日本電気株式会社　　2026 年 1 月

NEC Composable Disaggregated Infrastructure ソリューション適用時の AI 推論性能について（概要版）

近年、生成 AI や大規模データ解析の活用が急速に拡大する中、データセンターや研究機関では「柔軟なリソース拡張」「CAPEX/OPEX の最適化」が喫緊の課題となっています。

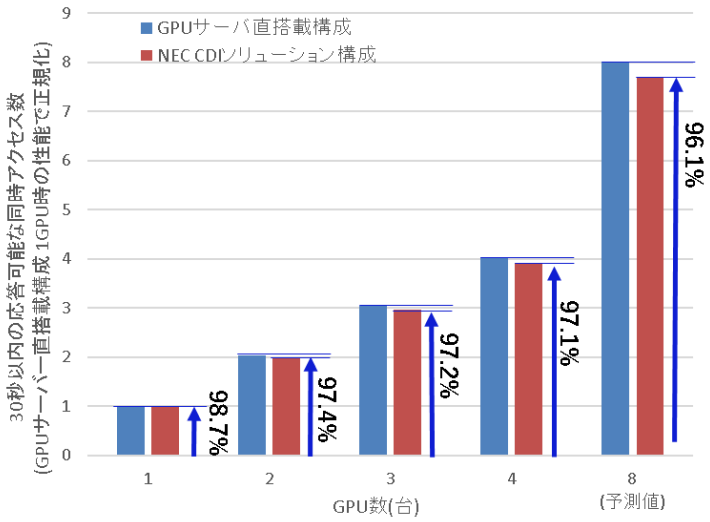
NEC の NEC Composable Disaggregated Infrastructure（CDI）ソリューション は、CPU や GPU などの計算資源を物理筐体から分離し、ネットワーク上にデータセンタースケールで分散配置できるコンピューティング基盤です。これにより、ピーク時のリソース不足や遊休資産の発生を抑え、設備投資・運用コスト（TCO）の最適化と省電力化に寄与します。

本ホワイトペーパーでは、NEC CDI ソリューションが「柔軟なリソース拡張」「CAPEX/OPEX の最適化」の課題をどこまで解消できるかを、AI 推論基盤モデル cotomi v2 による実測データと冗長構成下の耐障害性評価をもとに検証した結果をまとめています。本概要は、実測データで示したホワイトペーパーの要点を抜粋したものです。主な検証結果は次のとおりです。

1. 性能：従来構成比 97.1% を維持（1〜4GPU 実測）

分散接続でありながら、GPU 1〜4 台の構成で性能差は **1.3〜2.9%** と極小に留まり従来構成比で **97.1% を維持**できており、**実用上ほぼ等価**の推論性能を実現しました。 さらに、予測値ですが 8GPU 構成でも 96.1% の性能維持が見込まれ、大規模 GPU プールへの拡張性が確認できています。

➡ PCIe 直結と遜色ない性能を分散接続構成でも得られる。



図：AI 推論性能(2000 文字の推論を 30 秒以内に完了できる同時アクセス数)の比較

2. TCO：非同時ピークのユースケースで GPU 配備数抑制の可能性

GPU プール化と割当変更により、昼夜でワークロードが入れ替わるユースケース（例えば AI 推論と夜間バッチ処理等）で GPU を余剰なく使い切ることが可能です。効果は負荷プロファイルに依存しますが、ピークが時間的に重ならないケースでは総配備数を抑制できる可能性があります。さらに余剰 GPU の電力・冷却も発生しないため OPEX 削減にも寄与が可能と考えます。

➡ CAPEX（GPU 配備数削減）＋OPEX（余剰 GPU の電力・冷却未発生）の双方で大幅な効率化に期待。

3. ネットワーク冗長機能による高い可用性の実現

疑似障害発生時のネットワークトラフィックおよびアプリケーション性能の推移の結果から、NEC CDI ソリューションが提供するネットワーク冗長機能により、ネットワーク障害発生および復旧の前後でアプリケーション性能の顕著な悪化は観測されませんでした。PCIe 接続は本質的に「単一路」で動作し、経路冗長化を備えていないため、障害が発生した場合、アプリケーション停止やサービス劣化が即座に発生する構造的な弱点があります。一方、NEC CDI ソリューションでは、ExpEther による 100Gbps x 2 の完全二系統冗長を標準機能として備えており従来構成を上回る優位性が確認できました。

➡ NEC CDI ソリューションが提供するネットワーク冗長機能は、従来の GPU サーバ直接搭載構成では実現できない高い可用性が実現可能です。

ホワイトペーパー詳細編では以下の内容について詳細に記載しています。ご興味ある方はぜひダウンロードください。

ホワイトペーパー詳細編で記載している内容

- 評価構成の詳細（GPU 直結構成／CDI 構成、諸元・インターコネクト差分など）
- 評価方法：性能指標の定義と測定条件
- 8GPU 推定の前提（1〜4GPU 実測からロジスティック型モデルで推定）
- GPU プール化による運用（昼夜で負荷が入れ替わるユースケースの考察）
- 冗長評価の方法（疑似障害の与え方、切替挙動の確認観点）

以降のページには NEC Composable Disaggregated Infrastructure ソリューションの概要を記載していますので、合わせてお読みください。

NEC Composable Disaggregated Infrastructure ソリューションの概要

1. データセンターの課題

近年、生成 AI・高解像度映像処理・大規模データ解析といった計算負荷の急増により、データセンターでは以下の課題が顕在化しています。

- **GPU/CPU リソースの柔軟な配置が困難**

サーバはピーク時の最大負荷に合わせて構成されるため、通常時には多くの GPU・CPU が遊休状態となります。また、従来の GPU サーバでは個々の筐体にリソースが固定される構造のため、ワークロードに応じた柔軟なリソース再配置ができません。

- **過剰配置や電力浪費による OPEX 増大**

ピーク時のために大量の GPU を常時稼働させる必要があり、実際には使われないリソースにも電力・冷却コストが発生します。未使用 GPU の電源オフや省電力動作が行えず、電力費が増大します。

- **GPU 世代交代に伴う筐体ごとの交換コスト（CAPEX）が肥大化**

GPU などの高性能デバイスをサーバ筐体に直接搭載する従来方式では、デバイスだけを更新することが難しく、サーバ丸ごとの交換が必要です。このため設備投資（CAPEX）が大きく増加し、更新サイクルも短期化しています。

- **電源・冷却の局所集中による運用制約**

高性能 GPU の集積はラック単位の消費電力や冷却上限に達しやすく、データセンター内での配置制約が拡大。フロアや建屋を跨いだ柔軟な設備配置ができず、リソース増設のボトルネックとなっています。

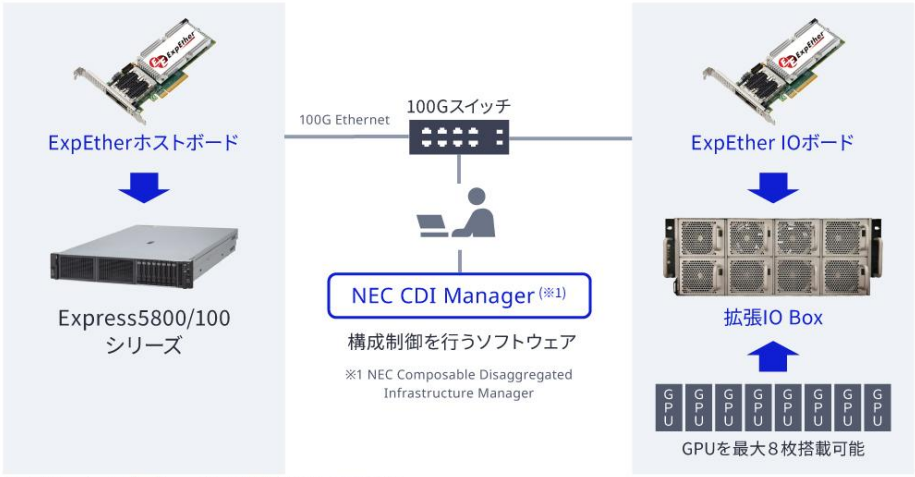
2. NEC Composable Disaggregated Infrastructure ソリューション

このようなデータセンターの課題に対し NEC は新たなコンピューティング基盤として NEC Composable Disaggregated Infrastructure(CDI) ソリューションを提供しています。NEC CDI ソリューションは、GPU や CPU などの計算資源を物理筐体から分離し、標準 Ethernet 上で データセンタースケールに分散配置・接続するコンピューティング基盤です。中核技術として NEC 独自の ExpEther を活用し、100Gbps Ethernet を用いた遠隔 I/O 拡張を高信頼・低遅延に実現します。これにより、フロア／建屋を横断し、各拠点の電源・冷却能力に応じた 柔軟な動的構成 を可能にします。

3. NEC CDI ソリューションの構成

本ソリューションは、100Gbps Ethernet の光ファイバー接続を可能にする 100G 版 ExpEther IP コアを搭載した ExpEther ボード、8 台の GPU を搭載可能な拡張 IO Box、効率的なリソース運用を可能とする NEC Composable Disaggregated Infrastructure Manager で構成されます。これらの製品により、筐体や設置場所の制約を受けず、コンピューティングリソースをデータセンタースケールで分散配置することが可能です。これにより、データセンターや企業内のコンピューティングリソースを、各拠

点の電源や冷却能力に応じて、異なるフロアや建屋を横断して柔軟かつ効率的に動的に構成できます。また、設備更新もデバイス単位で柔軟に行えるため、導入コストを抑えつつ高い拡張性を維持できます。



※ 本ソリューションの構成については、CDI構成ガイドをご確認ください。
CDI構成ガイドに記載のない構成で本CDIソリューションを実現したい場合は個別にお問い合わせください。

図 2　NEC CDI ソリューションの製品構成

製品名	製品概要
ExpEther ホストボード ExpEtherIO ボード	100G 版 ExpEther IP コアを活用しており、サーバおよび拡張 IO Box に搭載することで、100Gbps×2 のネットワークを活用し、最大 2km 離れた I/O デバイスの遠隔拡張を高信頼・低遅延に実現します。
拡張 IO Box	PCIe Gen5×16 スロットを 8 基搭載し、最大 8 枚の GPU を収容可能です。
NEC Composable Disaggregated Infrastructure(CDI) Manager	サーバと I/O デバイスの接続・切断や電源制御、稼働状況の監視を GUI で一元化。構成情報やスペックも可視化でき、認証・認可機能を備えた高信頼な運用基盤です。今後は性能・電力監視やトポロジー表示など、さらなる機能拡張も予定し、AI/HPC 分野の効率的なリソース運用を支援します。

4. NEC CDI ソリューションの特徴

本ソリューションの主な特徴を以下にまとめます。

計算資源のプール化と動的割り当て：時間帯・用途で変動する負荷に応じて、必要な GPU をサーバ間で柔軟に再配分。余剰資源は電源オフにより省電力化。

CAPEX/OPEX の削減：デバイス単位の設備更新が可能になり、サーバ丸ごと交換を回避。未使用資源の計画的な停止により運用コストを抑制。

電源・冷却制約の緩和：分散配置により、フロア／建屋を跨いだ最適配置が可能。電力・冷却の局所集中を避け、増設の自由度を向上。

高信頼性：ネットワーク冗長構成による障害時の自動フェイルオーバー、再送・輻輳制御、監視連携で冗長経路を確保し、分散環境でも継続運用を支える。

改版履歴

版	日付	変更内容
1	2026/1/30	初版作成

免責事項

本書の内容の一部または全部を無断で複写・改変・再配布することは禁止します。
本書の内容に関しては、将来予告なしに変更することがあります。
本書の作成者および作成に関連する部門は、本書の技術的もしくは編集上の誤記・欠落・瑕疵が存在する場合においても、一切の責任を負いません。
本書の作成者および作成に関連する部門は、本書の内容に沿った操作を行って生じた事象(障害・不具合、およびこれに限らず全ての現象)、ならびに、本書の内容に沿った操作を行ったにも関わらず記載と異なる動作・結果・障害が生じた場合に関して、一切の責任を負いません。

商標について

- ✓ 記載の会社名および商品名は各社の商標または登録商標です。
- ✓ Ethernet は、富士ゼロックス社の登録商標です。
- ✓ PCI-Express は PCI-SIG の登録商標です。
- ✓ その他、記載されている会社名、製品名は、各社の登録商標または商標です。

その他、本書の免責事項は「免責事項」の項を参照ください。