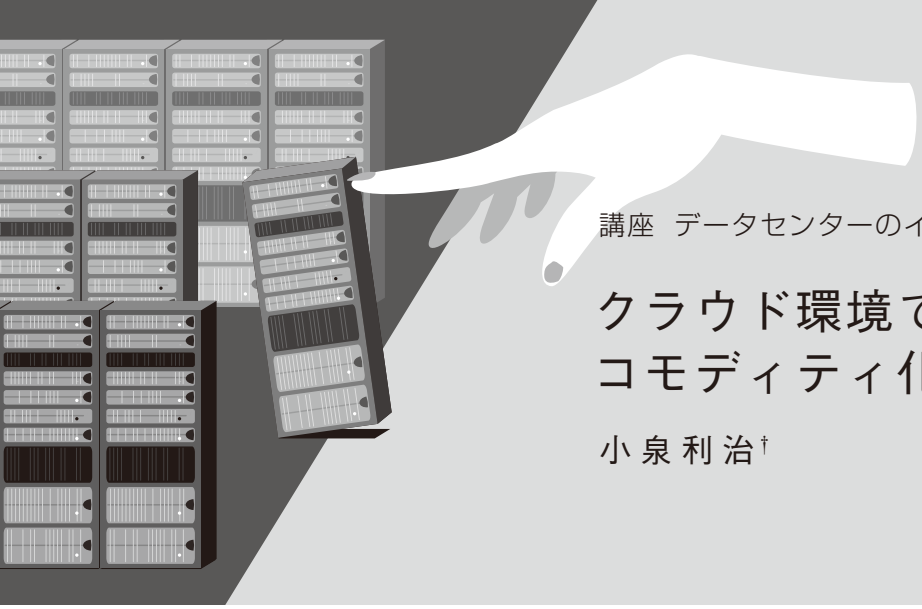
本稿により，ITを支えるデータセンターへの興味が少しでも高まれば，幸いである．

（2021年2月1日受付）

〔文　献〕

1) 経済産業省：“ディジタルトランスフォーメーションを推進するためのガイドライン（DX推進ガイドライン）Ver.1.0”（2018）
2) Z. Bauman: "Liquid Modernity", Polity Press (2000)/ジークムント・バウマン：“リキッド・モダニティ－液状化する社会－”，森田典正訳，大槻書店（2001）
3) 政府CIO：“2018年6月7日各府省情報化統括責任者（CIO）連絡会議決定『政府情報システムにおけるクラウドサービスの利用に係る基本方針』”（2018），https://cio.go.jp/sites/default/files/uploads/documents/cloud_%20policy.pdf
4) 日本データセンター協会：“データセンター運用ガイドブック”（2020）
5) PriMetrica, Inc.: "TeleGeography Submarine Cable Map"（2021），https://www.submarinecablemap.com/
6) ASHRAE TC9.9: "Thermal Guidelines for Data Processing Environments", Fourth Edition, ASHRAE, Atlanta Georgia (2015)
7) ICU-T L.1300 8 Cooling 8.3.2 High Efficiency Cooling Plant: "Recommendation ITU-T L.1300 - Best Practices for Green Data centers", International Telecommunication Union（Nov. 2011）
8) M. Salim, et al.: "Data Centers' Energy Auditing and Benchmarking-Progress Update", ASHRAE Trans., OR-10-013, Orlando (2010)
9) 環境省：“2020年9月11日『脱炭素イノベーションによる地域循環共生圏構築事業（一部総務省・経済産業省・国土交通省連携事業）』”（2020），http://www.env.go.jp/earth/earth/ondanka/energy-taisakutokubetsu-kaikeir03/matetr03-05.pdf
10) https://news.microsoft.com/ja-jp/features/200915-project-natick-underwater-datacenter/

中西　草太（なかにし　そうた）　2000年，関西大学社会学部産業社会学専攻卒業．2002年，同大学院修士課程修了．同年，CRCソリューションズ（株）（現 伊藤忠テクノソリューションズ（株））入社．以来，データセンターの構築・運用に従事．現在，同社ハウジングサービス課長



講座　データセンターのインフラ技術〔第３回〕

クラウド環境で使われる機材のコモディティ化

小泉利治†

1．まえがき

スマートフォンの普及により，インターネットを介した多くのサービスを利用することが一般的となっている．SNSを介した情報のやりとりや，宅配サービス，オンラインゲームなど日々の生活の中で，スマートフォンは必需品となっている．スマートフォン普及前は，顧客との接点は人もしくはテレビや新聞などの不特定多数の人に向けた情報発信だけであったため，企業は社内の業務を円滑に進めるためのシステムのみを運用することが多かった．

しかし，顧客ニーズの変化にともない，多くの企業が不特定多数のユーザに向けた情報やサービスの提供だけではなく，特定ユーザに向けたサービスを提供するために，今まで運用してきた業務効率化のシステムに加えて直接顧客のニーズに応えるための新たなシステムを構築し運用する必要が出て来た．これら新たなシステムは，顧客の需要により利用状況が変化するため，従来のオンプレミス環境にシステム構築するのではなく，柔軟にシステムのリソースを変更できるクラウド環境の活用が進んでいる．このようなニーズを即座にとらえ大きくビジネスを拡大させた企業としてFacebook, Amazon, Google, Microsoft等がある．スマートフォンユーザ向けのサービスや，企業向けのクラウドサービスを全世界に提供することで業績を大きく伸ばしている．自社サービスを提供するために巨大なインフラ環境を持つ企業がどのようなハードウェア機材を使用し，運用しているのか以降にて解説する．

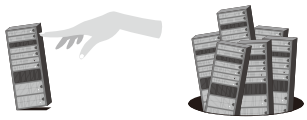
2．クラウド環境で求められるハードウェア機材の要件

企業向けのクラウドサービスやインターネットサービスを提供するためのインフラ環境（以降，これらインフラ環境の総称をクラウド環境と呼ぶ）は，利用者の増加に合わせて拡張が必要となる．小規模で始まったサービスであっても，そのサービスが普及するとインフラ環境も急激に拡大していくことになる．世界各国でサービスを提供するためのシステムともなると，世界中のデータセンターにインフラ基盤を持つような巨大な環境となる．従来型の業務系システムでは，そのシステムを使用する期間でどれぐらいの拡張が必要となるのか予め予測することができたので，システムの構築を行う前に将来的に必要なリソースを試算し，システムの拡張をせずに運用が可能なシステムを構築してきた．しかしクラウド環境ではシステム構築時に将来的にどれぐらいサービスが普及するか予測することが困難なことが多いため，インフラ基盤についても利用状況に合わせて拡張ができるように設計を行うことが必須の要件となる．

そのため，サーバ，ストレージ，ネットワークの増強が必要となった場合，筐体内にあるCPU，メモリー，ディスク，ネットワークポートを拡張するのではなく，サーバ，ストレージ，ネットワークスイッチの機器自体を拡張していくスケールアウト型のシステムにしていくことが一般的となっている．このスケールアウト型の拡張方式については，無限の拡張性を実現できるのと同時に可用性も高めることが可能となる．従来の業務系システムにて可用性を考慮する場合，アクティブ，スタンバイのクラスタ構成をとることが一般的であった．クラスタ構成では，アクティブなサーバに不具合が発生した場合は，スタンバイのサーバに切り替えて処理を再開することになるが，切り替え時に瞬断が発生することと，クラスタを構成するサーバ台数に制限があるため拡張性が必要なクラウド環境では使われない方式となっている．スケールアウト方式にて可用性を向上させる場合は，アプリケーションもしくはミドルウェアが1台のサーバ障害がシステムに影響を与えない仕組みとなっている必要があるが，クラウド環境のシステムでは数千，数万のサーバが並列に処理実行しており，可用性も担保できる仕組みが実装されていることがほとんどである．このようなスケールアウト型のクラウド環境においては，ハードウェアに求められる要件も変わってくる．システムとして可用性が担保されているのであれば，今までハードウェアにて実装されていた全コンポーネントの冗長化は不

†伊藤忠テクノソリューションズ株式会社
"Infrastructure Technology for Datacenters (3); Commoditization of Cloud Infrastructure" by Toshiharu Koizumi (ITOCHU Techno-Solutions Corporation, Tokyo)

映像情報メディア学会誌 Vol. 75, No. 4, pp. 495～499（2021）



要となる．過剰なハードウェアの冗長化は，クラウド環境ではアクティブ，スタンバイ構成により使えるリソースが半分となってしまうこともあり無駄と判断されることが多い．次に，クラウド環境ではインフラに関わるコストをできる限り抑えていく必要がある．提供するサービスの普及が進んでいくとインフラ環境も増設を繰り返していくことになるが，その増設の規模も大きくなることが多い．サービスの収益面を考慮するとできるだけシステムにかかるコストを抑えていきたいという要件は強くなる．システムに関わるコストでは，調達コストはもちろんであるが，メンテナンスや運用に関わるコストも考慮される．

従来型の業務系システムでは，ハードウェア機材の製造メーカもしくはシステムインテグレータ等が提供する保守サービスを契約し，オンサイト交換作業等のメンテナンス作業に自社エンジニアを使わないことが一般的である．上述の通り，従来型の業務系システムは，ハードウェア機材障害時にシステムに影響を与えることが多く，速やかに復旧する必要があることから24時間365日で対応できる体制を整備しておく必要があった．しかし，クラウド環境においては，ハードウェア機材の障害がシステムに与える影響は限定的となるため，即座にハードウェア機材を復旧させる必要がなくなる．多くのクラウド環境では，保守契約は締結せずに自社で保守用の交換部材を保持し，任意のタイミングで自社エンジニアがハードウェア交換，ハードウェア復旧作業を行うことで，メンテナンスコストを抑えることを実現していることが多い．また，運用時のコストという面でなるべく省電力な設計であることが求められる．数千，数万というハードウェアを運用した場合の電気代は膨大な金額となる．昨今のハードウェア機材の高スペック化に伴い，機材が消費する電力も増加傾向にある．また，GPUサーバ等の電力を多く消費するサーバの出現により電力はクラウド環境の運用コストにおいて大きなインパクトを与えるようになっている．機材の消費電力の増大はデータセンターの集積率にも影響を与えるため，なるべく高スペックで省電力を実現可能なハードウェア機材というのが求められる．これら要件を実現するために，多くのクラウド環境では，従来のハードウェアメーカからODM(Original Design Manufacturing)と呼ばれる，製造メーカからハードウェア機材の調達を行うことが一般的になっている．

3．ODMの台頭とODM製品の特長

ODMはハードウェアメーカからの委託を受けハードウェアの製造を行っていた．ODMと似た使い方をされる，OEM(Original Equipment Manufacturing)やEMS(Electronics Manufacturing Service)などもある(**表1**)が，OEMは製造(場合によっては設計も)をODMに委託し，自社ブランド製品として販売する事業モデルとなり，EMSは設計および製造を行うが自社ブランド製品としては販売を行わない事業モデルとなる．

表1　製造モデルの比較

	設　計	製　造	自社ブランド製品
ODM: Original Design Manufacturing	○	○	○
EMS: Electronics Manufacturing Service	○	○	×
OEM: Original Equipment Manufacturing	○	×	○

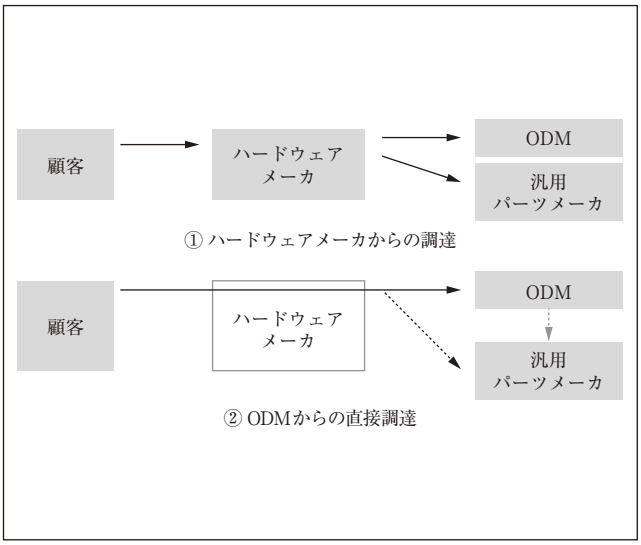


図1　調達モデル比較

ODMの多くは台湾企業である．中国と日本にアクセスが良いこともあり，日本の製造メーカの製造委託を受け，同じ言語でコミュニケーションができ，比較的安価に労働力の確保および工場の建設が可能な中国の工場でハードウェア機材の製造を行っている．安価にハードウェアを製造することが可能であることから，クラウド環境をもつ事業者が図1で示すようなODMから直接ハードウェア機材を購入することでコスト削減を実現した．

ODMからの直接機材調達には2つのコスト的なメリットがある．1つ目は調達コストを下げることができる点である．前述した通り従来企業にハードウェアを提供いていたメーカの多くは，ハードウェアの製造をODMに委託を行ってきた．昨今のODMからの直接調達はハードウェアメーカを介さない調達フローにより，ハードウェアメーカが乗せていたさまざまな管理コストや利益を削減できるため，大幅なコスト削減が可能となる．また，CPU，メモリー，ディスク等の汎用パーツについても同様に個別調達を行い，ODMが製造した機器にODM側で汎用パーツも組み込むといった取り組みも合わせて行うことで更なるコスト削減を実現している企業も多い．

2つ目のメリットとしては，自社の環境にあったハードウェア機器にカスタマイズすることが可能となることである．

従来型の業務システムで使用される汎用的なハードウェア機器は，クラウド環境で使われるハードウェアとしてはオーバスペックであることが多い．前述のとおりクラウド環境のシステムは，スケールアウト型の構成によりアプリケーション含めた可用性担保の仕組みが組み込まれているため，ハードウェア自体の冗長化をそれほど必要としない．現在一般的に販売されているハードウェア機器は従来の業務システムで使用されることを想定して設計されているものが多く，交換可能なコンポーネントがすべて冗長化されており，単一コンポーネントの障害がハードウェア機器のダウンに繋がらない仕組みとなっている．このような機器はクラウド環境ではオーバスペックであり，冗長化によるパーツ数，消費電力の増大により調達コスト，電力コスト増に繋がってしまう．これら課題を解決するために，オーバスペックとなっているスペックを取り除く自社仕様のハードウェアをODMに設計含めて依頼し，調達を行っている企業が増えてきている．ただし，自社仕様にカスタマイズされた機材を設計するためには，多くの費用が必要となるため，少なくとも数千クラスの機材調達を行うことが必須となる．また，製造においても個別の製造工程となるため，1回の調達量もそれなりのボリュームが必要となる．小ロットの製造はコスト増および長納期化に繋がるので注意が必要となる．このような，クラウド環境に適したハードウェアの設計を行うという点で参考になる取り組みがある．Facebook社が立ち上げたOpen Compute Projectである．

4．Open Compute Projectとは

Open Compute Projectは，2011年にFacebookが設立したハードウェアのオープン化を推進する非営利組織として発足したコミュニティである．Facebookは日本国内においても知名度の高いSNSを運営提供している企業である．Facebook社のサービスは全世界に提供され一日に18億を超えるアクセスユーザ数があり，継続的にユーザ数を増やしている．これらユーザの拡大やサービスの充実を実現するために世界中の多くのデータセンターにサービスを提供するためのインフラ環境を持ち運用している．Facebookも当初は汎用的なハードウェア機材をハードウェアメーカから購入していたが，急激なユーザの増加によりインフラ環境も多くの増設が必要となる中で，無駄を極力省いていくことを目的に，自社仕様のハードウェアをODMより調達するようになった．そして，自社仕様のハードウェアをオープン化し，自社仕様のハードウェアを同様の悩みを抱える他事業者にも使用してもらうことで自社仕様ハードウェアのコモディティ化に成功した．FacebookはOpen Compute Projectの中で自社仕様のサーバ，ストレージ，ネットワークスイッチ，ラックなど多くのハードウェア仕様を公開している（図2）．

このハードウェアスペックの公開は，新たにFacebook仕様のハードウェアを採用する事業者，Facebookの双方に大きなメリットをもたらすことになる．新たにFacebook仕様のハードウェアを採用する事業者にとっては，Facebookが大規模クラウド環境向けに設計したハードウェアを，設計費用をかけずに購入することができる．また，Facebookのシステムで運用実績がある機材ということになるので，安心感も得られるであろう．Facebookについても，他事業者が同じ仕様のハードウェアを採用することで大量生産が可能となりコスト削減を実現した．また，Open Compute Projectに加盟する同じような悩みを抱える事業者と共同での新たなハードウェア仕様化の推進や，他事業者が考えた優れた仕様を自社ハードウェアに取り込むことも可能となる．Open Compute Projectは従来のハードウェアメーカ主導のハードウェア設計からユーザ主導のハードウェア設計への変革を後押ししたコミュニティとして大きな役割を果たしている．

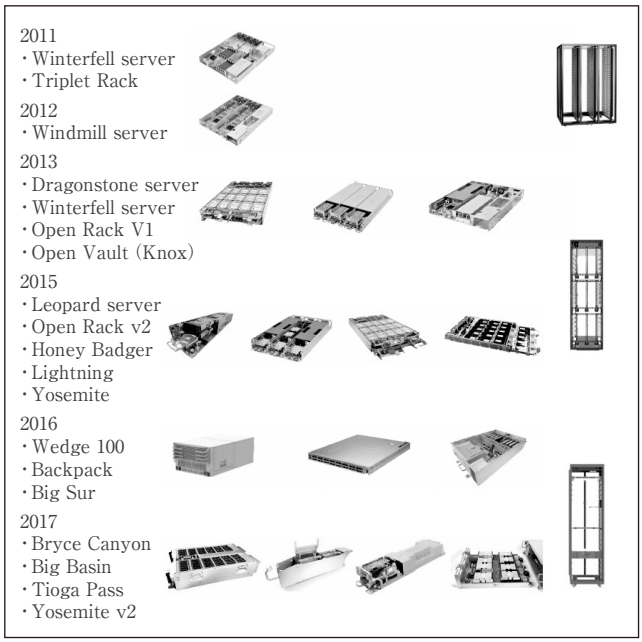


図2　Facebookが仕様公開したハードウェア

Open Compute Projectでは多くのハードウェア機器のスペックが公開されているが，代表的なハードウェアとしてラックシステム，サーバがある（図3）．共に省電力を考慮した設計となっており，ラックシステムには集中電源方式を採用することで，省電力化を実現している．また，サーバについても大規模クラウド環境で使用するハードウェアとしては無駄のない設計となっており，低消費電力と集約率の向上を実現している．

サーバ，集中電源ラックシステム等のハードウェア調達コストなどの設備投資コストが含まれるCAPEX（Capital Expenditure）で約5%～10%，電気代やデータセンタコストなどの運用コストが含まれるOPEX（Operating Expense）で約40%のコスト削減が可能となるケースもある（図4）．

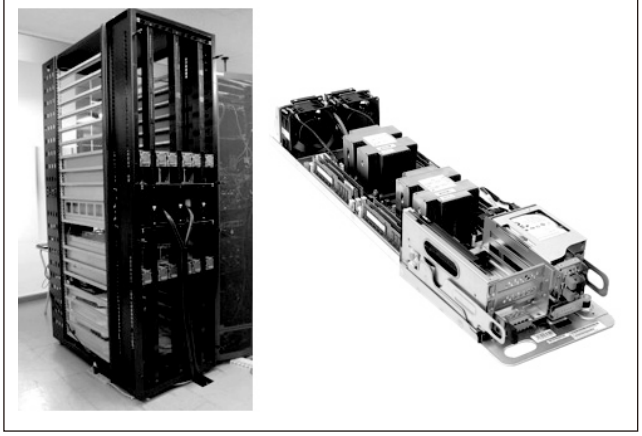


図3　Open Compute Project仕様のハードウェア

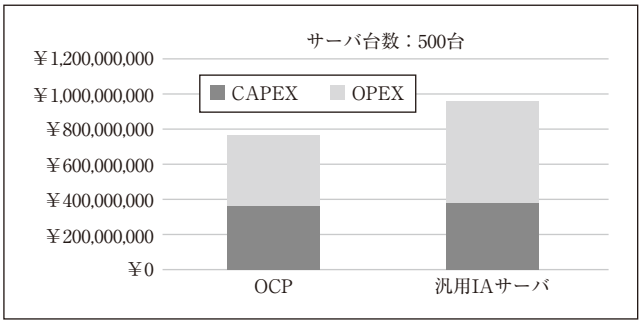


図4　コスト比較

また，日本国内においてもOpen Compute Project Japanが活動しており，日本国内でのOpen Compute Projectに関する情報発信および日本国内のデータセンターにOpen Compute Projectのハードウェアが使用できるようにラックシステムおよび集中電源の製品化の支援を行ってきた．耐震性能を考慮したラックシステムや単相電源をサポートした集中電源がすでに製品化されており，日本のデータセンターにも容易に設置ができるようになっている．

5．ODM製品の課題

クラウド環境を持つ事業者にとっては，調達コストや電力コストを抑えることができるODMからの調達は魅力的ではあるが，日本国内の事業者がこれら製品を採用する際はいくつかの課題が存在する．

1つ目は，ハードウェアの保守メンテナンスをどのような体制で行うかである．日本企業の多くはハードウェア機器の保守メンテナンスをハードウェアメーカもしくはシステムインテグレータが実施していることが多く，自社に保守メンテナンスを行うエンジニアリソースが存在しないことが多い．ODMから提供されるサポートとしては，交換用パーツの送付のみを行うセンドバックサービスとなるケースがほとんどであり，パーツ交換の作業まで行うオンサイトサービスが可能であったとしても，高額になることが多い．センドバックのサービスについても故障パーツを海外のパーツセンターに送付し，修復した後に返却するセンドバックサービスがほとんどであるため，ある程度のパーツを自社でストックしておく必要がある（表2）．

表2　保守レベル比較

	期　間	対応時間	対応レベル	言　語	パーツ返却
汎用製品	3年間	平日営業時間帯	翌日以降オンサイト	日本語	国　内
ODM	3年間	平日営業時間帯	センドバック	日本語	−

インターネットサービスを提供するインフラ環境については，前述したとおりスケールアウト構成によりある程度の余裕をもって計画的にメンテナンス作業を行うことが可能であるため，比較的容易に交換作業を行うためのリソースを確保することが可能であるが，日本国内でのクラウド事業者が持つインフラ環境は従来の業務システムに近いサービスレベルの提供を求められることも珍しくない．日本国内においても企業システムのクラウド化が進む中で，業務系のシステムをクラウドに移行するケースも増えてきている．日本国内のクラウド事業者は業務系システムのクラウド化をターゲットにしている場合も多く，ODMが提供するセンドバックのサービスだけでは足りないということも少なくない．将来のシステム拡張を見据え，保守メンテナンスの人員を確保し，自社リソースにて対応を行っていくことが一番コスト的なメリット得られる方法ではあるが，すぐには体制を整備するのは難しい場合は，国内でODM製品を販売している事業者に依頼することも可能であるため，相談してみるのもよいであろう．

2つ目は品質面である．ODMからハードウェア機材を直接購入にすることにより，ハードウェアメーカが介在しなくなりコスト削減を実現することができるが，従来はODMが製造したハードウェアを，ハードウェアメーカの品質管理により，品質を担保してきた．ODMからの直調達のモデルでは，この品質管理をODMもしくは調達を行う事業者が行う必要がある．品質管理については，ハードウェアメーカの役割であったため，ODMには知見が少ないケースがあるのと，日本企業が求める品質基準とはかけ離れている可能性もあるので注意が必要である．調達した機器の品質状況はトレースできるように管理を行い，品質が悪化傾向にある場合は，定量データを提示しながら改善要望を行う必要がある．

3つ目は調達規模である．1回の調達規模が数千台といったような規模であれば，ODMからの直接調達はメリットが多いが，自社仕様のハードウェアの設計，製造できる調達規模に達しているクラウド環境は日本国内においてはそれほど多くはない．ハードウェアを自社仕様として設計するとなると，数千以上の調達規模が必要となる．しかし，ODM各社はクラウド環境に適したOpen Compute Project

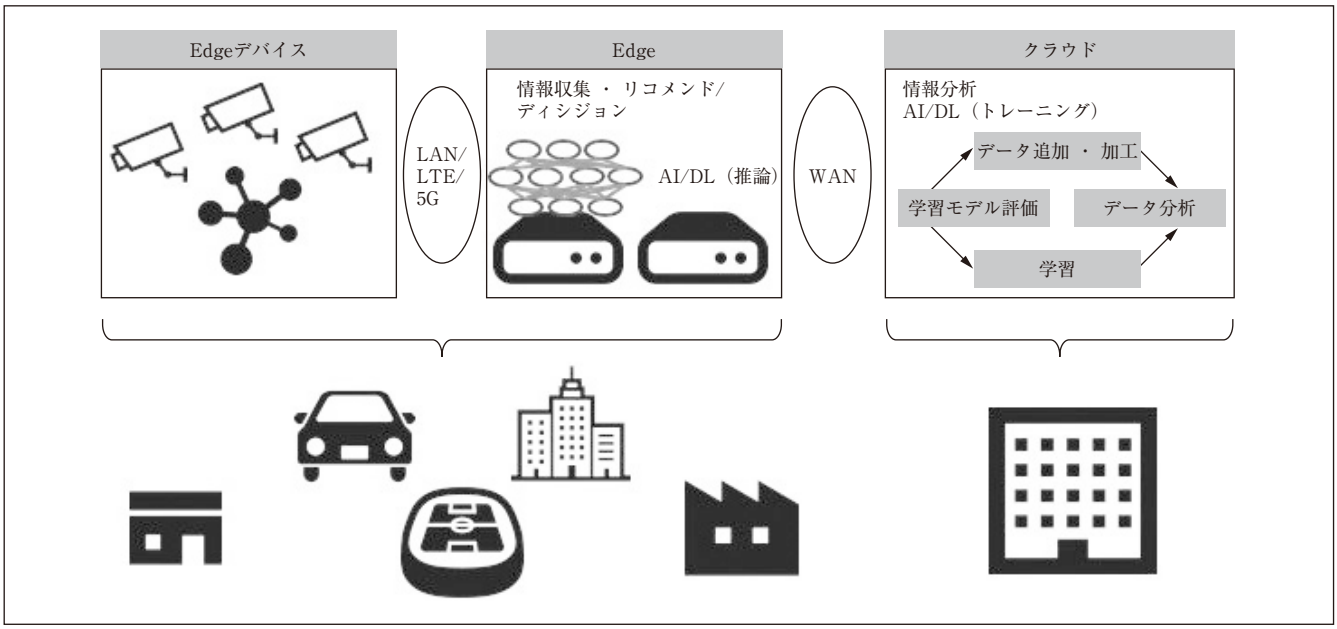


図5　エッジコンピューティング概略図

で設計された製品も含めて汎用的なハードウェアの提供も行っているので，ODMの汎用モデルであればODMからの直接調達のメリットを活かしながら，調達数を下げることも可能である．

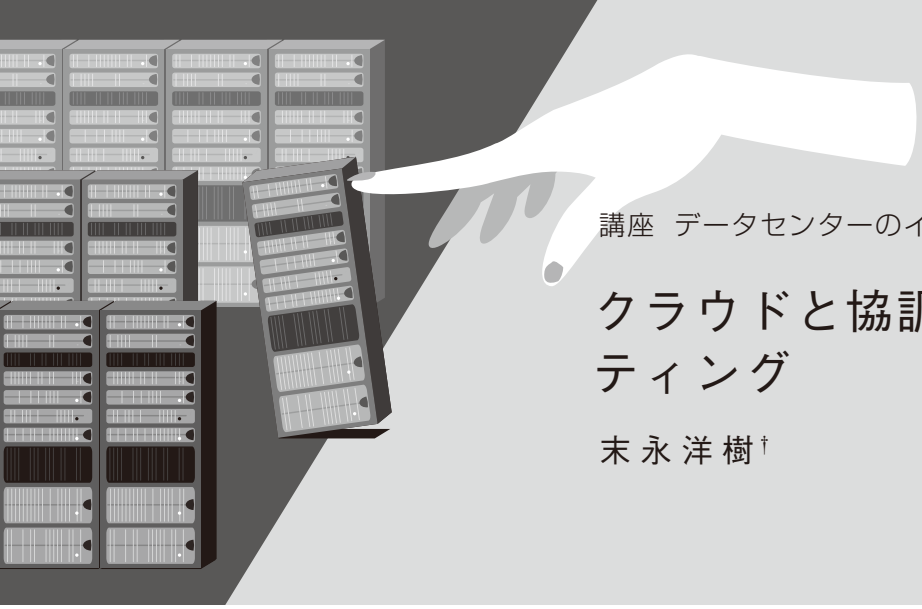
6．今後クラウド環境で使用されるハードウェアとは

現在多くのクラウド環境で使われているハードウェア機材はコモディティ化が進み，コモディティ化したハードウェアはODMがその製造を担い，製品として事業者に直接提供するというビジネスモデルになりつつある．このクラウド環境についても多様な活用モデルが生まれてきている．今後インフラ環境の構成を大きく変える可能性があるのがエッジコンピューティングであろう（図5）．

昨今のAI, Deep Learningの活用拡大により，店舗，工場，市街地，車などにカメラやセンサ等のエッジデバイスを設置し，エッジデバイスのデータを収集，情報分析をすることでリコメンデーションやディシジョンを行うことが可能となっている．エッジコンピューティングでは，リアルタイム性を求められることが多く，エッジデバイスになるべく近い場所で推論の処理を行う必要がある．今までクラウド環境で利用していた大規模なデータセンターではなく，店舗や施設にある小規模なマシンルームにサーバ，ストレージ，ネットワークスイッチといったハードウェア機材を設置する必要が出てくる．データセンターと比較すると電源環境や温度や湿度など厳密に管理されていない環境も多く，限られた設置スペースの中にハードウェア機材を設置する要件も増えてくる．今までのデータセンターに設置するハードウェアと異なり，小型化されたサーバや稼働温度，湿度のスペックが緩和されたハードウェアが求められることになる．また，データセンター側にもAI, Deep Learningの情報分析を行うために，従来のサーバより消費電力が高いGPUサーバを設置する必要が出てくる．インフラ環境で使用するハードウェア機材についても多様化が進むことになるが，ハードウェア機材を調達する規模も大きくなるため，ここでもハードウェアの製造はODMが主役となる可能性が高い．ハードウェアに多くの機能を求めるのではなく，ハードウェアはシンプルにそして安価にということが求められる流れの中で，多様化とコモディティ化を両立させることが求められる．そしてその要望をかなえているのが，現時点ではODMということになるので，このビジネスモデルは当面続くであろう．そして事業者はコモディティ化されたハードウェア機材を使い，顧客ニーズに応じたサービスを作り上げていくことが競合との棲み分けになっていくのであろう．

（2021年3月29日受付）

小泉（こいずみ）　利治（としはる）　2000年，伊藤忠テクノサイエンス（株）（現，伊藤忠テクノソリューションズ（株））入社．サン・マイクロシステム製品を始めとしたハードウェアメーカ製品の主管業務を従事．2015年より，オープンコンピュートプロジェクトジャパン（OCPJ）の副座長として活動中．



講座　データセンターのインフラ技術〔第４回〕

クラウドと協調するエッジコンピューティング

末永洋樹†

1．まえがき

2010年代の初頭頃から，クラウドコンピューティングを補完しうる仕組みとしてエッジコンピューティングの研究開発が活発化している．エッジコンピューティングを大まかに表現すると，データセンタに構築されたクラウドでもユーザが手元で使っている端末でもない場所で計算機リソースを提供するアーキテクチャである．インターネットのアーキテクチャでは，データを転送するノードとデータを処理するノードとは必ずしも区別する必要はない．データセンタとユーザの間には広大なデータ処理環境を想定しうる．このため，エッジコンピューティングは対象範囲の広い漠然としたアーキテクチャである．本稿では例を挙げつつ，具体的なイメージを持てるよう解説していく．

一方で，一部の例にとらわれることなくエッジコンピューティングを理解・評価するための考え方も紹介したい．対象範囲の広大さはエッジコンピューティングの本質である．特定の対象にとらわれすぎてしまうと価値を見誤る可能性がある．

2．なぜエッジコンピューティングが注目されたのか

エッジコンピューティングが注目された背景として，IoT関連の技術開発の進展が挙げられる．IoTではインターネットに接続する小さなモノが大量に生まれると考えられる．これらのモノが送信する個々のデータは必ずしも大きくはないが，モノの数が多いためデータの総量は多くなる．IoTにより生成される大量のデータは，クラウドのストレージと通信インフラに対する負荷となる．総務省通信白書によると，IoT機器の数は2016年に187.3億台，2019年に253.5億台，2022年に348.3億台（予測値）である[1]．IoTによる通信量，データ量の変化をふまえ，既存のクラウドコンピューティングに対しいくつかの懸念事項が指摘された．

† 株式会社IIJイノベーションインスティテュート技術研究所 研究企画室
"Infrastructure Technology for Datacenters (4); Edge Computing Integrated into the Cloud" by Hiroki Suenaga (IIJ Innovation Institute Inc, Tokyo)

データの送信先がクラウドであるとすると，小さなデータが大量にストレージへ流れ込んでいくこととなり，ストレージに対する要求性能が高くなりやすい．もちろん，本講座において過去に紹介してきたように，クラウドのデータセンタ技術はスケールアウトを前提とした構成を取っており，データの流量が増えたからと言ってすぐに対応できなくなるようなことはない．しかし，IoTの進化速度がストレージ技術の進化速度を上回ってしまった場合，コストが急激に上昇する可能性がある．一般に利用しているストレージデバイス（SDカードなど）を見ても，安く調達できる性能と容量との組み合わせはある程度決まっているものである．そこから外れてしまうと価格は急激に上昇する．これがクラウドで発生すると，IoTの利用コストが上昇してしまうと懸念された．

IoTで使われる通信路についても，データセンタ内部では複数のモノからの通信が合流して大きな流れとなりコストの問題を生じさせる（または充分なコストをかけられず，ユーザの利便性を損なう）可能性がある．他方，データセンタと逆側のモノに近い領域でも通信路に問題が生じる可能性がある．IoTデバイスはスマートフォンのように強力な広域ネットワーク接続能力を持っているとは限らない．仮に4Gや5Gなどの広域ネットワークへの接続機能を持っていたとしても，電力消費と価格を抑えるために性能を制限せざるをえないことも考えられる．このため，すべてのデータをIoTデバイスからクラウドへ直接転送するだけの通信容量の確保は現実的ではないのではないかと懸念された[2]．

モノに近い領域での通信では遅延が問題となる可能性もある．IoTで取り扱われるモノは，物理的な世界の様子を観察するセンサと，物理的な世界に影響を及ぼすアクチュエータとに分けて考えることができる．これらを組み合わせることで利便性を提供することがIoTの骨子である．このとき，システムの動作速度は通信の遅延に大きく制約される．車を例にすると，速度計はセンサであり，速度を制御するアクセルとブレーキがアクチュエータである．車が速度計の指示値をクラウドに送信し，前方の事故を検出したクラウドがブレーキ信号を車に返信するような自動運転システムがあったとしても，追突を回避できる保証はない．

映像情報メディア学会誌 Vol. 75, No. 5, pp. 621～626（2021）

インターネットはこのような極めて高い即時性・信頼性を提供できるネットワークとしては設計されていない．一方で，車の速度計の指示値の統計やブレーキ操作の統計を集積するような処理であれば即時性は必要ない．このような用途にはインターネットとクラウドは向いている．IoTでは既存のクラウド技術が得意とする処理と不得意とする処理が混在するケースが増えるのではないかと懸念された[3]．

通信技術の周辺ではセキュリティとプライバシーも懸念材料になりうる．IoTで扱われるデータには，企業内部の活動を推測しうるデータや，個人のプライバシーに関わるデータも含まれる．一つ一つのデータは，おそらく断片的であり一見すると価値がないようにみえるかもしれない．また，クラウドにおける情報漏洩事故はときおり報告されるものの，頻繁に発生しているわけではない．しかし，データの断片を丁寧に集めて再構成することは可能であり，そこから高精度の推論ができる可能性はある．インターネット上に送信するデータを匿名性の高い最低限のデータに絞ることができれば，こうしたリスクを低減できる．このような目的を実現する仕組みとしてエッジコンピューティングが注目されることもある[4]．

図1にエッジコンピューティングの大まかなコンセプトを示す．IoT機器やユーザ端末などの近傍にクラウドコンピューティング向けのデータセンタ設備（以下，クラウド設備）を簡略化したエッジコンピューティング設備を設置することでここまで見てきたような懸念に対応する．エッジコンピューティング設備にストレージを用意することで，クラウド設備に送信するデータ量の削減を見込む．削減の前提としてIoT機器の生成するデータのすべてが長期保存されるわけではない，もしくは長期保存すべきではないという仮定がある．利用後に保存せずに捨てるべきデータであれば，エッジコンピューティング設備で処理してしまうことでクラウド設備への送信を省略できる．これによりクラウド設備のストレージとネットワーク資源に対する負荷を軽減できる．

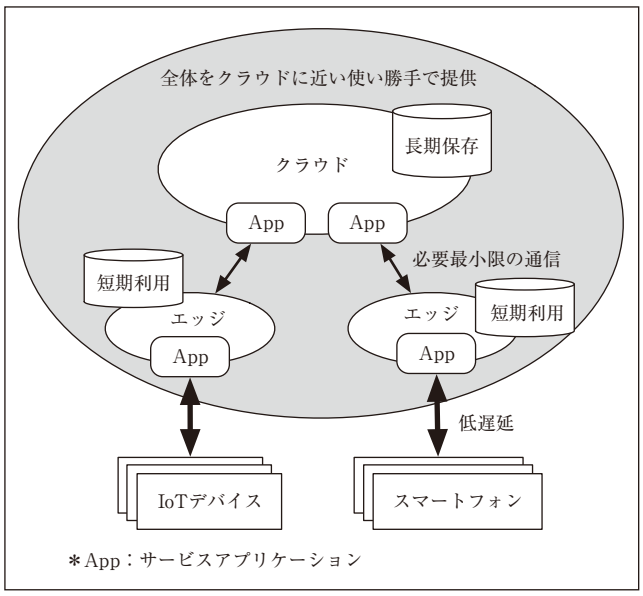


図1　エッジコンピューティングのコンセプト

エッジコンピューティング設備をIoT機器やユーザの近傍に設置することにより，通信遅延もある程度コントロールできる．クラウド設備との通信は通常インターネットを経由するため，通信遅延のコントロールは困難である＊1．しかし，IoT機器やユーザ端末のネットワーク構成を良くみてみると，品質をコントロールしやすいネットワーク技術を利用してインターネットゲートウェイに接続している場合がある．携帯電話網による接続が好例である．このIoT機器やユーザの端末が直結している網にエッジコンピューティング設備を設置できれば，通信遅延は網の仕様として与えられ予測可能な範囲に収まる．網の品質と信頼性はデータの秘匿にも寄与する．

3．エッジはどこにあるのか

ここまでクラウドの反対側の端，という程度の意味でエッジという言葉を使ってきた．これは曖昧な定義であり，実際に何を持ってエッジとすべきか明らかではない．例えば，クラウド事業者の立場では，クラウド設備とインターネットの境界部分はエッジであると考えることができる．一方，ユーザの立場からすれば手元の端末とインターネットの境界がエッジである．明確な境界をイメージさせるような言葉でいて，広く合意できるような実体はない．

エッジコンピューティングの対象とする「エッジ」の比較的わかりやすい定義として，物理現象とディジタルデータの境界部分を「エッジ」と考えるという方針がある[5]．例えば，カメラで撮影した映像をネットワーク経由で送信する場合を考える．カメラに対する入力は光の挙動であり，これは物理現象である．カメラからの出力は光の挙動をサンプリングしたディジタルデータであり，このシステムはエッジを構成すると考えられる．システムの送信したデータはメディアに記録されたり，クラウドに集積されたりする．データの集積場所は大元の物理現象からは地理的にも時間的にも遠く離れており，「エッジ」とは呼びがたい．図2のようにIoT機器やユーザ端末とクラウドとの間に複数の設備がある場合，物理現象に最も近い設備Cが「エッジ」になりうることは比較的わかりやすい．

さらに「エッジ」を拡大することもできる場合がある．先のカメラの例にカメラを遠隔制御する制御システムを追加

＊1　インターネットプロトコルと互換性のある閉域網サービスも存在する．閉域網であればインターネットと違い細かなコントロールが可能である．しかし長距離通信に利用する際のコストはインターネットよりも高く，IoT機器の接続においては選択しにくいと考えられる．IoT機器が直接接続する携帯電話網やFTTH網も閉域網である．インターネット接続では閉域網は直近のインターネットゲートウェイまでの短い区間でのみ利用する．

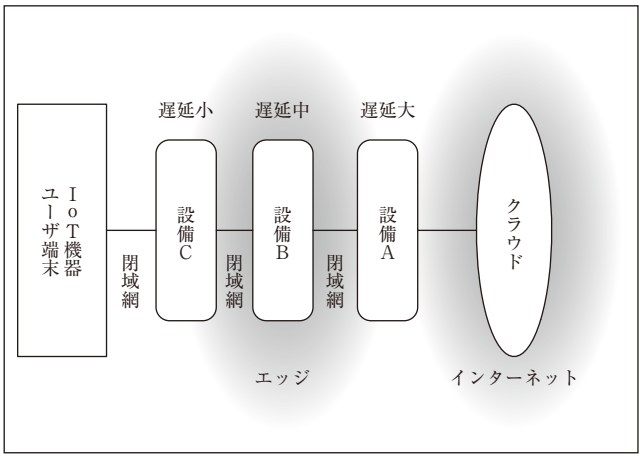


図2　エッジの範囲

した場合，この制御システムも「エッジ」を構成する可能性がある．この場合，カメラから受け取るディジタルデータの伝送遅延が問題となる．充分に短い遅延でデータを受け取れるのであれば，物理現象にリアルタイムに反応するシステムが構築できる．このシステムは実質的に物理現象に接していると考えられ，「エッジ」を構成する．伝送遅延が大きくなってくると物理現象にリアルタイムに反応することはできなくなり，「エッジ」とは言えなくなってくる．

「エッジ」の範囲は伝送遅延に関連すると言えそうである．どの程度の伝送遅延であれば「エッジ」と呼べるのかは用途によって異なる．このため，エッジコンピューティングのアーキテクチャにおいては許容できる伝送遅延を指定する仕組みが組み込まれていることが多い．これにより，用途に合わせたエッジを定義できる．図2において，設備Bの伝送遅延が充分に短いのであれば設備Bも「エッジ」の一部として利用できる．

物理現象とディジタルデータ以外の視点から「エッジ」を考えることも可能である．取り扱うデータを秘匿する必要がある場合，データの流通可能な範囲が「エッジ」の範囲を強く制限する．監視カメラで人物を検出するようなシステムを考えると，撮影した画像には高度の秘匿性が求められ，システム外には送信できない可能性が高い．外部への送信が許されるとすると，検出結果をYes/Noだけに絞るなど，部外者がみても意味を推測できないようなデータに限られるだろう．図2において，設備Bや設備Aの伝送遅延は大きい．一方で閉域網接続であり通信の秘匿や安定性は担保できる．伝送遅延が問題とならず，通信の秘匿・安定性が問題なのであれば設備Bや設備Aも「エッジ」として利用できる．

すべての要素を列挙することはできないが，他にも「エッジ」の尺度は考えられるだろう．エッジコンピューティングに触れる場合，どのような尺度を想定しているのかを明確に意識する必要がある．

4．エッジコンピューティングの具体例

エッジコンピューティングの標準化事例は複数あるが，本稿では標準化が良く進んでおり実際に触れる機会も期待できる事例としてMEC（Multi-access Edge Computingもしくは Mobile Edge Computing＊2）を例として取り上げる．

4.1　MEC（Multi-access Edge Computing）

MECは携帯電話網である5Gネットワークを想定したエッジコンピューティングの仕組みである．ETSI（欧州電気通信標準化機構）により標準化が行われている＊3．

MECは5Gの基地局でユーザの通信を中継するだけではなく，クラウドで提供されているようなアプリケーション実行環境を提供するための一連のフレームワーク，アーキテクチャ，APIである[6]．MECにおけるアプリケーションの配置は図3のようになる．わかりやすいアプリケーションとしては，動画配信サービス，地図サービス，見守りサービスなどが考えられるだろう．従来すべての通信がクラウドに向かっていたところの一部をMECアプリケーションで終端する．これにより，レスポンスの向上，クラウド側トラフィックの最適化，プライバシーを含むデータの局在化などが実現できる．

基地局が全国に多数あることを考えると，地理的な条件に合わせてサービスを最適化できる可能性もある．先に挙げた地図サービスの例でいうと，特定の街に設置された基地局配下のユーザはその街の地図を参照する可能性が高い．このことに注目して，基地局配下のユーザに限定してクラウドよりも豊富なスポット情報を提供するモデルが考えられる．これは自動運転の分野で期待されている．自動運転では，車に搭載したセンサにより鮮度の高い高精度地図を生成するダイナミックマップ技術が重要であるとされる．ダイナミックマップをMECで提供する場合，現在位

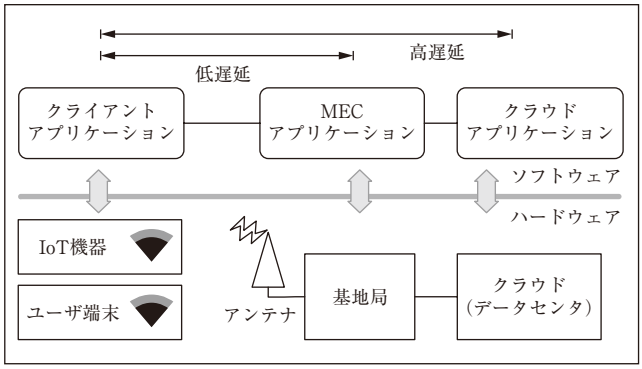


図3　MECアーキテクチャの概略

＊2　標準化の初期段階ではMobile Edge Computingと呼称されていた背景があり，いくつかの文書ではこの用語が使われている．新しい文書についてはMulti-access Edge Computingに統一されている．

＊3　5Gネットワークの仕様は3GPP（Third Generation Partnership Project）により標準化されているため，厳密には互いに独立した標準であるが，現時点ではMECと5Gは強く関連している．

置周辺についてはMECに集積された鮮度・解像度の高い地図情報が高速に提供され，遠くの街の情報についてはクラウドから解像度を落とした大まかな情報が提供されるというモデルが考えられる．MECで収集したダイナミックマップのデータは，データの解像度を落として軽量化した上でインターネットに集積するようにすればシステム全体の効率が向上するだろう．

ダイナミックマップの例は計算量の最適化という側面から捉えることもできる．車の量は膨大であり，各車によって生成されるデータも膨大である．あらゆる処理をクラウドでこなすこともちろん可能であるが，MECを利用する場合，一部のフィルタリング処理をMECに任せることでクラウドのデータ量・計算量を減らすという設計もありうる．データセンタの強力な設備を背景に持つクラウドの計算能力をMECで補うとしても効果は限定的だが，副次的な恩恵と考えれば有益である．

4.2 MECのアーキテクチャ

MECはハードウェアの仮想化技術*4を前提とし，仮想化された環境に任意のアプリケーションを分散配置するアーキテクチャとなっている[6]．図4に仮想化された環境とMECアプリケーションの関係を示す．MECは特定の仮想化技術やコンテナ技術に依存することはなく，利用する仮想化技術は原則として自由である．仮想環境に対してアプリケーションの動作に必要な要件を提示するためのデータ形式が標準化されており，どのような仮想化技術が利用されていたとしても動作条件をコントロールできるよう配慮されている．このデータ形式をAppD（Application descriptor）と呼ぶ[7][8]．

AppDには特定のハードウェア名や仮想化技術の名前が書かれるようなことはない．どのような計算リソース（CPUやメモリー）が必要なのか（virtualComputeDescriptor），どのようなストレージが必要なのか（virtualStorageDescriptor），どのような通信制御が必要なのか（appTrafficRule），通信遅延はどこまで許容できるのか（appLatency）などの抽象的な要素を指定する．MECシステムはAppDに書かれた要件を解釈し，要件を満たすようにアプリケーションを基地局に配置する．

アプリケーションの配置先となる基地局側の機能・性能は仮想化されたリソースとして準備されている．基地局側の仮想化アーキテクチャはNFV（Network Function Virtualization）としてやはりETSIで標準化されている．大まかには，NFVの枠組みで仮想化された基地局側の提供リソースと，MECの枠組みで仮想化された要求リソースのマッチングを図ることでアプリケーションの配置先が決定できる仕組みである．

提供リソースと要求リソースを比較検討して最適なアプリケーションの実行環境を選択する機構はエッジコンピューティングにおいて重要である．もちろん，人間の判

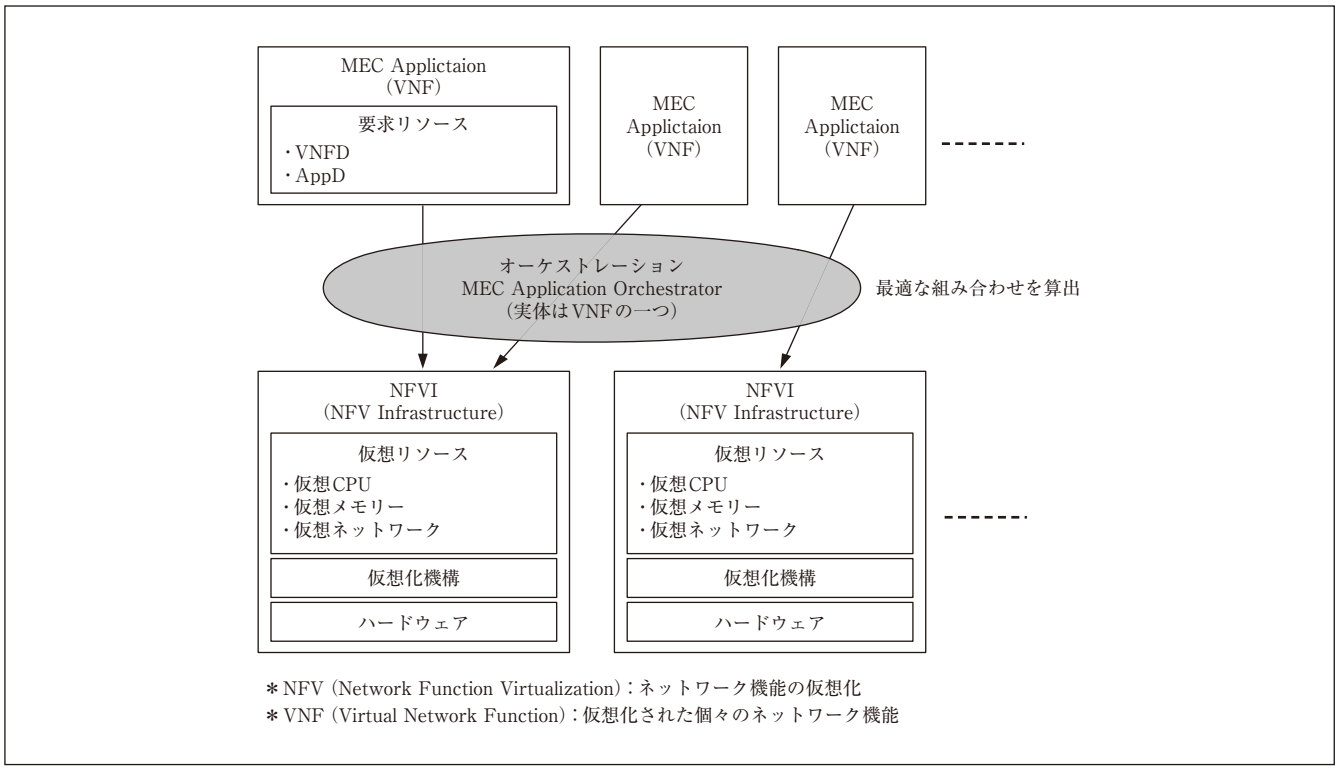


図4　MECアプリケーションの配置

＊4　身近な技術としては，VMwareによる仮想マシンやDockerによるアプリケーションコンテナがある．

断により，ユーザの近隣に独立した機材を配置して個別に環境を構築することも可能ではある．広義ではこうした手法もエッジコンピューティングと考えることもできる．しかし，ユーザの近隣に独立したコンピューティング環境を構築する手法は従来からオンプレミスと呼ばれており，エッジコンピューティングとは異なるアーキテクチャであると考えた方が理解しやすい．

5．オーケストレーション

MECの例でみたように，エッジコンピューティングにおいては要求リソースと提供リソースを付き合わせて最適なアプリケーションの配置を導き出す必要がある．この仕組みを一般にオーケストレーションと呼んでいる．オーケストレーションはクラウドコンピューティングのアーキテクチャにも存在する要素である．エッジコンピューティングのオーケストレーションを考える場合，クラウドコンピューティング向けのデータセンタも考慮に入れる（もしくはクラウドコンピューティング側がエッジコンピューティングを考慮に入れる）必要がある．

エッジコンピューティングの提供リソースはクラウドコンピューティングと比較して限られているため，要求が緩やかなアプリケーションについては貴重なエッジコンピューティング設備ではなくクラウド設備を優先的に使った方が良い．クラウド事業者の提供サービスはリソースの増強を簡単に実行できるよう設計されており，何かのイベントによる負荷上昇が予想される場合などには一時的にCPUパワーを増やしたりメモリーを増やしたりというオペレーションも比較的容易である．一方，エッジコンピューティングの場合，ユーザ近傍の計算資源の一つ一つに関してはクラウドほどの柔軟性は実現しにくい．クラウドのように予備リソースを配置しても有効活用できるとは限らないし，物理的に予備リソースを置くだけの場所がない可能性もある．例としてあげたMECについても，MEC環境でのスケールアウトはスコープ外としている[9]．

エッジコンピューティングには要求の厳しい処理だけを担当させ，他の処理はクラウドコンピューティングが担当すべきである．できることであれば，この処理の分割も自動で解決してほしいと考えると，エッジコンピューティングのオーケストレーションとクラウドコンピューティングのオーケストレーションは密に連携する必要がある．クラウドコンピューティングとエッジコンピューティングを同時にオーケストレーションの対象にしようと考えた場合，設備の提供資源の不均質さが課題となる．データセンタの場合，ある程度特性の揃った均質な計算機群を用意できる．対してエッジコンピューティングは特性が異なるエッジを使い分ける必要がある．この違いに対処する必要がある．

例として，クラウドコンピューティング向けのソフトウェアスタックを提供するOpenStackの場合，以下の点がエッジコンピューティングに対応するために不可欠な要素であると分析している[10]．

(1) 潜在的かつ広範に分散された複数のサイトの動的プールをサポートする能力
(2) 潜在的に信頼できないネットワークコネクション
(3) ネットワーク間をまたぐサイトにおいてリソース制限が解決困難となる可能性

上記のような課題が解決できれば，エッジコンピューティングとクラウドコンピューティングを区別する必要もなくなる．ユーザの立場ではエッジコンピューティングという言葉を特別に意識することもなくなるかもしれない．

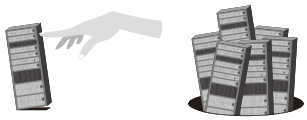
6．要求リソースを伝える難しさ

データセンタがエッジコンピューティングを意識した仕組みに変化することで，ユーザはクラウドコンピューティングと同じ感覚でエッジコンピューティングの利益を享受できるのではないかと考えられる．既存のクラウドコンピューティングとの違いは，必要であれば要求リソースをより厳格に提示できる点にある．今まで通りの緩やかな要求リソースを提示していれば，今まで通りのクラウドコンピューティングである．追加の要求を出したとき，はじめてエッジコンピューティングの利用が検討されることになる．

ここで，要求リソースの提示がユーザにとって困難な作業となる可能性がある．必要・不要の2択で表現できるような要求であれば比較的容易に検討できる．しかし，処理速度のような定量的な指標については専門知識がないと判断が難しい．外部の専門家に協力を求めるとしても，意思伝達がうまくできない可能性がある．定量的な見積もりの曖昧さは時として大きな事故に繋がる．特にエッジコンピューティングとIoTが組み合わさった場合，情報に関する事故だけではなく，ケガなどの人的被害が生じることをも想定する必要がある．

最後にエッジコンピューティングそのものからは少し外れることとなるが，ユーザの助けとなる資料として非機能要求グレードを紹介したい．機能の有無だけでは表現できないシステムに対する要求を非機能要求と呼ぶ．非機能要求は本稿で扱ってきたエッジコンピューティングの要求リソースを包含する概念である．この非機能要求に関する意思伝達の難しさを軽減する取り組みとして，IPA（情報処理推進機構）が非機能要求グレードを規定している[11]．非機能要求グレードの骨子は専門家でなくとも理解できる平易な表現で非機能要求を説明することである．一般的に必要と考えられる要求事項が典型的なモデルの例とともに列挙してあるため，検討漏れを防ぐ効果もある．

例として，システムの可用性はデータセンタ設備で提供するクラウドコンピューティングとエッジコンピューティングでは異なってくると考えられる．非機能要求グレードを参照すると，可用性を説明するための項目がいくつか挙



げられている．一つを取り出してみると，「システムの稼働時間や停止運用に関する情報」とある．内容は，

［レベル0］規定なし
［レベル1］定時内（9:00～17:00）
［レベル2］夜間のみ停止（9:00～21:00）
［レベル3］1時間程度の停止あり（9:00～翌8:00内）
［レベル4］若干の停止あり（9:00～翌8:55）
［レベル5］24時間無停止

という大まかなレベルで示されている．充分に平易な表現である．こうした平易な表現をもとに専門家との意思伝達を図ればエッジコンピューティングが必要なのか，クラウドコンピューティングが必要なのかをより的確に判断できるだろう．この非機能要求グレードそのものはエッジコンピューティングやクラウドコンピューティングを意識したものではなく，最適なツールとは言えない．しかし，将来的にはエッジコンピューティング周辺においても平易な語彙も同様に蓄積され，専門的な知識は必ずしも必要なくなっていくと考えられる．

7．むすび

2010年代から注目されはじめたエッジコンピューティングについて，背景，具体例を挙げて概観してきた．

エッジコンピューティングに関して常に問われるのは，「クラウドではだめなのか？」である．単純に理屈上の効率性からすれば，より多くの計算機をよりユーザに近い場所に分散配置できるエッジコンピューティングが高効率になりうるのは明白である．しかし，エッジコンピューティング環境におけるオーケストレーションは複雑になりがちで，その複雑さのもたらすコストがメリットに見合うのかどうかは決して明白ではない．

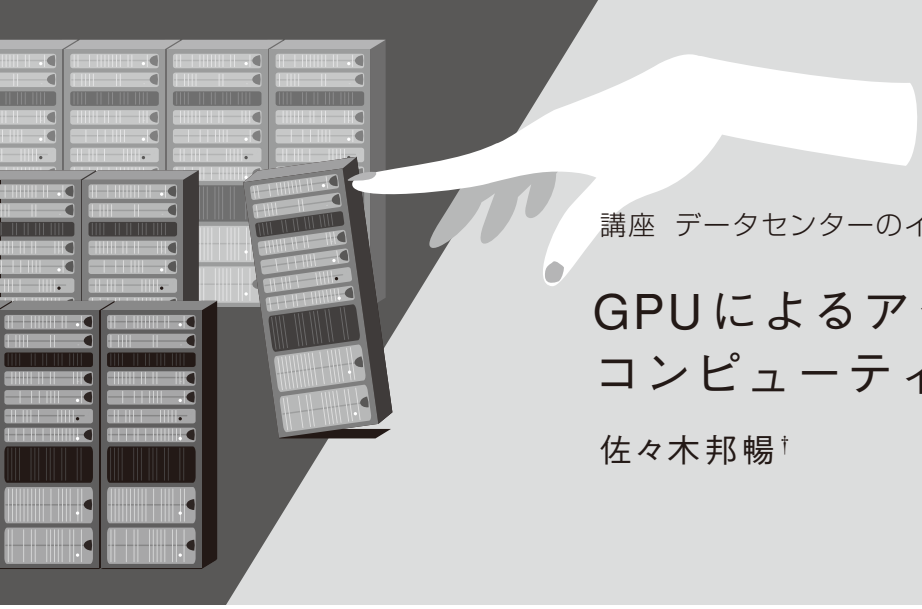
本稿はクラウドではだめなのか？ という問いを常に頭におきながら，クラウドとエッジの違いがどこにあるのかを解説してきた．最後に，やや脱線気味ではあるものの，非機能要求について紙面を割いた．今後エッジコンピューティングに出会った際には，それで何ができるのか？ ではなく，どれだけ上手にできるのか？ という観点から既存のシステムとの違いを考えていただければ幸いである．

（2021年6月28日受付）

〔文　献〕

1）総務省：“令和二年版情報通信白書”，総務省，p.76（2020）
2）A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, M. Ayyash: "Internet of Things: A survey on enabling technologies, protocols and applications", IEEE, IEEE Communications Surveys & Tutorials, 17, 4, pp.2364-2366 (2015)
3）F. Bonomi, R. Milito, J. Zhu, S. Addepalli: "Fog computing and its role in the internet of things", ACM SIGCOMM, MCC '12: Proceedings of the first edition of the MCC workshop on Mobile cloud computing (2012)
4）S. Parikh, D. Dave, R. Patel, N. Doshi: "Security and Privacy Issues in Cloud, Fog and Edge Computing", Elsevier BV, Procedia Computer Science, 160 (2019)
5）IIC: "The Industrial Internet of Things Distributed Computing in the Edge", Industrial Internet Consortium (IIC), p. 6 (2020)
6）ETSI: "Multi-access Edge Computing (MEC); Framework and Reference Architecture", ETSI (2019)
7）ETSI: "Multi-access Edge Computing (MEC); MEC Management; Part 2: Application lifecycle, rules and requirements management", ETSI, p.29 (2019)
8）ETSI: "Network Functions Virtualization (NFV) Release 3; Management and Orchestration; VNF Descriptor and Packaging Specification", ETSI, p.24 (2019)
9）ETSI: "Mobile Edge Computing (MEC); Deloyment of Mobile Edge Computing in an NFV environment", ETSI, p.25 (2018)
10）OpenStack Foundation: "Cloud Edge Computing: Beyond the Data Center", OpenStack Foundation, p.3 (2018)
11）IPA：“非機能要求グレード2018システム基盤の非機能要求に関するグレード表”，IPA（2018）

末永（すえなが）　洋樹（ひろき）　2000年，日本大学理工学部電気工学科卒業．2003年，北陸科学技術大学院大学中退．同年より，（株）インターネットイニシアティブにて小規模オフィス向けのルータ開発に従事する．2016年より，フォグコンピューティングおよびエッジコンピューティングの研究開発に従事．



講座　データセンターのインフラ技術〔第５回〕

GPUによるアクセラレーテッドコンピューティング

佐々木邦暢†

1．まえがき

アクセラレーテッドコンピューティングとは，コンピュータの主たる演算装置であるCPU（Central Processing Unit）に加えて，GPU（Graphics Processing Unit）やFPGA（Field Programmable Gate Array）等のデバイスを利用してCPUの処理を肩代わりさせ，システムを高速化することを指す．その用途として，以下のようなものが挙げられる

(1) HPC（High Performance Computing）やAI（Artificial Intelligence）・機械学習処理の高速化
(2) ストレージの圧縮・重複排除
(3) ネットワークの暗号化・複合化

本稿では，この中でもHPCや機械学習処理の高速化において重要な要素となったGPUによるアクセラレーテッドコンピューティング（以下，GPUコンピューティング）の状況と，それを支える技術について歴史を簡単に振り返り，最新の状況をお伝えする．

2．GPUの特性，CPUとの違い

2.1　GPUとは

コンピュータの性能向上に伴って高精細な3Dグラフィックスを高速に描画することが可能となり，ビデオゲームや映画のようなエンタテインメントから，自動運転車の仮想空間上での走行試験といった産業用途まで，さまざまな用途で活用が進んでいる．

このような3Dグラフィックスでは，画面に表示される人や建物，自動車などのオブジェクトを3次元のモデルで表現する．モデルは小さな平面のパネルを組み合わせて構築され，その頂点座標群に対して，移動や回転などの座標変換（Transpose）や，光源に応じた陰影計算（Lighting）が適用されることで，2次元のディスプレイに表示される画像ができあがる．

高精細な画像を得るためには多数のパネルからなる精緻なモデルが必要で，それを構成する頂点の数は場合によっては数千万以上となる．また，陰影処理では，画面のすべてのピクセルについて表示する色を計算する必要がある．ピクセル数はFull-HD解像度なら200万，4Kなら800万ピクセルに及ぶ．

数千万の頂点座標に対する座標変換と，数百万のピクセル毎の陰影計算は，非常に負荷の高い演算である上に，ユーザの操作などに応じてリアルタイムに表示内容を変化させる必要のある3Dグラフィックスの表示においては，これらの演算にかけられる時間は限られている．動きが滑らかに見えるためには，少なくとも毎秒30回，VR（Virtual Reality）コンテンツではいわゆる「VR酔い」を避けるために最低でも60回程度の画面書き換えが要求されるためだ．

したがって，3Dグラフィックスの処理には高い演算能力が必要となるが，これをCPUのみで行うことは難しい．そのために，グラフィックス関連の演算をCPUに代わって処理するGPUが発展してきた．

2.2　CPUとGPU

CPUではさまざまな種類のアプリケーションが実行されるため，特定の処理だけに特化するわけにはいかない．しかし，GPUはグラフィックス処理に特化することができるために，CPUのような汎用性を追求する必要はない．そのためCPUとGPUは大きく異なる特性を持つ．

3Dグラフィックスの処理に必要な前述の座標変換処理は頂点座標毎に独立しており，また陰影計算は画面のピクセル毎に独立している．したがって，これらは並列に処理することができる．そのため，GPUは並列演算性能を高めることに注力している．CPUはシングルスレッド性能を高めるためのアウトオブオーダ実行や分岐予測といった機能を備えるが，GPUはこれらを持たず，その分，多くの演算器を搭載することにリソースを割き，高い演算性能を獲得している．

例として，2020年にリリースされたデータセンタ向けCPU（Intel Xeon Platinum 8380H）とGPU（NVIDIA A100）の倍精度（64-bit）演算性能を比較すると，CPUの1.79 TFLOPSに対し，GPUは19.5 TFLOPSであり，10倍強の差がある．

GPUがこの高い演算性能を効果的に発揮するのは，同じ

†エヌビディア合同会社
"Infrastructure Technology for Datacenters (5): GPU Accelerated Computing" by Kuninobu Sasaki (NVIDIA Japan, Tokyo)

命令を多数のスレッドで並列実行するタイプの処理で，同時に実行可能なスレッド数は，CPUの場合は数十だが，GPUではこれが数千レベルとなる．

また，多数の演算器を生かすためには大量のデータを供給し続ける必要があり，そのために高帯域幅のメモリーを搭載するのもGPUの特徴である．最新のNVIDIA A100 GPUはHBM2eメモリーにより2TB/sを超える帯域幅を持つ．

3．GPUコンピューティングの歴史

前述の通り，GPUはグラフィックス描画のために並列演算性能を進化させてきた．ここで少しその歴史を振り返りながら，画面描画に特化したデバイスが汎用的なアクセラレータに進化した過程を見てみたい．

3.1　ハードウェアT&L（Transpose and Lighting）の実現

GPUの歴史は古く，1970年代のグラフィックスコントローラまで遡ることができるが，「GPU」という名称が初めて使われたのは，1999年に発売されたエヌビディアのGeForce 256である．これは，座標変換（Transpose）および陰影計算（Lighting）処理とレンダリングの機能を1チップに収めた初めての製品で，NVIDIAはこれをGPU（Graphics Processing Unit）と命名した．現在まで続くGeForce製品はここから始まる．

3.2　プログラマブルGPU

2001年に登場したGeForce 3は，初の「プログラマブルGPU」であった．これは頂点座標の変換を行うバーテックスシェーダや，ピクセル単位での陰影計算を担うピクセルシェーダの機能を自由にプログラムできる「プログラマブルシェーダ」を搭載したもので，汎用的なGPUコンピューティングに繋がる重要な進化であった．Microsoftの初代Xboxには，このGeForce 3のカスタム版が搭載されていた．

プログラマブルシェーダによってGPUの処理に自由度が生まれると，この非常に高い並列演算性能を，グラフィックス以外の用途にも活用できないかという考えが当然出てくる．その初期には，シェーダ記述言語のCgやGLSL（OpenGL Shading Language），HLSL（High Level Shading Language）を利用した演算も行われた．計算したいデータをテクスチャに見立ててGPUへ転送し，ピクセルシェーダで計算するといった手法である．しかし，シェーダ記述言語はあくまでグラフィックス処理を目的としており，汎用計算目的に使いやすいものではなかった．

3.3　ユニファイドシェーダとCUDAの登場

2006年，エヌビディアは新たな世代のGPUチップ「G80」とそれを搭載した製品であるGeForce 8800を発表した（図1）．G80ではバーテックスシェーダとピクセルシェーダをまとめた「統合シェーダ」アーキテクチャが採用されており，処理の自由度がさらに高まった．

また，G80と同時に発表された，GPUコンピューティングのためのプログラミングモデル「CUDA（Compute Unified Device Architecture：クーダ）」により，開発者が慣れ親しんだC言語に近い形で，GPU上での汎用計算を記述できるようになった．これにより，グラフィックス処理の枠を超えた，本格的なGPUコンピューティングの時代が始まり，翌2007年には，G80を搭載したデータセンタ向け製品「Tesla」も登場した．

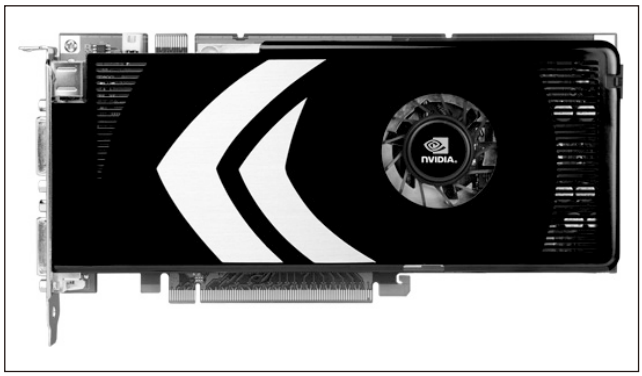


図1　GeForce 8800

また，2008年にはGPUに限らずさまざまなデバイスをサポートすることを目指した，クロスプラットフォームの並列コンピューティングAPIであるOpenCL（Open Compute Language）の最初の仕様も発表され，AMDやエヌビディアもこれに対応している．

4．スーパーコンピュータとGPU

GPUの高い並列演算性能は，高性能計算（HPC）の世界でその実力を示すことになる．ここでは，半年ごと（6月と11月）に更新されるスーパーコンピュータの世界ランキングTOP500およびGreen500に着目して，GPUコンピューティングの進化を見てみたい．

GPUで高速化されたスーパーコンピュータが初めてTOP500リストに登場したのは，2008年11月のことである．東京工業大学のTSUBAMEがTesla S1070を170台搭載し，29位にランクインした．Tesla S1070は，G80の進化型であるGT200チップを搭載した1Uラックマウント型のGPUユニットであった．

2009年，エヌビディアは「Fermi」と呼ぶ新たなGPUアーキテクチャを発表する．Fermi世代のTeslaは，HPCで重要な倍精度浮動小数点演算の性能が大幅に強化された．その結果，2010年11月のTOP500リストでは，Tesla M2050を搭載した中国のTianhe-1 AがGPUスパコンとして初めて首位を獲得．また，東工大のTSUBAME 2.0はノード毎に3基，全体で4,224基のTesla M2050を搭載し，TOP500リストの4位という世界トップクラスのスーパーコンピュータとなった．

その後，2012年11月にはFermiの後を継いだKeplerアーキテクチャによるTesla K20X GPUを18,688基搭載した米国エネルギー省オークリッジ国立研究所（ORNL）の

図2　Summit

Titanが首位を獲得．Titanを含め50システムがNVIDIA GPUを搭載しており，全体の10%がGPUスパコンとなった（余談だが，TITAN VやTITAN RTXといったGPU製品の名称は，このTitanスーパーコンピュータに由来する）．

2018年6月には27,648基のTesla V100 GPUを搭載したORNLのSummit（図2）が1位に，これによってアメリカは4年半年ぶりにTOP500の首位を中国から奪還し，Summitは2020年6月に日本の「富岳」が登場するまで，その座を維持した．

本稿執筆時点における最新版である2021年6月のリストでは，TOP500[1]の全500システム中140システム，上位10システムに絞ると過半数の6システムがGPUを搭載している．また，電力性能比（消費電力1Wあたりの演算性能）のランキングであるGreen500[2]では，上位10システム中，実に9システムがGPUを搭載している．なお，GPUを搭載していない1システムは日本のPreferred Networksによる MN-3で，これは独自開発のMN-Coreというアクセラレータを搭載してGreen500の首位を獲得した．

CPUのシングルスレッド性能の向上が伸び悩む現在，アクセラレーテッドコンピューティングは高い演算性能・電力性能比を実現するために欠かせない技術となっている．

5．AI・ディープラーニングへの活用

5.1　GPUとディープラーニング

前章で紹介したスーパーコンピュータの領域に限らず，現在GPUコンピューティングは大きな広がりを見せているが，その原動力となっているのは，やはりディープラーニング（深層学習）での活用である．

GPUを活用したディープラーニングが実績を出し始めたのは10年ほど前になる．

例えば，2011年のSeide[3]では，GPUを使ったディープラーニングで音声のテキスト変換（speech-to-text）の精度を高めたことが報告されている．

また，ディープラーニングを一躍有名にした出来事の一つに，2012年のImageNet Large Scale Visual Recognition Challenge（ILSVRC）でトロント大学のSupervisionチームが圧勝した事例があるが，彼らの作成したAlexNetというモデルは，文献[4]に次のようにあるとおり，GeForce GTX 580を2基使用してトレーニングされた．

“*Our network takes between five and six days to train on two GTX 580 3 GB GPUs.*”

では，なぜGPUはディープラーニングの高速化に有用なのだろうか？

画像分類用の畳み込みニューラルネットワークを例に取ると，典型的には前段の畳み込み層で画像の特徴を抽出し，後段の全結合層でそれを分類する．前述のAlexNet（図3）は，5段の畳み込み層と，3段の全結合層を持つ．

畳み込み層，全結合層ともに，入力ベクトル（をまとめた行列）に対してパラメータ行列をかける行列積の演算を大量に行うことになる．これは，頂点座標に変換行列をかけるといったGPU本来の処理に類似した，GPUが最も得意とするタイプの演算であり，GPUの高い並列演算性能が遺憾なく発揮されることとなった．

5.2　複雑化するディープラーニングモデル

AlexNetは8層のニューラルネットワークであったが，

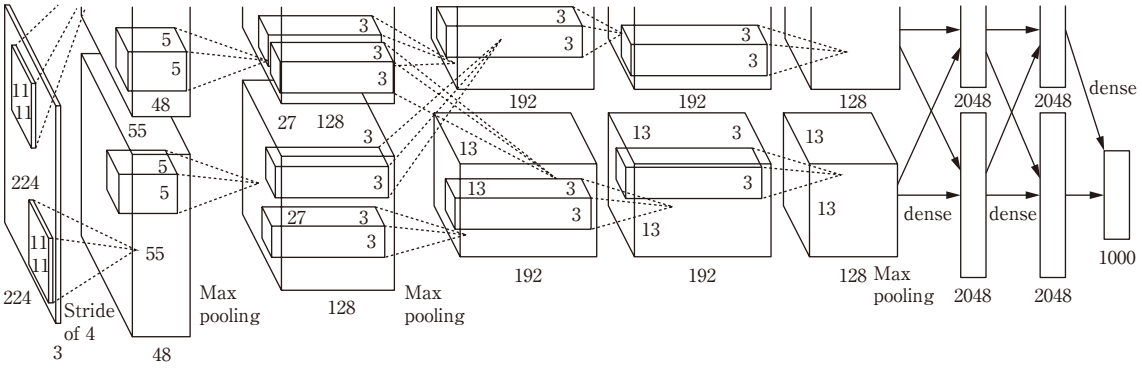


図3　AlexNetの層構成

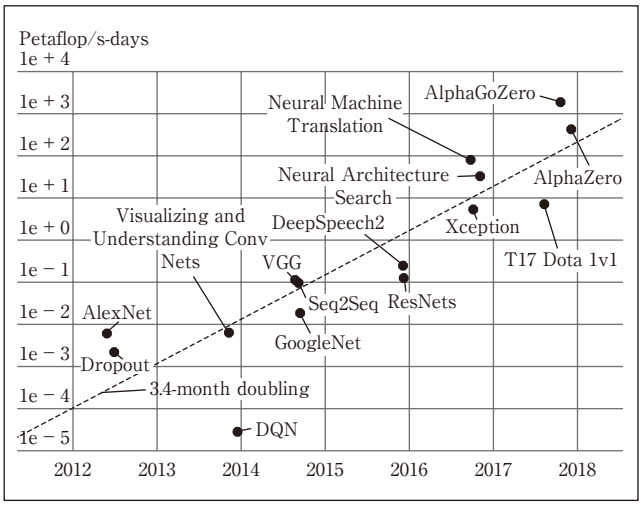


図4　複雑化するディープラーニングモデル

その後のディープラーニングの研究進展によって，モデルは急速に大規模化する．一例として，AlexNetの3年後，2015年のILSVRCで首位に輝いたMicrosoft ResearchのResNet[5]は152層の複雑な構造を持っていた．

ここ数年は特に，自然言語理解（NLU）関連でモデルの複雑化が著しい．モデルのパラメータ数（トレーニング時に調節する必要のある重みの数）に着目すると，2012年のAlexNetが6,000万パラメータだったのに対し，2018年に登場し，多くのNLUタスクでそれまでの最高スコアを更新したBERT[6]のLARGE版は3.4億，2019年に公開され，生成される文章の品質が高いことで悪用も懸念されると話題になったGPT-2[7]が15億，GPT-2をさらに発展させ2020年に公開されたGPT-3[8]に至っては1,750億のパラメータを持つ．

また，OpenAIのブログ記事“AI and Compute”[9]によれば，2012年のAlexNetから2018年のAlphaGo Zeroまでの間に，計算負荷は30万倍に増加している（図4）．

5.3　大規模ディープラーニングインフラと分散学習

このような流れの中で生まれてきたのが，複数のGPU，複数の計算機を束ねたクラスタ環境でディープラーニングモデルをトレーニングする分散学習という手法である．「機械学習のHPC化」とも言えよう．ここで，いくつか例を示す．

5.3.1　ResNet-50チャレンジ

2017年6月にFacebookの研究チームがTesla P100 GPUを256基（ノードあたり8基の32ノードクラスタ）使い，ResNet-50のトレーニングを1時間で完了させたのは一つの大きな出来事であった．この「ResNet-50をILSVRC2012データセットで90エポック，精度75%」が一つの基準となり，以後次々に記録が塗り替えられていくことになる．

2017年の11月にはPreferred NetworksがTesla P100を1,024基搭載する自社のスーパーコンピュータ“MN-1”で同じトレーニングを15分で完了させた．MN-1はノードあたり8基のTesla P100を搭載し，ノード間をFDR InfiniBandで接続した構成で，2017年11月のTOP500リストの91位にランクインしている．

この「ResNet-50チャレンジ」はその後もTencent, SONY, Google等が記録を更新し，2019年3月には富士通が産総研のスーパーコンピュータ“AI Bridging Cloud Infrastructure（ABCI）”で1.2分の記録を樹立している．この時のABCIは，1,088ノードで合計4,352基のTesla V100 GPUを備えていた．現在は，最新のNVIDIA A100 GPUを搭載する「計算ノード（A）」も加わって“ABCI 2.0”にアップグレードされ，依然として日本最大のGPUスーパーコンピュータである．

5.3.2　ゴードン・ベル賞

ゴードン・ベル賞は，HPC分野で顕著な功績を挙げた研究に対し，アメリカ計算機学会（ACM: Association for Computing Machinery）から授与される賞である．ここ数年は，ディープラーニングを大規模に活用した研究が受賞するケースも目立ってきた．

2018年には当時TOP500ランク1位だったGPUスーパーコンピュータである米国のSummitと，同じく5位であったスイスのPiz Daintを使い，気象画像の分析にディープラーニング（画像のセグメンテーション）を適用した研究[10]が受賞している．

昨年（2020年）は．米国とアメリカの研究者を中心とした研究[11]が受賞した．これは，分子動力学の計算にディープラーニングを活用する「深層ポテンシャル分子動力学（DPMD：Deep Potential Molecular Dynamics）」の例で，前述のSummitスーパーコンピュータの4,560ノードを使って大規模実行することで，第一原理分子動力学法と同等の精度で，1億原子規模の計算を可能にしたものである．

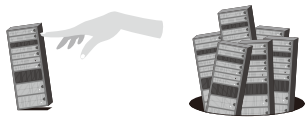
5.3.3　エヌビディア社内のGPUクラスタ

大規模なディープラーニング向けインフラの例として，エヌビディア社内の例を紹介したい．

エヌビディアは2016年，新たなPascalアーキテクチャに基づくTesla P100 GPUと，それを8基搭載した「DGX-1」というサーバ機を発表した．DGX-1は社内にも多数導入され，2016年11月のTOP500リストにはDGX-1の125ノードクラスタであるDGX SATURNV[12]がランクインした（「サターンファイブ」という名称は，アポロ計画で人類を月へ運んだサターン5型ロケットから）．

その後もSATURNVは成長を続け，2018年に登場した，Tesla V100 GPUを16基搭載するDGX-2, 2020年には，NVIDIA A100 GPUを8基搭載するDGX A100といったDGXシリーズの後継機が追加されることで，現在は総計2,000ノードを超える大規模GPUサーバ群となっている．

SATURNVは社内のさまざまなチームが製品開発に利用しているほか，その運用管理から得られたノウハウを元に



DGX SuperPODというリファレンスアーキテクチャを公開．これは，20ノードの「スケーラブルユニット（SU）」を単位としてモジュール方式でクラスタを拡張可能なアーキテクチャで，ストレージやネットワーク装置の構成も定義されている．顧客はこれに基づきDGXの性能を引き出すクラスタを自社に構築可能である．

5.4 混合精度演算

5.4.1 ディープラーニングと数値の精度

前述の分散学習は，複数のGPUを使うことで大規模なモデルを高速にトレーニングすることを可能にする手法であった．これとは別に，近年注目されているのが，演算に使用する数値の精度（ビット数）をディープラーニング向けに最適化する，「混合精度演算」である．

GPUの「本来の」用途であるグラフィック向けの演算では，座標などの数値は32ビット幅の浮動小数点数で表現される．これを単精度浮動小数点数，あるいは単にFP32と呼ぶ．

一方，HPC向けにはより広い範囲を細かく表現できる，64ビット幅の倍精度浮動小数点数（FP64）が必要となる．エヌビディアがFermi世代のGPUでこの倍精度演算を強化したことでGPUコンピューティングが本格化したのはすでに述べたとおりである．

しかしながら，ニューラルネットワークの重みパラメータ等は，倍精度を必要としないため，AI・機械学習領域でのGPU利用はFP32を中心に始まり，近年はさらに小さな数値型で演算を行うようになっている．

InterSpeech 2014でのMicrosoft Researchによる発表[14]では，データ並列の分散SGD（Stochastic Gradient Descent：確率的勾配降下法）においてGPU間でやりとりされる勾配を1ビットで表現する方式を提案し，これはMicrosoftのディープラーニングフレームワークであるMicrosoft Cognitive Toolkitに1-bit SGDとして実装された．2015年のIBMによる発表[15]では，半精度の固定小数点数での演算で精度の低下なくトレーニングを高速化できることが示され，2017年の大山らによる論文[16]では，fp8でのトレーニングについて述べられている．

5.4.2 ハードウェア側の対応

このように，ディープラーニングの演算には，従来からのFP32と，よりビット数の少ない数値型を合わせて使う混合精度演算[17]が一般化しつつあり，GPU等のハードウェア側でもその対応が進んでいる．

Googleが2017年に発表した第2世代のTPU（Cloud Tensor Processing Unit）[18]は，独自の16ビット浮動小数点数であるbfloat16に対応した．これは，後にAMD, Arm, Intel等のプロセッサでもサポートされることとなった．

同じく2017年にエヌビディアが発表したVoltaアーキテクチャGPUでは，ディープラーニングの学習および推論フェーズにフォーカスした新たな演算器「Tensorコア」が導入された．TensorコアはFP16を入力とし，4×4行列の積和演算に相当する128演算を1サイクルで実行する能力を持つ．これによりTensorコアを搭載するNVIDIA V100 GPUのFP16理論演算性能は125 TFLOPSに達し，前世代のTeslaP100のFP16性能21.2 TFLOPSの6倍弱に向上した．

2020年に登場した最新のNVIDIA Ampereアーキテクチャでは，Tensorコアがさらに強化され，TF32（TensorFloat 32）という新しい数値型をサポートした．これは，FP32の仮数部を10ビットに短縮した19ビットの内部表現を使い，Tensorコアによる高速な演算を可能にするという機能である．TF32はTensorコア内部でのみ利用され，入力と出力は一般的なFP32であるため，Tensorコアに対応していない従来型のプログラムを，修正せずにそのまま高速化できる利点がある．

6．クラウドコンピューティングとGPU

ここまで紹介したとおり，GPUコンピューティングの用途が当初の科学技術計算から，AIの発展を支える機械学習領域へと広がってきたことで，GPUインフラの需要がさらに高まっている．この状況を踏まえ，大規模なデータセンタで大量のサーバを運用することに熟達したクラウドサービスプロバイダ各社から，さまざまなタイプのGPUリソースが提供されている．ここではその一部を紹介する．

6.1 GPU IaaS

各社の最も基本的かつ汎用的なサービスが，GPUを搭載したサーバ（仮想マシンが中心だが，仮想化されていないベアメタルサーバを提供するクラウドもある）を従量課金で利用できるIaaS（Infrastructure as a Service）である．Amazon Web Services（AWS）のEC2，Microsoft AzureのAzure Virtual Machines, Google CloudのCompute Engineといったサービスがこれにあたる．

最初にGPUインスタンスを提供したのはAWSで，2010年にCG1インスタンスをリリースしている．これは，Fermi世代のデータセンタ向けGPUであるTesla M2050を搭載したもので，Tianhe-1AやTSUBAME 2.0といった世界有数のスーパーコンピュータと同じGPUをクラウドで利用することができた．

その後，Kepler世代のTesla K80，Pascal世代のTesla P100，Volta世代のV100，Turing世代のT4，Ampere世代のA100と，GPUの進化とともに，各社のGPUインスタンスもその種類を増している．比較的新しいVolta世代以降のGPUに関して，提供状態を**表1**に示す．

これらクラウド各社のGPUインスタンスを利用して，実際にどの程度の規模のシステムを構築できるのか，具体的な情報は公開されていないが，最新のスーパーコンピュータTOP500リスト（2021年6月）には，Microsoft AzureのND A100 v4インスタンスによって構築されたクラスタが4システム，26～29位にランクインしている[19]．ND A100 v4は，

表1　クラウド各社のGPUインスタンス提供状態

	Volta	Turing	Ampere
	V100	T4	A100
AWS	○	○	○
Microsoft Azure	○	○	○
Google Cloud	○	○	○
Oracle Cloud	○		○
IBM Cloud	○		
Alibaba Cloud	○	○	○
Tencent Cloud	○	○	
Baidu Cloud	○	○	

最新のNVIDIA A100 GPUをノードあたり8基搭載し，ノード間は低遅延，広帯域のInfiniBand HDRで接続されている．公開されているコア数から逆算すると，ノード数は164となり，少なくともこのサイズのクラスタはクラウドで構成することができるとわかる．

なお，IaaSのサービスでは，各社がGPU対応のマシンイメージ（マシンの雛形）を提供していることが多い，例えばAWSのDeep Learning AMIや，Microsoft AzureのData Science Virtual Machine等である．また，エヌビディアも各クラウド向けに，GPUドライバや，CUDA Toolkit, NVIDIA Container Toolkitといったよく利用されるソフトウェアをインストール済みのマシンイメージを提供している[20]．これらのイメージを利用することで，GPUインスタンスのセットアップに煩わされることなく，本質的な作業に素早く取りかかることができる．

6.2　マネージドKubernetes

コンテナベースのワークロードをある程度大規模な（複数台のサーバが稼働する）環境で実行する場合，コンテナオーケストレーションツールであるKubernetes[21]が一般的になってきた．Kubernetesは，それ自体の構築・運用にある程度のノウハウが必要になるが，クラウド各社が提供するマネージドKubernetesサービスを利用することで，Kubernetesクラスタの構築部分をサービスに任せ，利用の敷居を下げることができる．

AWSのEKS, AzureのAKS, Google CloudのGKE, Oracle CloudのOKEといったサービスでは，GPUインスタンスの利用にも対応しており，コンテナベースのGPUワークロードを容易にKubernetes上で利用することが可能になっている．

7．むすび

本稿では，GPUの発展を振り返り，グラフィックスの描画を高速化するデバイスが，HPCやAI分野に欠かすことのできない演算アクセラレータとなった経緯と，スーパーコンピュータやクラウド環境での大規模な利用例等についてお伝えした．

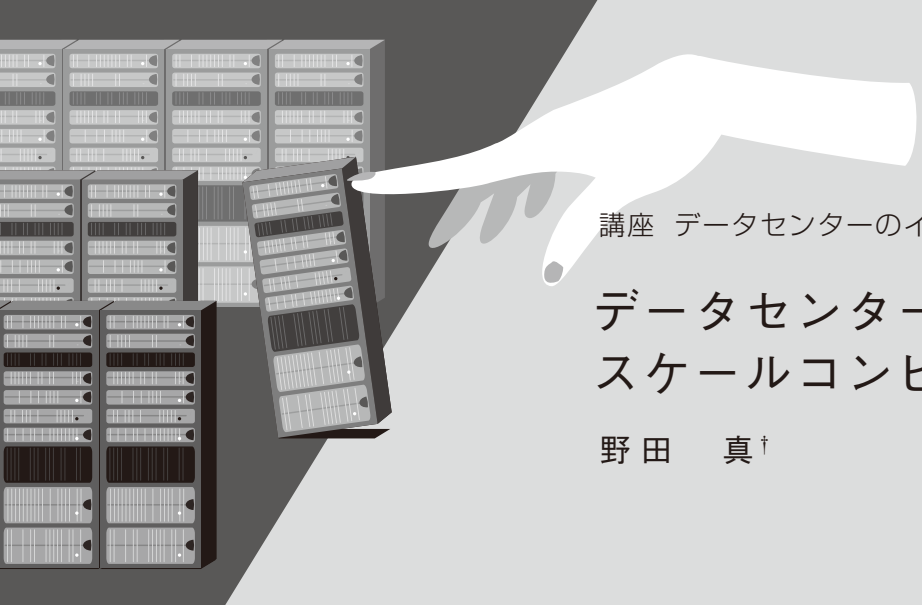
GPUによるアクセラレーテッドコンピューティングについて，ささやかながら有用な情報となっていれば幸いである．

（2021年9月1日受付）

〔文　献〕

1）TOP500, https://top500.org/
2）Green500, https://www.top500.org/lists/green500/
3）Seide, Frank, G. Li and D. Yu: "Conversational speech transcription using context-dependent deep neural networks", Twelfth annual conference of the international speech communication association (2011)
4）Krizhevsky, Alex, I. Sutskever and G.E. Hinton: "Imagenet classification with deep convolutional neural networks", Advances in neural information processing systems 25, pp.1097-1105 (2012)
5）He, Kaiming, et al: "Deep residual learning for image recognition", Proceedings of the IEEE conference on computer vision and pattern recognition (2016)
6）Devlin, Jacob, et al: "Bert: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint arXiv:1810.04805 (2018)
7）GPT-2: 1.5B Release, https://openai.com/blog/gpt-2-1-5b-release/
8）Brown, B. Tom, et al: "Language models are few-shot learners", arXiv preprint arXiv:2005.14165 (2020)
9）"AI and Compute", https://openai.com/blog/ai-and-compute/
10）Kurth, Thorsten, et al: "Exascale deep learning for climate analytics", SC18: International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE (2018)
11）Jia, Weile, et al: "Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning", SC20: International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE (2020)
12）DGX SATURN V, https://www.top500.org/system/178928/
13）BFloat16: the secret to high performance on Cloud TPUs, https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus
14）Seide, Frank, et al: "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns", Fifteenth Annual Conference of the International Speech Communication Association (2014)
15）Gupta, Suyog, et al: "Deep learning with limited numerical precision", International conference on machine learning. PMLR (2015)
16）大山洋介，野村哲弘，佐藤育郎，松岡聡：“ディープラーニングのデータ並列学習における少精度浮動小数点数を用いた通信量の削減”，情処学研究報告，2017-HPC-158，30，pp.1-10（2017）
17）Micikevicius, Paulius, et al: "Mixed precision training", arXiv preprint arXiv:1710.03740 (2017)
18）Cloud Tensor Processing Unit (TPU), https://cloud.google.com/tpu/docs/tpus
19）PIONEER-EUS - NDV4 CLUSTER, https://www.top500.org/system/179969/
20）Using NGC on Public Clouds, https://docs.nvidia.com/ngc/ngc-deploy-public-cloud/index.html
21）Kubernetes, https://kubernetes.io/

佐々木邦暢（ささきくにのぶ）　1998年，東海大学工学部動力機械工学科卒業．日本マイクロソフト（株）でクラウド・HPC関連のコンサルティングを担当中，東工大TSUBAME 2.0構築案件でGPUコンピューティングに出会う．2016年より，エヌビディア（同），データセンター向けGPUサーバやコンテナ技術などインフラ領域のプリセールスエンジニアとして活動中．



講座　データセンターのインフラ技術〔最終回〕

データセンター スケールコンピューティング

野田　真†

1．まえがき

現在，第三次AI（Artificial Intelligence）ブームの真っ只中で，世の中の幅広い産業がその恩恵を受けている．身近なところでは，多くの人々が日々お使いになる検索エンジンや推薦エンジンの最適化，クレジットカード決済時の不正利用の検知，各家庭での生活を便利にしてくれるスマートスピーカをはじめとするスマート家電，さらには医療機関における画像診断のサポートなど，枚挙に暇がない．さらに今後は，自動車の完全自動運転，生産現場の産業用ロボットのリアルタイム制御，日常会話中の即時通訳，難病治療に対するより効率的な創薬など，さらなる発展が期待されている．これらを支えるのはAIの中でも，主に深層学習（ディープラーニング）が中心である．深層学習は，人間の脳の神経系を模したニューロンの層を重ねて，多層で複雑なニューラルネットワークを構成することで，表現力と学習力を高めたモデルを実現できる機械学習の一手法だ．ただ，前述したような多岐にわたる恩恵をもたらす可能性を秘めた深層学習であるが，高度な問題設定に対する，より精度の高い推論を行うべく，日々そのモデルが複雑化・巨大化しているのが現状だ．例えば，最近の文章生成言語モデルとして非常に有名なGPT-3（Generative Pretrained Transformer 3）は，約1,750億個ものパラメータを持ち，一般的なデータセンタークラスサーバ単体の演算能力では，モデルの学習に途方もない時間を要する．この例のように，今後ますます複雑化・巨大化が想定される深層学習を含む機械学習のモデルに対応していくためには，従来からのサーバを一つの単位としてコンピューティングを考える「サーバスケール」のアーキテクチャではなく，データセンターそのものを一つのコンピューティングの単位と捉える「データセンタースケールコンピューティング」が不可欠となる．本講座では，このデータセンタースケールコンピューティングについて，アーキテクチャや構成要素とその詳細，実現例や今後の方向性を解説する．

†エヌビディア
"Infrastructure Technology for Datacenters (final study); Data Center Scale Computing" by Makoto Noda (NVIDIA, Tokyo)

2．データセンタースケールコンピューティングとは

データセンターのアーキテクチャは，初期の物理サーバをベースにした静的なシステム，そして次に仮想化された動的なシステム，現在では，ストレージなどのリソースを，サーバから分離した上，抽象化して利用する動的かつコンポーザブルなシステムへと進化している．また，このようなシステム上で稼働するワークロードのメインストリームも，旧来のウェブやメールなどのシンプルなクライアントサーバ型のワークロードから，現代社会での利活用が進むAIを支えるためのワークロードへ変遷している．この変遷の裏には，1990年台からのインターネットの進展によってアクセスが可能となった，かつてない量と質のビッグデータ，人工知能に携わる研究者達の絶え間ない研鑽によるアルゴリズムやモデルの改良，そしてGPU（Graphics Processing Unit）などのハードウェアの進化に起因する，ムーアの法則を超越した高速演算が可能となったことで，AIが一気に発展したという背景がある．結果として，深層学習を中心とするAIの学習および推論が，昨今のデータセンターワークロードのメインストリームの一部となった．ただし，これらのワークロードは，超膨大なデータ量と複雑で巨大なモデルの上に成り立つものであり（図1），デー

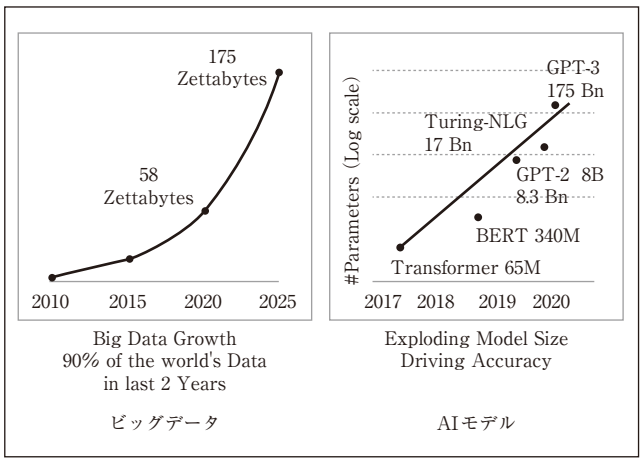


図1　爆発するデータ量と巨大化するモデル[1)]

映像情報メディア学会誌 Vol. 76, No. 1, pp. 70～77（2022）

タセンター内の個々のサーバで処理し，現実的な処理時間内で演算を完了させることは困難である．そのため，同ワークロードを実行するソフトウェアプログラムも，それぞれのサーバ単体で実行される形態から，クラウドネイティブにコンテナ化され，個々の小さな疎結合サービスを軽量プロトコルで集約する形で実現するマイクロサービスアーキテクチャで実装された，データセンター全体の資源を駆使して実行される，分散コンピューティング形態のソフトウェアプログラムへと進展した．

このような状況下，データセンターには，データセンター全体にわたる計算資源，具体的にはCPU（Central Processing Unit），GPU，そしてストレージなどのリソースを適材適所に組み合わせて，ソフトウェアプログラムが要求するタスク処理に最適なコンピューティングプラットフォームを提供することが求められる．このように，これまでのサーバ単位のコンピューティング，あるいは限定されたクラスター内での複数サーバにわたる分散コンピューティングをさらに推し進め，データセンターそのものを一つのコンピューティング単位とする新たな潮流こそが，データセンタースケールコンピューティングである．なお，このデータセンタースケールコンピューティングの具現化には，データセンター全体にわたって存在する各種計算資源を，高速かつプログラマブルに組み合わせることができるデータセンターファブリックが重要な鍵となる．そこでデータセンターのインフラストラクチャ層の処理，つまり，仮想化基盤の処理，ネットワークプロトコルやストレージプロトコル処理，暗号化処理などを，アプリケーション層の処理から分離し，そのようなインフラストラクチャ層の処理に専念するプロセッシングコンポーネントが登場してきている．現時点では，業界内で統一された呼称は定まっていないものの，多くのベンダーが，DPU（Data Processing Unit）と銘打って，データセンターにおける第三のプロセッサーというコンセプトで定義，展開しており，同名称が徐々にデファクトスタンダードとして定着しつつある．

3．データセンターにおける第三のプロセッサー：DPU

それでは，DPUの役割を明らかにするべく，データセンタースケールコンピューティングを構成するその他のプロセッサーとの違いを含めて改めて整理する．CPUは高度で複雑な命令セットを用いて基本となる逐次処理を担う汎用コンピューティング向け，GPUは多くのコアを備えて超高速の並列処理を行うアクセラレーテッドコンピューティング向けである．そしてDPUは，データセンターファブリックのインフラストラクチャ上で，データの移動や仲介処理を最適化する（図2）．

ベンダー間で多少の違いはあるものの，DPUは基本的に以下の三つの要素を組み合わせたプログラマブルなSOC

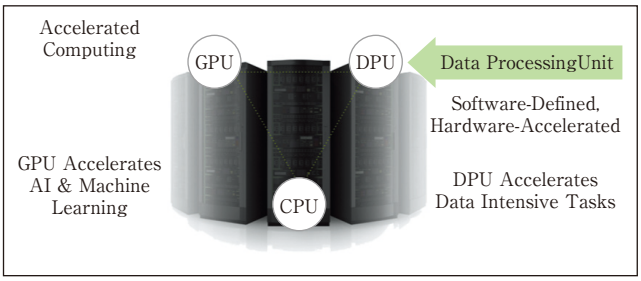


図2　データセンタースケールコンピューティングとCPU，GPU，DPUの役割

（System on a Chip）として実装されるのが一般的である．

(1) マルチコアCPU：ハイパフォーマンスでソフトウェアプログラマブルなコア．一般的に広く使用されているArmアーキテクチャなどが用いられ，その他の後述するSOCコンポーネントと密接に統合され，主にDPUにおけるコントロールプレーンの役割を担う．

(2) 高速I/Oのためのインタフェース：CPUやGPUやストレージ，そして外部のデータセンターファブリックの間で，アプリケーショントラフィックをラインレートで効率的に送受信する．

(3) プログラマブルなアクセラレーションエンジン：ASIC（Application Specific Integrated Circuit）やFPGA（Field Programmable Gate Array）で実装されるコンポーネントで，ネットワーク，セキュリティ，およびストレージなどに関する特定の処理をオフロードしてアクセラレーションすることで，システムのパフォーマンスを大幅に向上させる．

以降，より具体的なイメージをつかんでいただくために，マーケットに現存する最新のDPUの実装例を交えながら解説していく．本講座ではエヌビディア社のNVIDIA BlueField DPUシリーズを取り上げる（図3，図4）．

※なお，前述した（3）については，具体例なしにはイメージがし辛いため，NVIDIA BlueField-2 DPUにおいて実装されている代表的なアクセラレーション機能を以下に例示する．

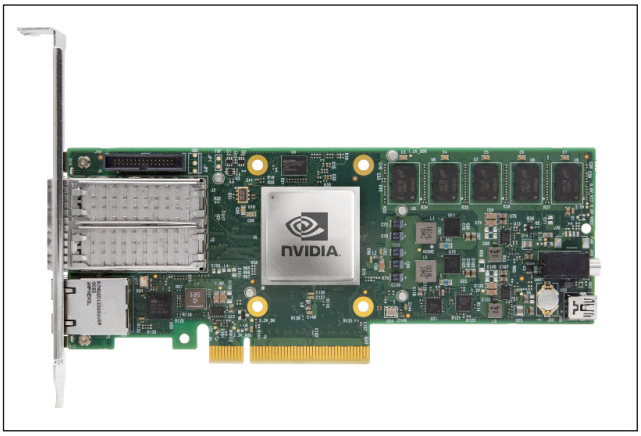


図3　NVIDIA BlueField-2 DPU外観

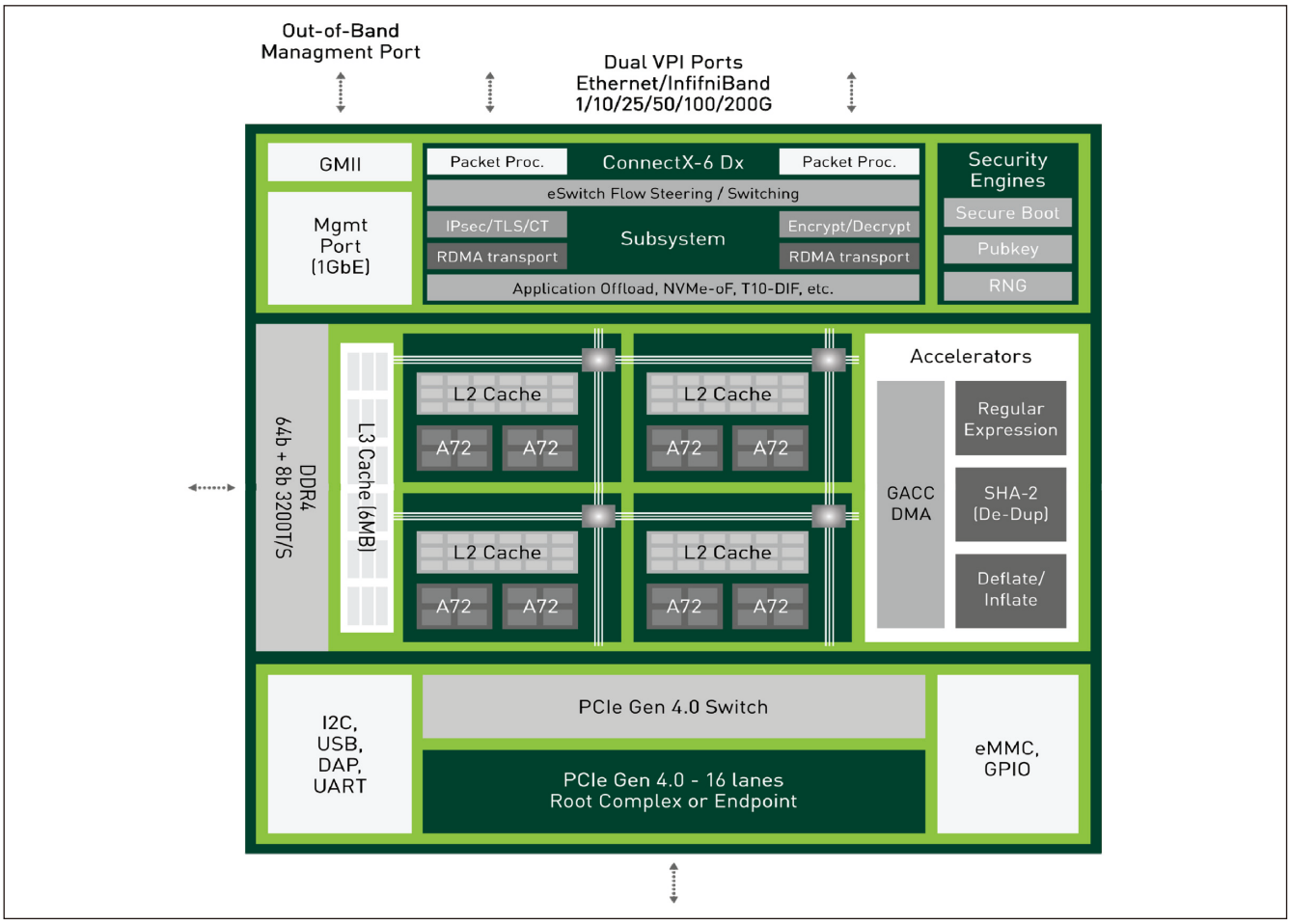


図4　NVIDIA BlueField-2 DPUブロック図

・OVS (Open Virtual Switch) [2] を実装するためのデータパケット解析，照合，操作
・RoCE (RDMA over Converged Ethernet) [3] のためのRDMA (Remote Direct Memory Access) データ転送アクセラレーション
・CPUをバイパスし，ネットワーク接続されたデータを（ストレージやその他のGPUから）GPUに直接供給するためのGPU-Directアクセラレータ
・RSS (Receive Side Scaling)，LRO (Large Receive Offload) [4]，チェックサム処理などのTCP (Transmission Control Protocol) アクセラレーション
・VXLAN (Virtual eXtensible Local Area Network) [5] などのネットワーク仮想化（オーバレイ）処理とVTEP (VXLAN Tunnel End Point) オフロード
・マルチメディアストリーミング，コンテンツディストリビューションネットワーク，および新たな4Kや8KでのVideo over IPを実現する，パケットペーシングによるアクセラレータ
・5G無線ネットワークのための時刻同期に関するアクセラレータ
・IPSEC (Security Architecture for IP) [6] およびTLS (Transport Layer Security) [7] 向けの暗号化アクセラレータ
・SR-IOV (Single Root I/O Virtualization) [8] など，および準仮想化に対応する仮想化サポート
・セキュアな分離：RoT (Root of Trust)，セキュアブート，安全なファームウェアのアップグレード，認証されたコンテナーおよびアプリケーションのライフサイクル管理

4．DPUによるデータセンタースケールコンピューティングの最適化

ここからはDPUがデータセンタースケールコンピューティングで果たす，重要な最適化シナリオを解説していく．

4.1　強固なセキュリティの実現

昨今，サイバー空間でのコラボレーションの増加や，マルチテナント環境のクラウドコンピューティングなど，意図せずにサイバーリスクに晒される場面も増えてきており，ゼロトラストセキュリティを徹底し，データセンター内のデータ通信時においても高い機密性を確保することが重要である．現在のデータセンターの多くは，外部との通信のためにDMZ (DeMilitarized Zone) エリアを設置したり，外部ネットワークと内部ネットワークの間にファイアウォールを置くなどの対策を行ってセキュリティを確保し

ている．しかし，大容量データを高いセキュリティを保って処理する場合，ファイアウォールやゲートウェイサーバ側の負荷が増大することになり，最悪の場合，同サーバが通信のボトルネックになりかねない．また，これらをより性能の高い機材に置き換えることができたとしても，現在そしてこれからのデータセントートラフィック流量と等価にスケールを続けていくのは非常に困難である．さらに，サイバーリスクの影響範囲を極小化するために，データセンター内の演算処理を行う個々のサーバ自身の仮想化基盤と統合する形態で，高度なセキュリティ機能を持たせるマイクロセグメンテーションという概念がある．このケースの場合も，大容量のトラフィックと高粒度のセキュリティ処理そのものが原因で，同サーバのCPUの負荷が増大して，本来のアプリケーション処理に計算資源を充分に確保することが困難であることは明白である．ここで，DPUの出番となる．データセンター内の各サーバのDPUのマルチコアCPUやアクセラレーションエンジンを駆使して，従来ファイアウォールやゲートウェイサーバが担っていたセキュリティ機能をオフロードする．これにより各サーバのCPUやGPUの負荷を上げることなく，高速かつ強固なデータセンターセキュリティが実現可能となる．また，サーバホストCPUのみでのセキュリティ処理では実現しえなかった，アプリケーションが稼働するホスト，CPUとインフラストラクチャ，DPUを完全に分離してセキュリティ処理を行うゼロトラスト通信も可能となる（図5）．これは，データセンターにてベアメタルインスタンスなどのサービスを提供する際に，各サーバのインフラストラクチャ層とアプリケーション層それぞれを，別の管理ドメインへ配置可能であることを意味する．

4.2　高度なストレージ仮想化

DPUに対して，分散ファイルシステムなどで必要となるクライアントのエージェント機能をオフロードすることによって，サーバのCPUリソースを消費することなく，ストレージを利用可能となり，サーバ側の計算資源をアプリケーションワークロードにてより多く活用することができる．また，ストレージのクライアント機能がDPUで動作することになるため，サーバ側のOS（Operating System）の制限がなくなり，レガシーなOSであっても，モダンなストレージシステムを利用することが可能となるであろう．

DPUでは，これに加えて仮想化ストレージのエミュレーションの役割を担うことも期待されている．リモートにあるストレージ，例えば，NVME（Non-Volatile Memory Express）[9]などの高速ストレージを，あたかもサーバ内のPCIe（Peripheral Component Interconnect Express）に接続されたローカルデバイスとしてエミュレーションする（図6）．これにより，ストレージ資源の集約によるデータセンター内のリソース最適化や，従来は実現が容易ではなかったリモートストレージをブートデバイスとして利用することなどが可能となる．

そして，これらDPUが提供するストレージ機能と，前述したDPUが備えるセキュリティ機能に包含される暗号化処理を併用することで，CPU側のオーバヘッドを最小限に保ちつつ，機密性の高いストレージ運用を実現することができる．

4.3　高性能なネットワーキング

クラウドネイティブでマイクロサービスアーキテクチャのワークロードの特徴の一つに，データセンターの横方向を往来するトラフィック，いわゆるイーストウエストトラ

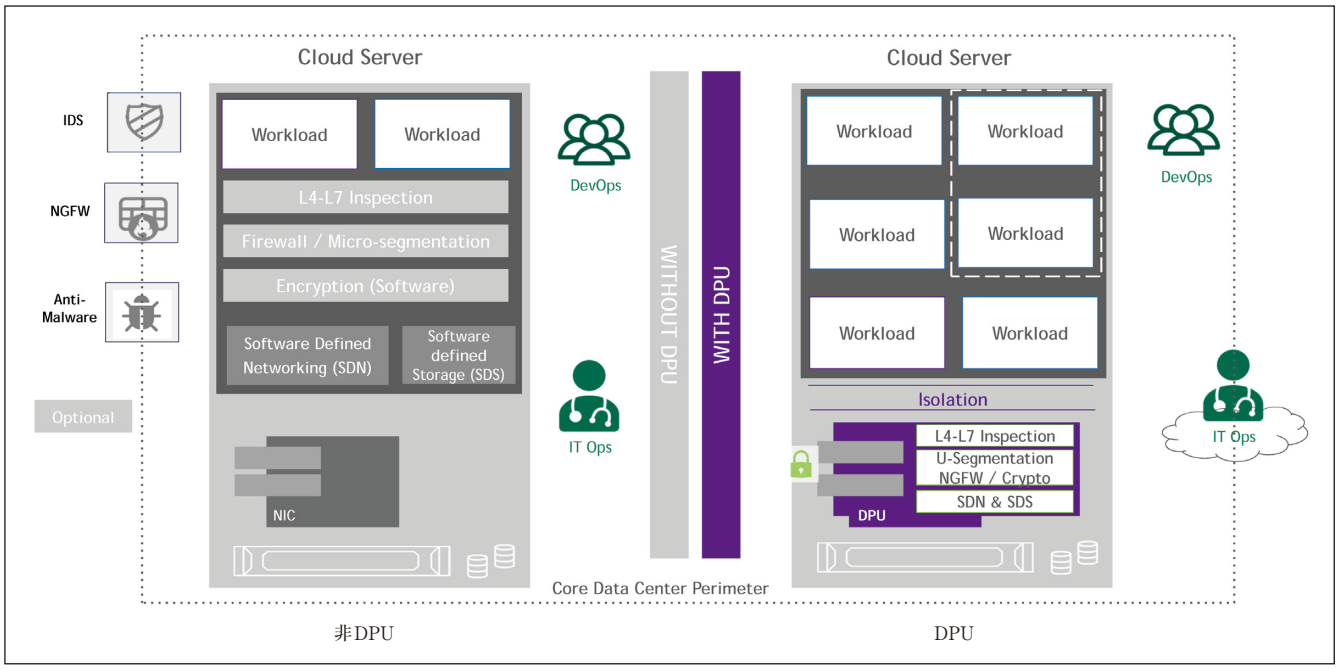


図5　DPUによるデータセンターのセキュリティ高度化

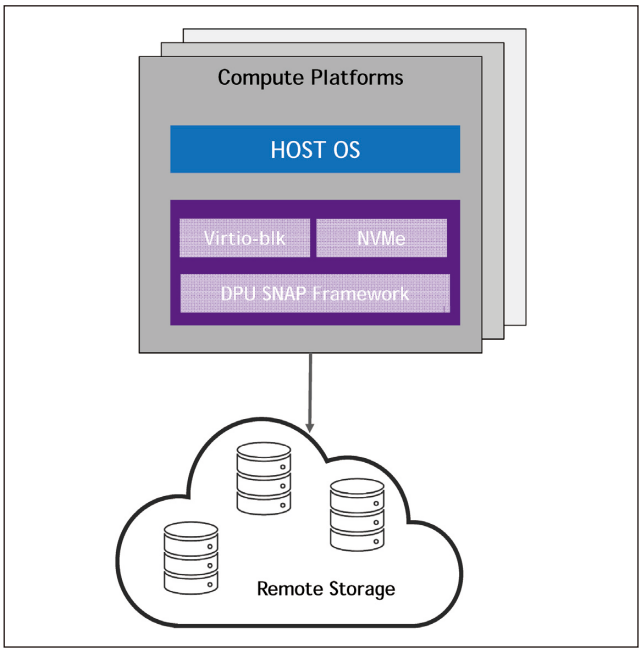


図6　DPUによるストレージのエミュレーション

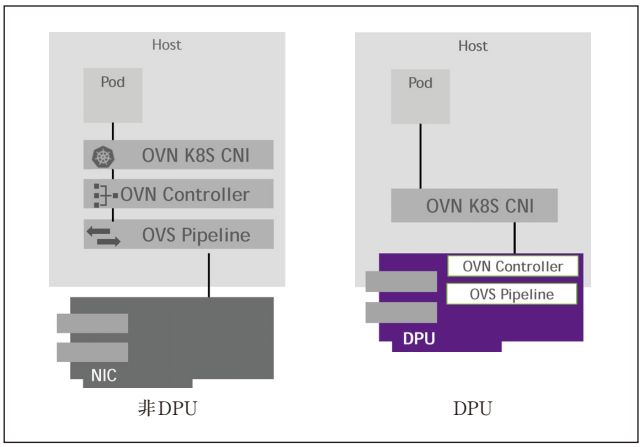


図7　仮想ネットワーク処理のオフロード

フィックの割合が大きくなることが挙げられる．ネットワーク仮想化においては，このようなトラフィックを，高スループットかつ低レイテンシーで処理しつつ，どれだけサーバのCPUオーバヘッドを極小化するかが重要となる．前者のスループットについては，DPUを活用し，ネットワーク仮想化レイヤ，すなわちOVSやOVN（Open Virtual Network）などの処理を，DPU側のアクセラレーションエンジンへオフロードして最適化することで，サーバのホストCPU側からネットワーキング処理を分離し，システムスループットの大幅な向上が期待できる．以下の図で，従来はCPU側のカーネルで行われていたOVSの転送処理を，DPUのアクセラレーションエンジンへオフロードするイメージを示す（図7）．

加えて，参考データとして一般的なデータセンタークラスサーバのソフトウェア処理のOVSスイッチングのパ

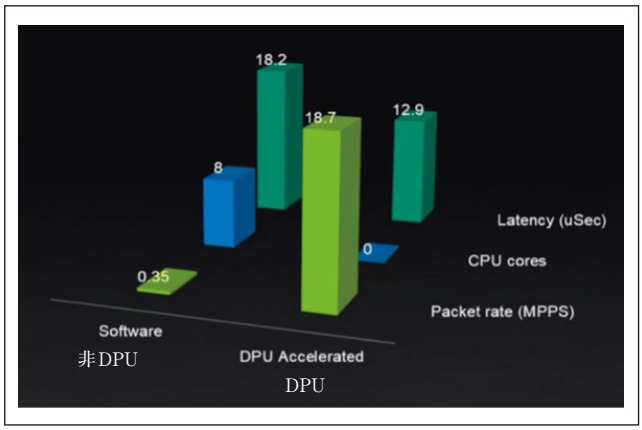


図8　OVSスイッチングのパフォーマンス比較

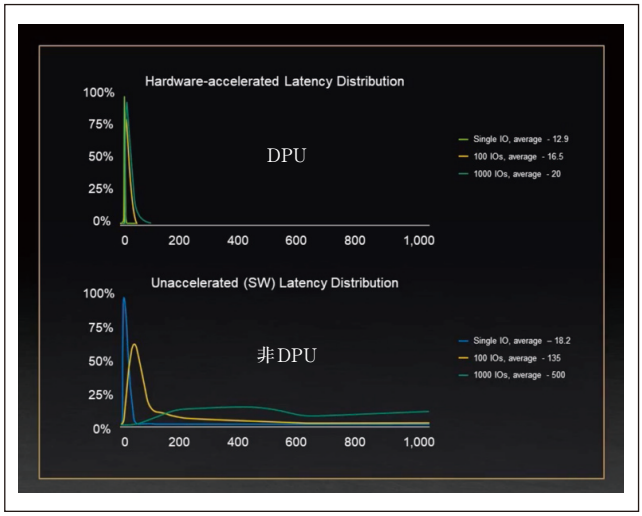


図9　レイテンシーのばらつきの比較

フォーマンスと，DPU処理（NVIDIA BlueField-2 DPU）のスイッチングのパフォーマンス比較を示す（図8）．

次に，後者のレイテンシーについてであるが，データセンター内にわたって点在するコンピューティングリソースを利用するにあたり，アプリケーションから要求されるパフォーマンスを達成するためには，ゆらぎの少ない確定的なレイテンシーで処理を完了する必要がある．そのためには，マイクロサービスを実行する個々のサーバでの処理遅延を，不確定要素による遅延変動が少ない，DPUのアクセラレーションエンジンで均一化（図9）しつつ，それらサーバ同士のプロセッシングコンポーネント間で，高速にデータを交換するRDMAを用いて，遅延なくデータ伝送する仕組みが重要となる．DPUを介して，GPU同士で直接データを交換するGPU Direct RDMAのイメージを例示する（図10）．

5．DPUの具体的な活用事例

DPUによって，仮想ストレージや仮想ネットワーク，セキュリティなどのデータセンターのインフラストラクチャ処理の大部分を，サーバホスト側のリソースからオフロー

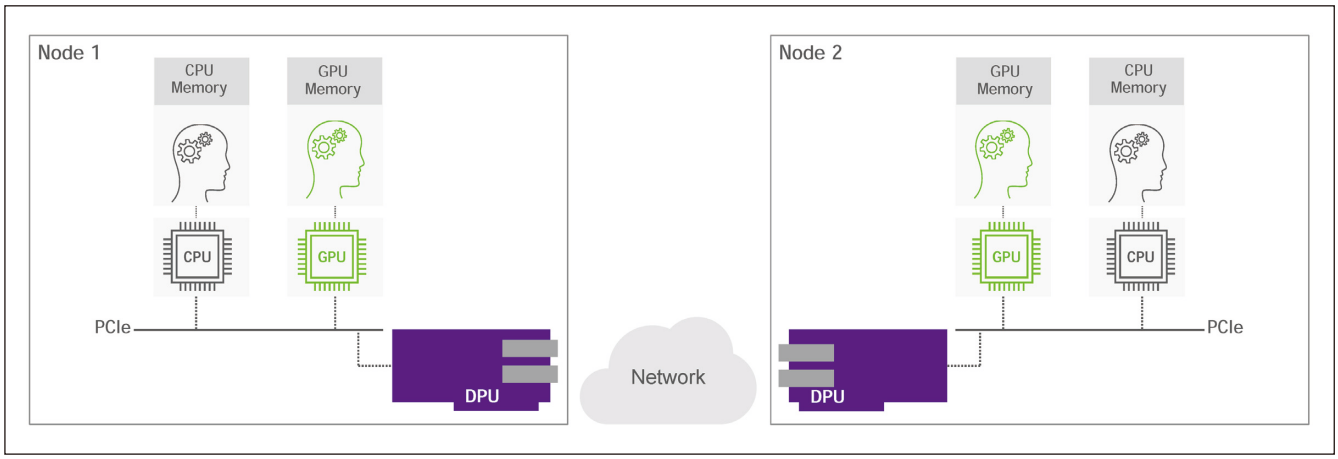


図10　GPU Direct RDMAの概要

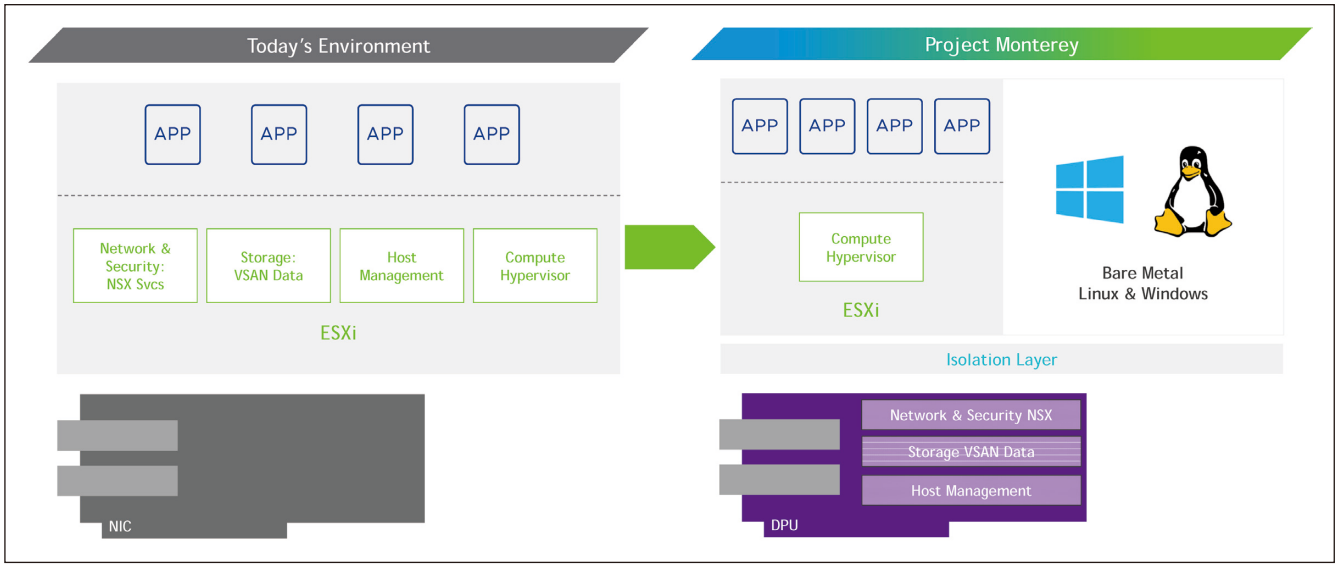


図11　VMware社Project Montereyの概要

ドして高速化することが可能となる．言い換えると，サーバホスト側の計算資源を，収益に直結する本来の仕事，つまりアプリケーションワークロードの処理に専念させることに他ならない．クラウド業界のハイパースケーラー達は，自社製の専用コントローラなどで同様のアプローチを具現化し，このような最適化をいち早く進めているが，DPUを活用することで，今では誰もがハイパースケーラーと同等のメリットを享受できる．ここでは，二つの先進的な活用事例をご紹介したい．

5.1　プライベートクラウドでの事例

図11の例では，エンタープライズで広く利用されているVMware vSphereの仮想化処理の大部分をDPUへオフロードし，最適化することで，サーバのホストCPUリソースを，仮想マシンやコンテナなどのアプリケーションワークロードに振り向けている．VMware社はこのような仕組みをProject Montereyとして掲げ，エヌビディア社をはじめとするDPUを供給するパートナーとともに開発を進めている．

5.2　クラウドサービス基盤での事例

別のユースケースとして，クラウドサービス基盤の最適化が挙げられる（図12）．ここでは，一例としてエヌビディア社が提供するGeForce NOWというクラウドゲーミングサービスを挙げる．GeForce NOWは，クラウド上のGeForceサーバに，世界中のゲームユーザがインターネットを介して接続できるゲーミングサービスだ．このサービスでは，ゲームの処理とレンダリングは，ローカルのユーザデバイスではなく，クラウド側のサーバのGPUで実行される．さらに，同サーバへのゲームのダウンロードやパッチ処理などは，すべてクラウド上で行われるので，ローカルのユーザデバイスは，そのような煩雑な処理から解放され，GPUなど特別なハードウェアを準備する必要なしに，いつでもどこでも高品質なゲームを楽しむことができる．このようにクラウド側に集約することによるメリットが大きい一方，遅延なくクラウド側からユーザのローカルデバイスへゲームをストリーミングすること，不特定多数のユーザが利用するクラウド側の仮想化環境において，セ

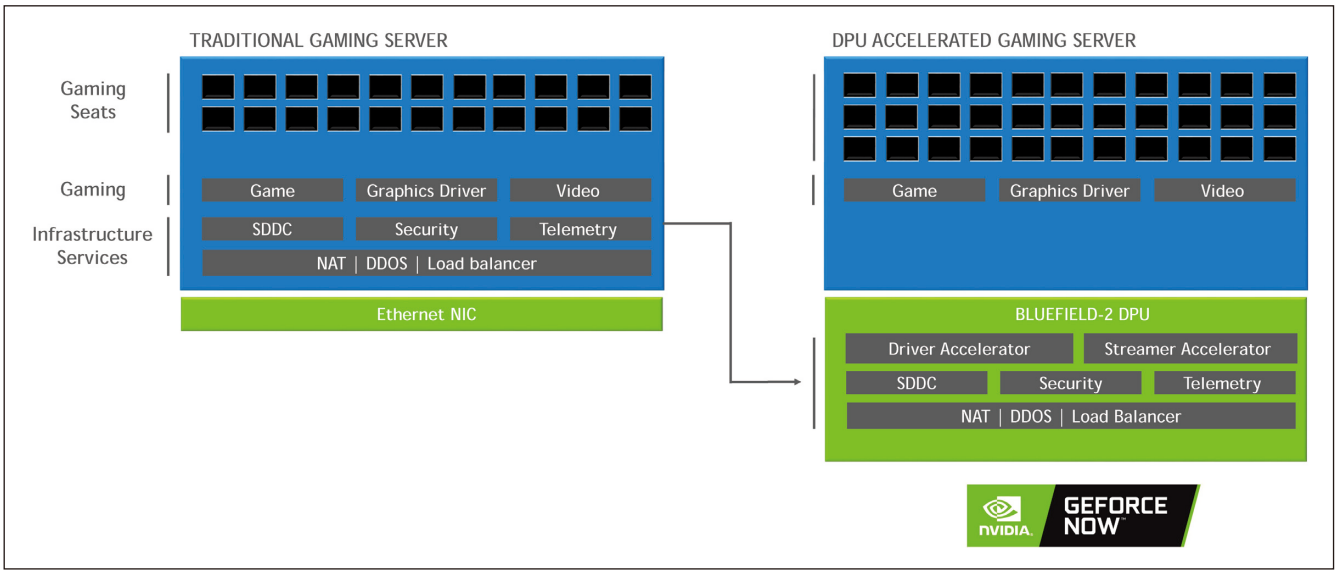


図12　クラウドサービス基盤へのDPUの適用例

キュアなマルチテナント性を担保すること，そしてアプリケーションの収容効率を高めてサービス全体の収益性を最大化していくこと，などのチャレンジもある．エヌビディア社はこのGeForce NOWのクラウドサービス基盤に，DPUを取り入れて，インフラストラクチャ層とアプリケーション層の分離を推進している．

この適用により，これまで以上に，効率良くセキュアでハイパフォーマンスなクラウドゲーミングを提供することが可能となり，より顧客満足度の高いサービスへの昇華が期待されている．

6．データセンタースケールコンピューティングに関するその他の考慮点

ここからは，データセンタースケールコンピューティングが今後更なる発展を遂げる際に，必須となるであろう二つの関連要素について考察する．

6.1　各サーバ内のCPUおよびメモリーシステムの強化

これまでデータセンター全体にわたるコンピューティングに焦点を当てて解説してきたが，一方で個々のサーバ内の処理についても，AIを中心としたワークロードへの変遷に伴って，これまでとは違った配慮が必要となってきている．冒頭の通り，昨今のAIモデルは，複雑性と規模を増しながら，大量のパラメータを調整し対話型AIを改善したり，数十テラバイト単位のデータを入力することで推薦エンジンを強化している．こうした大規模データ指向の演算要件が，個々のサーバ内のシステム限界にも直結しつつあるのだ．より具体的には，ホスト内のCPUとGPUの能力を最大限発揮し，高い精度での高速な並列演算を実行するためには，大容量メモリーへの高速アクセスが不可欠である．そのため，DPUによるインフラストラクチャの強化と歩調を合わせて，サーバホストのCPUおよびホストメモリーシステムについても更なる強化が必要となるであろう．すなわち，省電力で高性能なCPUプロセッサー，プロセッシングコンポーネント同士（GPUとCPU間など）の広帯域インタコネクト，さらに大容量の次世代メモリー（LPDDR5x[10]など）を組み合わせることによって，前述したサーバ内のボトルネックを解消し，データセンタースケールコンピューティングの更なる性能向上が期待できるはずだ．

6.2　リアルタイムのサイバーセキュリティ対策

昨今では，データセンターに限らず，企業ネットワークなどにおいてもゼロトラストのセキュリティモデルは常識となっている．このモデルの場合，データセンターのあらゆるトランザクションをリアルタイムで監視し，データセンターファブリックへの不正侵入を検知し，詳細を解析した上で，脅威を即座に検出して対策を取る必要があるが，この一連の流れをデータセンターの超高速のトラフィックレートに追随して実行しなければならないという課題が存在する．そのため，多くの場合は検査すべきデータの内，サンプリングされたほんの一部のデータしか検査がなされていないのが現状だ．この課題に対する解決法として，最先端のAIを取り入れた効率的な解析と，DPUのアクセラレーションエンジンによる検査データの広範化を組み合わせていく方法が有効だと考えられる．

具体的には，脅威の検査，特にログの精査に際して，最新の機械学習（特に自然言語処理などの深層学習）を取り入れることで，従来の手法と比較して，より効率的に暗号化されていない機密データの漏洩，フィッシング攻撃，マルウェアなどの以前は容易に特定ができなかった脅威や異常を特定，捕捉，対策を実行するエンジンが実現可能となるであろう．このようなセキュリティエンジンに対し，超高速のラインレートでデータセンターファブリックのトラフィックを供給したり，メタデータなどのテレメトリー

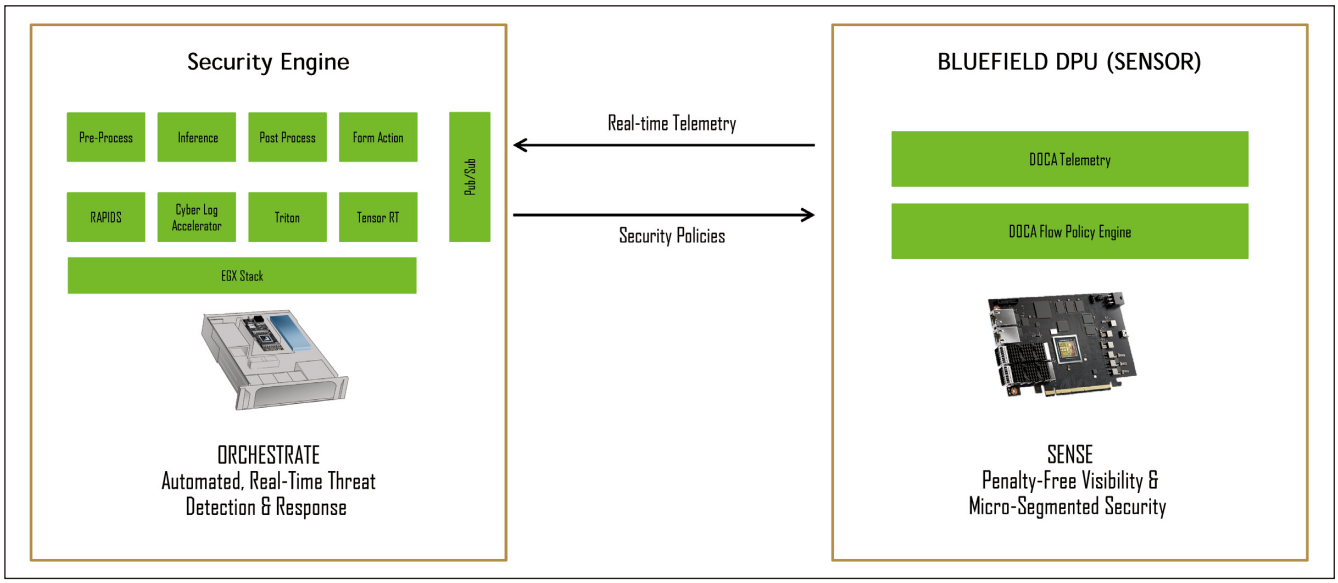


図13　AIとDPUによるリアルタイムサイバーセキュリティのイメージ

データを供給する役割を期待されているのが，それぞれのサーバに搭載されるDPUである．具体的にはDPU側からストリーミングされる検査対象データを，Kafkaなどの分散メッセージシステムを用いて集約し，前述したセキュリティエンジンの機械学習に入力し推論させ，正確で信頼性の高い脅威検出を実行する．検出後は，即座にサーバ側へフィードバックし，脅威が検出されたノードのDPUにて，特定の通信を遮断するなど，脅威の拡大を食い止める対策をリアルタイムに施す（図13）．

このアプローチにより，従来の境界型で部分的かつ受動的な脅威検出から，分散型で包括的かつ能動的な脅威検出となり得るため，データセンタースケールコンピューティングを，コアからエッジまで一貫してゼロトラストのセキュリティモデルで保護することが可能となるであろう．

7．むすび

本講座では，データセンタースケールコンピューティングが必要とされる背景，アーキテクチャ，主要な構成要素とその詳細，適用ユースケースを解説した．加えて，今後，同コンピューティングの更なる発展に必要となるであろう，個々のサーバシステムに関する考察，さらにセキュリティ面での進むべき方向性と実現案を提示した．これからのデータセンターは，IoTの本格化によりAIに入力されるデータが増大し，必要とされる演算量が指数関数的に上昇する一方，カーボンニュートラルへの対応など，地球規模での持続可能なディジタル社会の実現に向けた配慮が求められる．そのためにも，現代の情報インフラストラクチャの根幹と言っても過言ではない，世界各地に存在するデータセンターにおいて，CPU，GPU，そしてDPUが三位一体となったデータセンタースケールコンピューティングを展開し，高性能，高可用，そして周辺環境に充分配慮した高エネルギー効率を実現することは，現代の私達にとってだけでなく，これからの先の世代に対しても重要な責務であろう．本講座が，読者の皆様の多くが携わっていらっしゃる放送業界におけるデータセンタースケールコンピューティングの発展に寄与できれば幸いである．

（2021年9月10日受付）

〔文　献〕

1) IDC, The Digitization of the World (May 2020)
2) Linux Foundation, What is Open vSwitch : https://www.openvswitch.org/
3) Mellanox, RDMA over Converged Ethernet : https://docs.mellanox.com/pages/viewpage.action?pageId=12013422
4) Microsoft, Introduction to Receive Side Scaling : https://docs.microsoft.com/en-us/windows-hardware/drivers/network/introduction-to-receive-side-scaling
5) IETF, RFC7348: https://datatracker.ietf.org/doc/html/rfc7348
6) IETF, RFC 6071: https://datatracker.ietf.org/doc/html/rfc6071
7) IETF, RFC5246: https://datatracker.ietf.org/doc/html/rfc5246
8) Intel, PCI-SIG Single Root I/O Virtualization (SR-IOV) Support in Intel (r) Virtualization Technology for Connectivity: https://www.intel.com/content/dam/doc/white-paper/pci-sig-single-root-io-virtualization-support-in-virtualization-technology-for-connectivity-paper.pdf
9) nvmexpress, What is NVME?: https://nvmexpress.org
10) JEDEC, LPDDR5 & LPDDR5X : https://www.jeDec. org/category/technology-focus-area/mobile-memory-lpddr-wide-io-memory-mcp

野田（のだ）　真（まこと）　2002年より現在まで19年以上，一貫して通信に携わる．国内通信キャリヤにて基幹インフラ業務，通信系商社にて通信に関わる研究開発等に従事した後，2009年より，外資系通信機器ベンダーへ．エアヴァーナ，アリスタネットワークスにおけるテクニカルセールスを経て，2020年より，エヌビディアにて，通信領域のデベロッパーリレーションズに従事．主に通信へのAIとアクセラレーテッドコンピューティングの普及に邁進中．