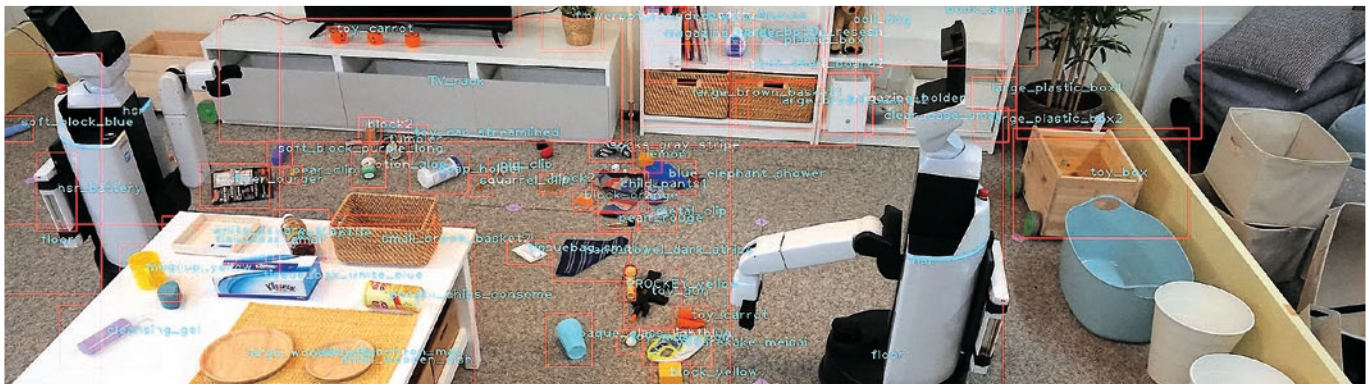
Cisco Nexus 9000 シリーズ 導入事例

# 株式会社 Preferred Networks

## 深層学習計算基盤のインターコネクトネットワークをイーサネットに統合 二重投資の回避、ボトルネック解消を同時に実現



### 製品 & サービス

- Cisco Nexus 9364C
- Cisco Nexus 9336C-FX2

### 課題

- 深層学習計算基盤「MN-1」「MN-1b」は、ストレージアクセスに 10Gbps イーサネットを採用しており、ストレージアクセスでボトルネックが発生しやすかった
- 次に構築する「MN-2」でボトルネックを解消する必要があったが、ノード間通信に利用している InfiniBand と高速イーサネットの二重投資は回避したかった
- マルチノードの深層学習計算基盤では短時間のバーストトラフィックがノード間で発生するが、その可視化も行いたかった

### ソリューション

- マルチノード深層学習で必要な RDMA をイーサネット上に実装する「RoCEv2」を採用することで、ネットワークをイーサネットに統合
- Cisco Nexus 9000 シリーズでリーフ & スパイン型ネットワークを構成することで十分な帯域を確保
- Cisco Nexus 9000 シリーズの Network Processing Unit に組み込まれたハードウェアベースのストリーミングテレメトリによって、短時間のバーストトラフィックを可視化

### 結果～今後

- ネットワーク統合によって二重投資を回避しながらボトルネックの解消に成功
- シスコ製品は安定性が高く、ネットワークエンジニアも慣れ親しんでいるので問題発生時にも迅速な対応が可能
- バーストトラフィックの可視化によって何がボトルネックになっているかが把握しやすくなり、次の投資判断が行いやすくなった

深層学習の研究開発に強みを持ち、日本を代表する企業と数多くの共同プロジェクトを推進している Preferred Networks。計算基盤を自社開発している同社が、大規模クラスタ ―「MN-2」に採用したのが Cisco Nexus 9000 シリーズです。以前はノード間通信に InfiniBand、ストレージアクセスに 10Gbps イーサネットを利用されていましたが、これらをイーサネットに統合。二重投資を行うことなく、ボトルネックを回避しやすいネットワークを実現しました。

**MN-2 構築でまず求めたのは、ネットワークのボトルネックを解消すること。しかし InfiniBand と高速イーサネットへの二重投資は避けたいと考えていました。**

―― 株式会社Preferred Networks 執行役員 計算基盤担当VP 博士（情報理工学） 土井 裕介 氏

深層学習（ディープラーニング）などの最先端技術の社会実装を目指し、2014 年 3 月に設立された Preferred Networks（以下、PFN）。同社のビジョンは「現実世界を計算可能にする」というもの。ソフトウェアとハードウェアを高度に融合することで、現実世界の課題解決を目指しています。具体的には、深層学習と多様な専門分野の深い知識を掛け合わせた最先端技術の研究開発に独自の強みを発揮。その問題解決能力は多くの企業から高く評価され、日本を代表する製造企業やプラント企業、医療機関などと共同プロジェクトを推進しています。

また独自開発した深層学習フレームワーク「Chainer」は、「Define-by-Run」というモデル記述手法を取り入れた深層学習フレームワークの先駆けとして、その後登場した「PyTorch」にも大きな影響を与えています。

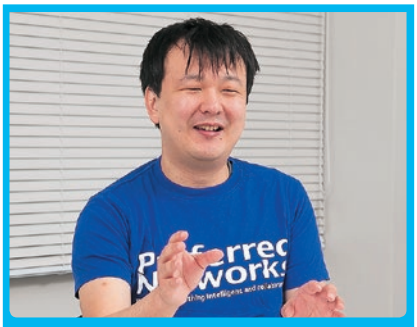
## 課題

### 民間企業トップクラスの計算基盤を自社開発

深層学習のためのハードウェアを自社開発していることも PFN の大きな特徴です。第 1 弾となる大規模並列コンピュータ「MN-1」の稼働を開始させたのは 2017 年 9 月のこと。NVIDIA 製 Tesla P100 GPU を 1024 基搭載したマルチノードコンピュータは、民間のプライベートな計算基盤としては国内最大級のものでした。

同年 11 月には、MN-1 上で分散深層学習パッケージ「ChainerMN」を動かし、一般に広く使われている画像分類データセット「ImageNet」の学習を 15 分で完了させ、当時世界最速を記録。また同時期に LINPACK 性能測定で約 1.39 ペタ FLOPS を記録し、2017 年 11 月時点のス―


株式会社Preferred Networks
執行役員 計算基盤担当VP
博士（情報理工学）
**土井 裕介** 氏



株式会社Preferred Networks
リサーチャー
博士（情報理工学）
**浅井 大史** 氏

パーコンピュータ性能ランキングを示す TOP500 リストで、産業領域のスーパーコンピュータにおいて世界 12 位、国内 1 位として登録されました。
その後、2018 年 7 月には NVIDIA Tesla V100 32GB GPU を採用し、MN-1 を拡張した「MN-1b」を稼働。さらに 2019 年 7 月には、NVIDIA V100 Tensor コア GPU を搭載した最新のマルチノード型 GPGPU（General-purpose computing on GPU：GPU による汎用計算）計算基盤「MN-2」を稼働しています。

### 第 2 世代の計算基盤構築で通信ボトルネックを解消

最新の計算基盤である MN-2 の計画が始まったのは 2016 年の夏頃。「当初は、社外サービスを利用するか自社で構築するかという議論をスタートし、2018 年秋に自社で構築することを決定しました」と PFN の土井 裕介氏は話します。
その後、スペックの具体化に着手。まず採用する GPU を決め、CPU やノードの実装方法、排熱方法などを検討していきました。ここで、重要なテーマの 1 つに挙げられたのがノード間やノードとストレージを接続するネットワークです。
従来の MN-1 や MN-1b では、ノード間をつなぐネットワークに InfiniBand、ノードとストレージをつなぐネットワークにオンボードの 10Gbps イーサネットを使用していました。InfiniBand をストレージアクセスに利用できなかった理由は、分散ファイルシステムとして採用していた HDFS（Hadoop Distributed File System）と InfiniBand との相性が良くなかったため。また、オンボードのイーサネットを使用することになったのは、各ノードに装備されていた PCI スロットに制約があったからでした。「そのためノード間通信は十分に高速である一方で、ストレージアクセスにボトルネックが発生しやすい状況でした」と PFN の浅井 大史氏は説明します。

## 信頼性の高さ、最新プロトコルへの対応の早さ、ハードウェアベースのストリーミングテレメトリがシスコを採用した理由です

## ソリューション

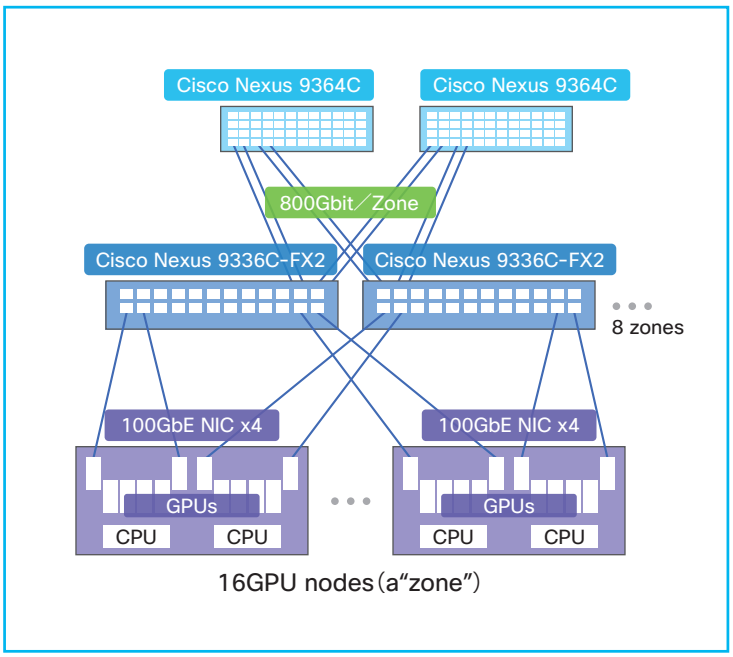
### 二重投資を回避するためにネットワークをイーサネットに統合

MN-2 におけるネットワークの課題は、ストレージアクセスのボトルネックだけではありませんでした。もう 1 つは、投資にまつわる課題です。「ボトルネックを解消するために高速イーサネットを採用して、InfiniBand と二重投資になってしまうことは避けたいと考えていました」と土井氏は言います。
InfiniBand でストレージアクセスも統合するのか、それともイーサネットでノード間通信を統合するのか――。PFN が最終的に選んだのは、高速イーサネットの上でこれまで InfiniBand で行ってきた通信を実現し、両者を統合することでした。
InfiniBand の核となるテクノロジーは、通信先ノードのメモリに直接データを書き込む RDMA（Remote Direct Memory Access）ですが、これをイーサネットのリンクレイヤーと一般的な IP/UDP レイヤーの上に乗せた「RoCE（RDMA over Converged Ethernet）v2」を採用することにしたのです。
「2018 年 10 月にスーパーコンピュータのカンファレンスに出席した際、RoCEv2 と InfiniBand を比較するチュートリアルに参加し、その時のデータから RoCEv2 でネットワークを統合できそうだと感じことが採用を後押ししました」と浅井氏は説明します。

### シスコをパートナーに選んだ 3 つの理由

MN-2 のイーサネット採用を決めた同社が、声をかけたのがシスコです。
すぐに Cisco Nexus シリーズの検証機を借り受けて、検証を行い、手応えを得た同社は、MN-2 全体のスペックを決定。8 基の GPU を搭載したノードに 100Gbps の NIC を 4 枚実装し、これらを Cisco Nexus 9336C-FX2 に接続、その上で Cisco Nexus 9364C に接続するという、リーフ &

### MN-2 のネットワーク構成イメージ




スパイン型アーキテクチャの採用を決めました。

シスコ製品を採用した理由について浅井氏は次のように説明します。

「第 1 に信頼性の高さ。私は長年にわたってネットワークの世界を見てきましたが、Cisco Nexus に搭載されている NX-OS には豊富な実績があり、非常に高い安定性を実現していると評価しています。2 つ目は、最新プロトコルへの対応が早いこと。RoCEv2 のようなカッティングエッジな技術にも、十分に対応してもらえると期待しました。そして 3 つ目が、Network Processing Unit（NPU）に組み込まれたハードウェアベースのストリーミングテレメトリが利用できることです。マルチノードの深層学習計算基盤では『All-Reduce』と呼ばれる分散計算のための通信が一定の周期で必要になり、短時間のバーストトラフィックがノード間で発生します。ハードウェアベースのストリーミングテレメトリであれば、このようなトラフィックも可視化しやすくなります」。

MN-2 の構築は 2019 年 6 月に完了。翌月から稼働を開始し、MN-1、MN-1b と合算した処理能力は、約 200 ペタ FLOPS に達しています。

### MN-2 外観

## 結果～今後

### 二重投資を回避しながら高速なネットワークを実現

ノード間通信とストレージアクセスをイーサネットに統合したことで、ネットワークへの二重投資を回避でき、MN-2 の投資効率は MN-1 や MN-1b よりも向上しています。「今回の構成では、各ノードに NIC を 4 枚実装することで帯域を広げていますが、これが可能になったのも投資効率が向上したからです」（浅井氏）。

ケーブリングもシンプルになりました。InfiniBand とイーサネットを併用する形に比べて、ケーブル数が半分に抑えられるからです。「トータルでのパフォーマンスも発揮しやすくなりました」と浅井氏は続けます。

深層学習では、学習内容によってノード間トラフィックが大きい場合もあれば、ストレージアクセスのトラフィックが大きい場合もあります。両者が独立したネットワークであれば、いずれかのボトルネックによって発揮できる性能が制約されることになります。しかしネットワークが統合されていれば帯域の配分を柔軟に行えるため、このようなボトルネックの発生を回避しやすくなります。「投資を集中することで、今後さらに規模が大きくなった場合でも、容易に拡張していけます」と土井氏は言います。

### 安定性が高く法定点検での再起動も安心して実施可能

イーサネットは HDFS との相性も良く、トラブルが発生した場合の原因究明や解決が行いやすいことも、大きなメリットだと評価されています。

「これまでネットワークのトラブルはほとんど発生していませんが、仮に発生した場合の対処を想定すると、シスコ製品が最適だと考えています。当社には Linux のエンジニアも多いため、当初はホワイトボックスでネットワークを構成する案もありましたが、Linux のエンジニアがネットワーク機器としてのホワイトボックスを使いこなせるとは限りません。一方、ネットワークがわかるエンジニアであれば、たいていはシスコ製品を使いこなせます。教育コストを最小化しつつ、対応力を高めておくことができるからです」（浅井氏）。

MN-2 を設置している施設は、年 1 回の法定点検で全ての電源を落とす必要がありますが、ネットワークにシスコ製品を使用していれば、この時も安心だといいます。シャットダウンコマンドを打たなくても電源を落とすことができ、再起動すればきちんと動き出すからです。「ホワイトボックスでは、この安心感は得られないはずです」（浅井氏）。

## バーストトラフィックの可視化で投資判断も容易に

ハードウェアベースのストリーミングテレメトリによって、「All-Reduce トラフィック」のモニタリングも容易になりました。

「深層学習の計算を高速化する上でボトルネックになりやすいのは、この All-Reduce トラフィックです。十分な帯域を用意しておかないと、このトラフィックが発生する時間が長くなり、各ノードがいかに計算を高速化しても全体の性能向上に寄与しなくなります。ただ、All-Reduce トラフィックはバーストトラフィックなので、平均値が示されるソフトウェアベースのテレメトリでは把握しきれません。それに対してハードウェアベースのストリーミングテレメトリであれば、数十ミリ秒のバーストトラフィックも可視化でき、何がボトルネックになっているのかがはっきりとわかります。それを参照することで、今後、どこに投資すればいいのかの判断も行いやすくなりました」(土井氏)。

## ネットワークに関する議論が行えることも大きな成果

MN-2 の稼働開始によって、医療系や資源系のプロジェクト、研究領域など、あらゆる領域で PFN のビジネスはさらに加速されています。また、MN-2 の構築にシスコが参加したことで、ネットワークに関するより深い議論ができるようになったことも大きな成果です。

「MN-1 や MN-1b では、実現可能性に関する安全性を考慮した結果、保守的な設計を行う必要がありましたが、MN-2 の設計では、シスコの知見も活用することで、よりチャレンジングな設計ができたと考えています。今後もシスコの Network Processing Unit や製品開発に期待しています」と土井氏は言います。

実際、PFN は、計算基盤をさらに強化するために深層学習の特徴である「行列演算」に最適化した専用チップ「MN-Core」を開発。それを搭載した「MN-3」を 2020 年 5 月に稼働させるなど、さらなる挑戦を続けています。この最新基盤も、ネットワークは MN-2 と同様の構成となる予定。ボトルネックを回避しやすい安定したネットワークを手に入れたことが PFN の挑戦と飛躍を支えています。

## 株式会社Preferred Networks

| | | | |
|---|---|---|---|
| **所在地** | 東京都千代田区大手町 1 - 6 - 1 大手町ビル | **従業員数** | 約 300 名 (2020 年 3 月現在) |
| **設立** | 2014 年 (平成 26 年) 3 月 | **URL** | https://preferred.jp/ |

2014 年に設立された、深層学習の研究・開発を手掛けるスタートアップ企業。深層学習とロボティクス技術のビジネス活用を目指し、様々な分野でイノベーションの実現を推進している。ソフトウェア開発に強みを持つだけでなく、深層学習向けプロセッサーや大規模な計算基盤も自社で開発。ソフトウェアとハードウェアを両輪に、革新的かつ本質的な技術を自ら作り上げていくことで、未知なる領域にチャレンジし続けている。




シスコシステムズ合同会社
〒107-6227 東京都港区赤坂9-7-1 ミッドタウン・タワー
http://www.cisco.com/jp


2159-2005-01A-BL