

AI時代を支えるGPUサーバ群を高速に接続するネットワークを オープンな技術／オープンな仕組みで構築したい

Highlights

導入・構築のポイント

- (1)マルチテナント環境をシンプルに実現できるスイッチ・アーキテクチャ
- (2)テナント間の帯域割当を公平に行なえる
- (3)GPUサーバが求める帯域幅とボトルネックを作らない均等なレイテンシを提供できる

生成AIを始めとするAI技術の急速な進化発展を受けて、GPUコンピューティングのニーズも高まっています。GPU自体の進化も早いペースで進行しており、高性能化と同時に消費電力量や発熱量も増大しているため、多数のGPUを集めて大規模な演算処理を行なうためには相応のファシリティを用意しなくては対応が難しくなっています。こうしたニーズに応えるのがクラウド／データセンター事業者によるGPUコンピューティング・サービスです。さくらインターネットは「高火力」という印象的なネーミングでGPUサーバを提供しており、国際的に激化するAI開発競争のための基盤を提供するという重要な役割を担っています。そして、多数集められたGPUサーバを接続するためのネットワーク基盤として、Arista NetworksのDCS-7800シリーズ・スイッチが採用されました。



Customer Data

さくらインターネット株式会社

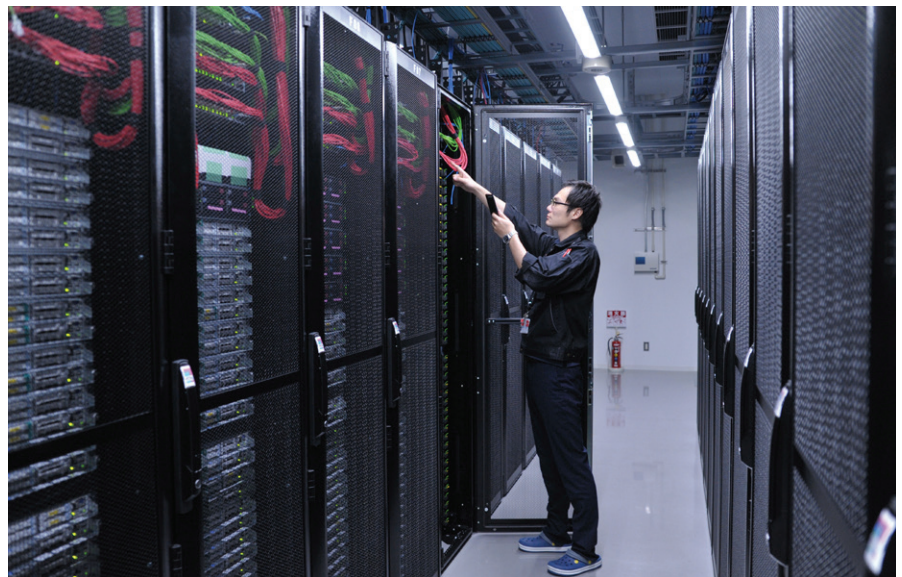
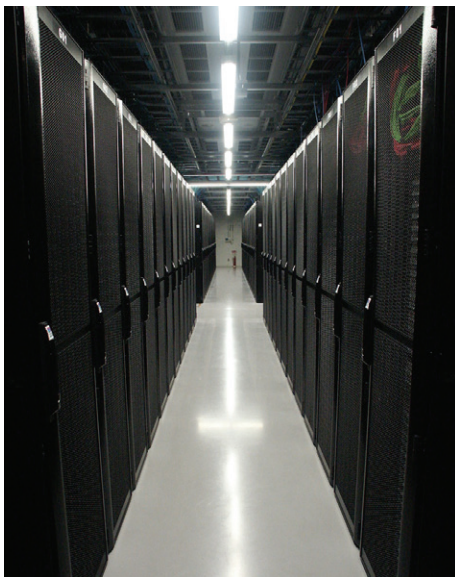
所在地 大阪府大阪市北区梅田1-12-12

創業 1996年12月



さくらインターネットは、まだクラウドという言葉が現在のような意味で使われ始めるよりも前、インターネットの黎明期である1996年創業され、当初は個人ユーザー向けのレンタルサーバ事業からスタートしました。同社は「『やりたいこと』を『できる』に変える」というビジョンを掲げ、「高い熱量を持って挑戦するすべての人たちが、自分のやりたいことを叶えられるような社会をインターネットとともにつくる」ことを目指しています。同社はトレンドの最先端を切り拓くような取り組みを次々と行なってきた実績を誇りますが、「深層学習など大量の計算資源を必要とするコンピューティング需要の高まりを受け、『高火力コンピューティング』をコンセプトとした演算能力に特化したサービスへの取り組みを開始」することを発表したのは2016年1月でした。その後のAI関連の演算ニーズの高まりを先取りし、同社は市場のニーズを踏まえてさらなるGPUサービスの拡充を目指しました。そして2023年6月に、経済安全保障推進法に基づく特定重要物資である「クラウドプログラム」の供給確保計画に関する経済産業省の認定を受け、「AIに関わるコンピューティングリソースを安定供給確保することは、日本のデジタル社会を発展させるために必要不可欠である」との認識から、生成AI向けクラウドサービス「高火力」の開発が決定されました。

2024年1月31日にはNVIDIA H100 GPU搭載サーバをベアメタルで提供する生成AI向けクラウドサービスとしてベアメタルシリーズ「高火力 PHY(ファイ)」の提供が開始されました。



さくらインターネットの「高火力」への取り組み

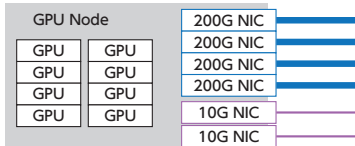
高火力サービスを統括するさくらインターネット株式会社 クラウド事業本部 クラウドサービス部 副部長の須藤 武文氏は「高火力」の出発点について、「社長の田中から『GPUをサービス／ビジネスに組み込むことはできないか』というリクエストを受けて、私がサービスに落とし込んだ形です」と言います。その際に考慮したのがGPUの消費電力の大きさで、「GPUが普通の他のプロセスよりも大きな電力を使うのであれば、それを手元に置いて使うのは早晚難しくなり、データセンターに置く必要が出てくるだろう。そのときにデータセンターにある計算機をどのようなインターフェイスでどのような形で提供するか。われわれはそのサービスを商品としてどのように作るかを考え、それがその後の機械学習の進化や生成AIの登場などの技術の進化とうまく噛み合った形です」と振り返りました。

今回提供開始された「高火力 PHY」はさくらインターネットの石狩データセンターに設置されています。石狩データセンターは2011年11月に稼働開始した郊外型データセンターで、北海道の寒冷な

気候を活かし、外気冷却を大幅に採り入れた高効率施設です。生成AIなど現在の高性能GPUをフル活用するワークロードでは、大量のデータをGPUに投入し、膨大な量の演算を繰り返すこととなります。細かなやりとりをレスポンス良く行なうことが求められるインタラクティブなアプリケーションとは異なり、データセンターまでの距離が遠いことによる遅延の影響はほぼなく、むしろ外気冷却を活用することでGPUが発生する膨大な熱量を低コストで冷却できる点が大きな強みとなります。そして、最新世代のGPUの演算性能を無駄なく活用するためには、GPUサーバー間を遅滞なく集団通信させることができる広帯域ネットワークが不可欠です。多数のGPUサーバを接続した大規模なGPUクラスタのネットワーク基盤には、Arista NetworksのDCS-7800シリーズが採用されました。



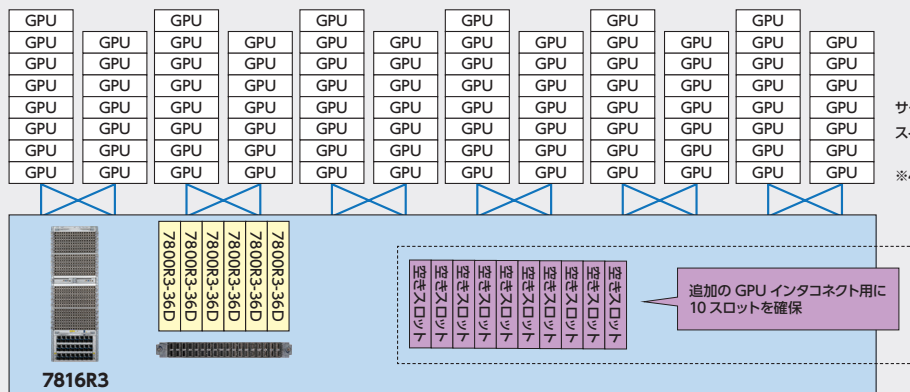
GPU クラスタ ネットワーク



各 GPU ノードは GPU x8 / 200G NIC x4 搭載
※ノード 1 台に対しラインカード 4 枚へ配線

Box 型製品での Leaf-Spine 構成ではなく単一シャーシにすることで様々なメリット

- より優れたサービスの均等性
- 非常にシンプルなマルチテナント (EVPN-VXLAN 等のプロトコルが不要)
- Leaf-Spine 間の 400G オプティクスが不要になることによる運用効率改善
- 運用台数が少ないことによる運用効率の向上



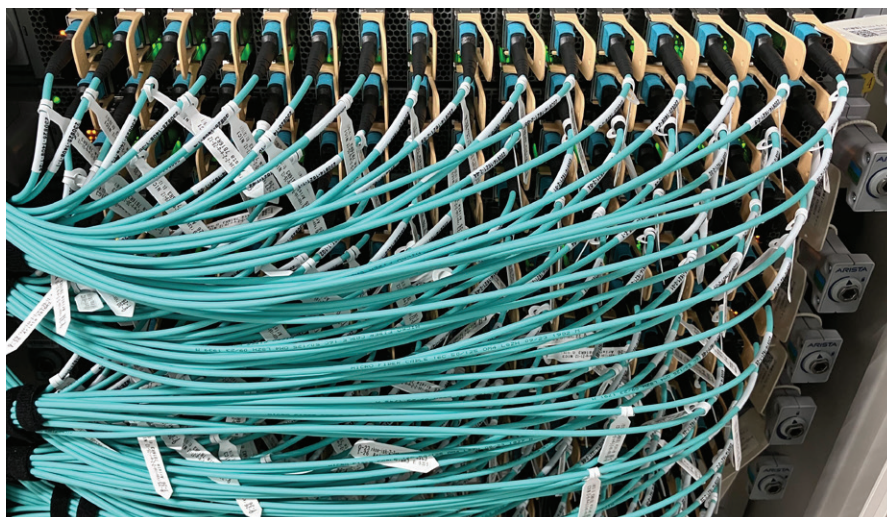
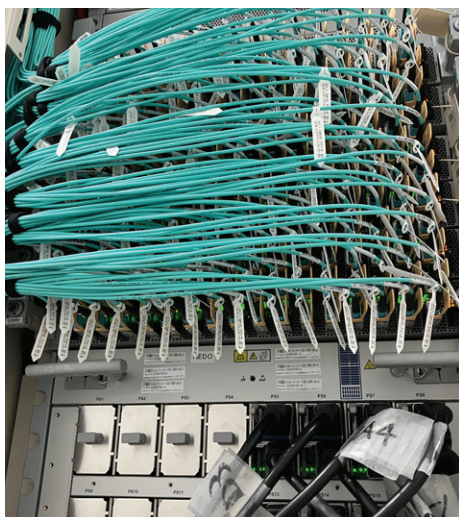
サーバ側：200G×360 ポート
スイッチ側：400G x180 ポート
(200G ヘブレイクアウト)
※400G-SR8 QSFP-DD x 180

「高火力 PHY」のシステム構成

「高火力 PHY」は2024年1月に提供開始されましたが、この環境の構築は2023年6月頃からスタートしており、システム規模を考えると極めて短期間でサービスインにこぎ着けた形となっています。システム構築を担当したさくらインターネット株式会社 クラウド事業本部 クラウドサービス部 インフラ開発リーダーの井上 喬視氏は、「高火力 PHY」の構成について「当社では現在サービスとして提供している『さくらの専用サーバー PHY』に連なる、物理サーバをクラウド化して提供するサービスを創業当初から続けていますので、基本的にはその構成の延長線上にあり、そこにさらにGPU特有の機能を付け加えた形です」と言います。同氏が最も重視したポイントとして、「インフラをマルチテナントで提供することが基本となるサービスである以上、GPUを接続するためのシステムについてもマルチテナンシーが実現できること」に加え、「6か月という超短期間で構築する必要があったため、なるべくシンプルなものにしたかった」という2点が挙げられました。

「高火力 PHY」では、「NVIDIA H100 TensorコアGPU」を8機搭載したGPUサーバを70台弱提供しています。各GPUサーバには200GbE NICが4ポート搭載されており、サーバ接続のポート数は300弱となります。この数の接続に対し、マルチテナントで必要となるトラフィックの分離をシンプルに実現したいという意図から、ネットワーク構成は一般的なLeaf-Spineではなく単一シャーシに収容し、スイッチのラインカードのポートとサーバ側のNICを直接接続しています。

Arista Networksのスイッチ・アーキテクチャでは、ラインカードを跨ぐような大規模な接続構成でも



シャーシ内の全てのポートをスイッチング接続するため、ネットワーク分割のための上位レイヤのプロトコルを利用する必要がなく、ポート間の通信がすべて独立に制御されます。井上氏は「EVPNやVXLANといったネットワーク分割のためのプロトコルを利用する場合は構成や検証に時間がかかり、運用負荷も掛かってしまいますが、今回の構成ならそうした手間がほとんど掛かりません」とそのアドバンテージを指摘しました。また、Arista Networksのスイッチはポート間のリソース割当のスケジュールの公平性に優れており、均質なサービスを提供できる点もマルチテナント環境では重要なポイントだとの指摘もありました。

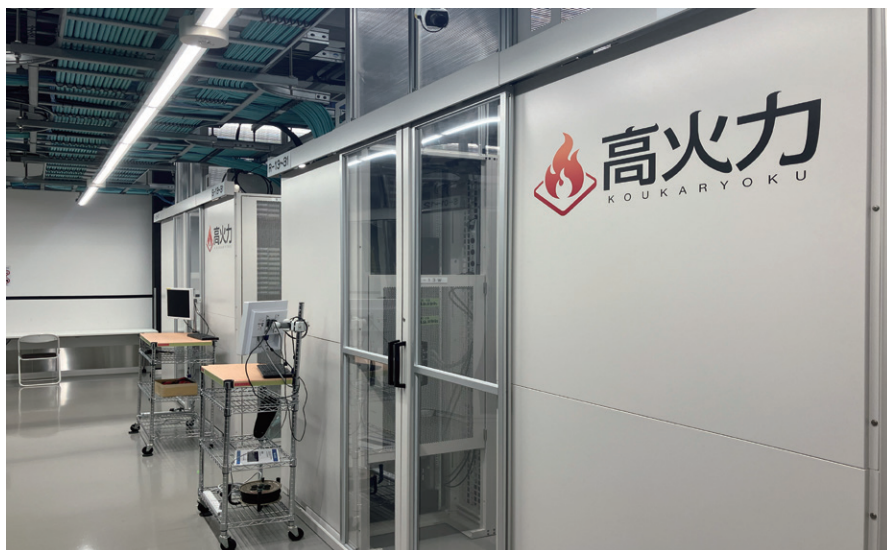
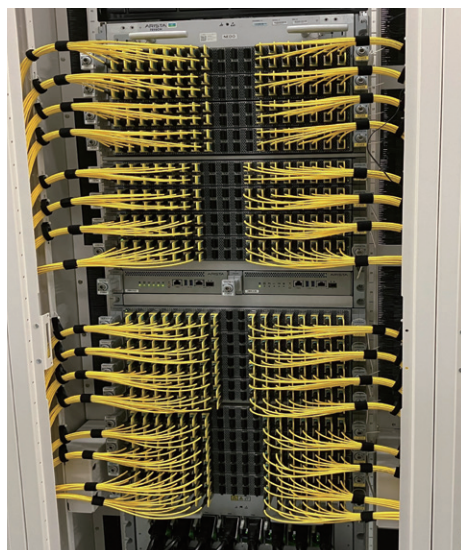
さらに同氏は、「大規模な演算を行なうユーザーさまでは多数のサーバを使って集団通信を行ないませんが、このときに一番遅いサーバに全体が引きずられてしまう傾向があるため、公平性の観点からすると、全体のレイテンシがほぼ均等であることが望ましいのです。リーフスパイン構成の場合においては、別のスイッチに収容されるなどした場合、離れた場所になる可能性も考えられますが、こうした場合でも

どのサーバをどのような組み合わせでお渡ししてもレイテンシに差が出ないこともこの構成のメリットです」と言います。

「高火力 PHY」の更なる発展

2023年6月の発表から約半年後の2024年1月にサービス提供にこぎ着けた「高火力 PHY」ですが、ユーザーからのニーズも高く、早くも追加の構築が始まっています。現時点での予定では2024年6月に全2,000基のGPUが提供可能になります。追加分では接続インターフェイスが400GbEに増速されるほかは、基本的なシステム構成は同一のものが踏襲される予定です。さらに次の構想も既に発表済で、GPUに最新世代の「NVIDIA HGX B200システム」など約10,000基を集積して総演算能力約18.9EFLOPSを実現する計画となっています。

市場では生成AIを始めとするAI技術の広範な活用がさまざまな産業分野で大きなトレンドとなっており、イノベーションを妨げないためにも十分な演算リソースを確保することの重要性が高まっている状況です。GPUはもちろん、多数のGPUを効率よく接続するためのネットワーク技術の重要性も高まる中、



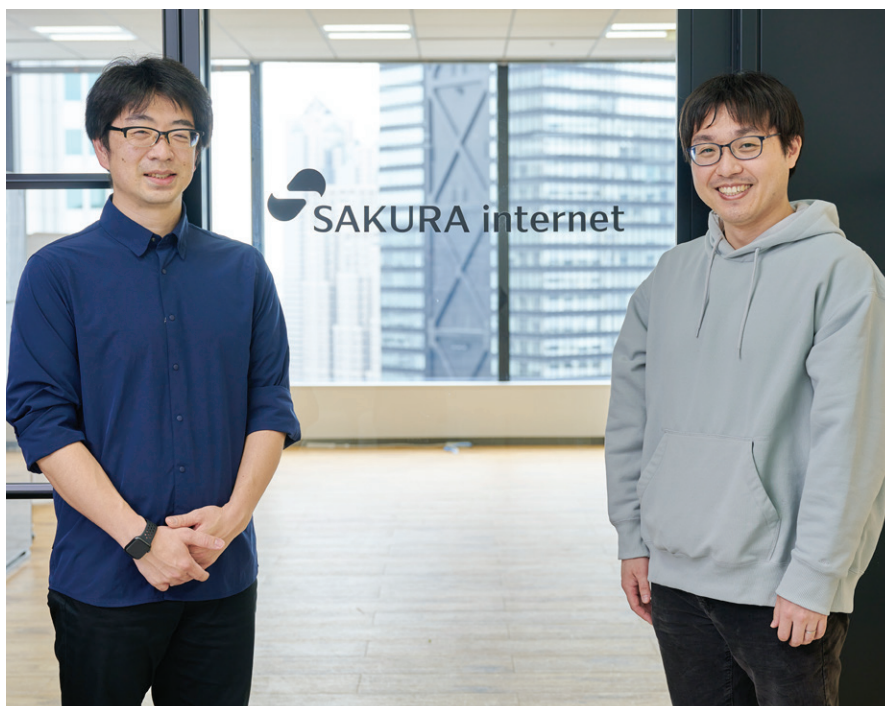
シンプルな構成で高水準のサービスを実現できるArista Networksの製品技術が「高火力 PHY」のサービスを支えています。

現在のAI系ネットワークは、技術革新の最先端だと井上氏は語ります。「以前はテレコム市場で新技術の導入が始まってそれが他に広がっていく流れだったと思いますが、今はGPUやAI系のネットワークで次々と技術革新が起こっていく流れだと思っています。今後ともArista Networksとも情報交換など連携を密にして技術革新に取り組んでいきたいと思っています」とのことです。なお、この点に関連して須藤氏は「クラウドサービスはインターネットの向こう側で何が動いているのかを推し量ることが極めて困難です。経済合理性としては使うだけで済ませる方が有利ですが、クラウドサービスを使う能力をどれだけ向上させても、それを作る能力を獲得するには至りません。海外の技術に全面的に依存するだけという状況にはリスクもあり、国内で作ることも重要だという国の方向性は我々の考えとも一致しています」と語っています。

国家レベルのIT戦略の重要な基盤として整備されている「高火力 PHY」にArista Networksが採用され、最先端のGPU/AIネットワークをさくらインターネットと連携して構築する形になったことで、今後も両社の連携から新たなイノベーションが生まれていくことに期待がかかります。

左から
さくらインターネット株式会社
クラウド事業本部
クラウドサービス部
副部長
須藤 武文氏

クラウド事業本部
クラウドサービス部 インフラ開発
リーダー
井上 喬視氏



お問い合わせ先
contact-japan@arista.com
URL
www.arista.com/jp

東京
〒100-0004
東京都千代田区大手町1-7-2
東京サンケイビル 27F

Tokyo
Level27, Tokyo Sankei Building,
1-7-2 Otemachi, Chiyoda-ku,
Tokyo 100-0004, Japan

大阪
〒530-0001
大阪市北区梅田2-2-2
ヒルトンプラザウエスト19F

Osaka
Level19 Hilton Plaza West Office Tower,
2-2-2 Umeda Kita-ku,
Osaka 530-0001, Japan

