ホワイトペーパー
Cisco public

AMD EPYC 第 4世代 および第 5世代プロセッサを搭載した Cisco UCS M8 プラットフォームのパフォーマンス調整

目次

本書の目的

Basic Input/Output System（BIOS）は、システムのハードウェア コンポーネントをテストおよび初期化し、ストレージ デバイスからオペレーティング システムを起動します。一般的な計算システムには、システムの動作を制御するいくつかの BIOS 設定があります。これらの設定の一部は、システムのパフォーマンスに直接関係します。

このドキュメントでは、AMD EPYC™（第 4 世代および第 5 世代 プロセッサを搭載した Cisco Unified Computing System™（Cisco UCS® ）M8 サーバーに有効な BIOS 設定について説明します。Cisco UCS X215c M8 コンピューティングノード、Cisco UCS C245 M8 ラックサーバー、および Cisco UCS C225 M8 ラックサーバーの最高のパフォーマンスとエネルギー効率の要件を満たすように BIOS 設定を最適化する方法について説明します。

このドキュメントでは、AMD EPYC第 4 世代および第 5世代 プロセッサを搭載した Cisco UCS M8 サーバーで、さまざまなワークロードタイプに対して選択できる BIOS 設定についても説明します。BIOS オプションを理解することで、適切な値を選択して最適なシステム パフォーマンスを実現できます。

このドキュメントでは、AMD EPYC 第 4 世代 および第 5 世代プロセッサに基づく Cisco UCS M8 サーバーの特定のファームウェア リリースの BIOS オプションについては説明しません。ここに示す設定は一般的なものです。

このドキュメントの内容

システム BIOS のパフォーマンス オプションを設定するプロセスは、その設定オプションの中に解りにくい選択肢が含まれていることがあり、作業が複雑で困難になる場合があります。ほとんどのオプションは、サーバを最適化するために、省電力を優先するか、パフォーマンスを優先するかを選択する必要があります。このドキュメントでは、第 4 世代および第 5 世代 AMD EPYC ファミリ CPU を使用する Cisco UCS M8 サーバーで最適なパフォーマンスを実現するのに役立つ一般的なガイドラインと推奨事項を示します。

AMD EPYC 9004 シリーズ プロセッサ

AMD EPYC 9004 シリーズ プロセッサは、革新的な Zen 4 コアと AMD Infinity アーキテクチャで構築されています。AMD EPYC 9004 シリーズ プロセッサは、コンピューティング コア、メモリ コントローラ、I/O コントローラ、信頼性、可用性、有用性（RAS）、およびセキュリティ機能を統合型のチップ（SoC）に組み込みます。AMD EPYC 9004 シリーズ プロセッサは、以前成功した AMD EPYC プロセッサの実績のあるマルチチップ モジュール（MCM）チップレット アーキテクチャを保持しながら、SoC コンポーネントをさらに改善しています。SoC には、Zen 4 ベースのコアを含むコア複合（CCX）を含むコア複合ダイ（CCD）が含まれています。

AMD EPYC 9004 シリーズ プロセッサは、新しい Zen 4 コンピューティング コアに基づいています。Zen 4 コアは 5nm プロセスを使用して製造され、前世代の Zen コアに比べてサイクルごとの指示（IPC）のアップリフトと周波数の向上を提供するように設計されています。各コアでは、前世代に比べて L2 キャッシュが大きくなり、キャッシュ効率が向上しています。

各コアは同時マルチスレッディング（SMT）をサポートします。これにより、2 つの独立したハードウェア スレッドを個別に実行し、対応するコアの L2 キャッシュを共有できます。

コアコンプレックス（CCX）では、最大 8 つの Zen 4 ベースのコアが L3 または最終レベルキャッシュ（LLC）を共有します。同時マルチスレッディング（SMT）を有効にすると、単一の CCX で最大 16 の同時ハードウェアスレッドをサポートできます。

AMD EPYC 9004 シリーズ プロセッサには、AMD 3D V キャッシュ ダイ スタッキング テクノロジーが含まれており、9700 シリーズ プロセッサとチップレットのより効率的な統合を可能にします。AMD 3D チップレット アーキテクチャは L3 キャッシュ タイルを垂直にスタックし、ダイごとに最大 96MB の L3 キャッシュ（およびソケットごとに最大 1 GB の L3 キャッシュ）を提供すると同時に、すべての AMD EPYC 9004 シリーズ プロセッサ モデルとのソケット互換性を提供します。

AMD 3D V-Cache テクノロジーを搭載した AMD EPYC 9004 シリーズ プロセッサは、銅線と銅線のハイブリッド ボンディング「バンキングレス」のチップオンワファプロセスに基づく業界をリードするロジックスタッキングを採用しており、200 倍以上で現在の 2D テクノロジー（およびはんだバンプを使用した他の 3D テクノロジーの相互接続密度の 15 倍）。これにより、低遅延、高帯域幅、電力および熱効率が向上します。

CCD は、更新された I/O ダイ（IOD）を介して、メモリ、I/O、および相互に接続します。この中央の AMD Infinity ファブリックは、CCX、メモリ、および I/O を相互接続するためのデータ パスと制御のサポートを提供します。各 CCD は、専用の高速グローバル メモリ インターコネクト（GMI）リンクを介して IOD に接続します。IOD はキャッシュの一貫性を維持し、さらに、xGMI または G リンクを介して潜在的な 2 番目のプロセッサにデータ ファブリックを拡張するためのインターフェイスを提供します。AMD EPYC 9004 シリーズ プロセッサは、最大 4 つの xGMI（または G リンク）をサポートし、最大速度は 32Gbps です。IOD は、DDR5 メモリ チャネル、PCIe Gen5、CXL 1.1+、および Infinity Fabric リンクを公開します。IOD は、DDR5 メモリをサポートする 12 個のユニファイド メモリ コントローラ（UMC）を提供します。

各 UMC は、チャネル（DPC）ごとに最大 2 つのデュアル インライン メモリ モジュール（DIMM）をサポートできます。ソケットごとに最大 24 の DIMM があります。第 4 世代 AMD EPYC プロセッサは、ソケットあたり最大 6TB の DDR5 メモリをサポートできます。前世代の AMD EPYC プロセッサと比較して、メモリ チャネルが追加され、高速になったことで、メモリ帯域幅が追加され、コア数の多いプロセッサに電力を供給できます。2、4、6、8、10、および 12 チャネルでのメモリインターリーブは、さまざまなワークロードおよびメモリ構成への最適化に役立ちます。

各プロセッサは、4 つの P リンクと 4 つの G リンクのセットを持つことができます。OEM マザーボードの設計では、G リンクを使用して、2 番目の第 4世代 AMD EPYC プロセッサに接続するか、追加の PCIe Gen5 レーンを提供できます。第 4 世代 AMD EPYC プロセッサは、最大 8 セットの x16 ビット I/O レーンをサポートします。つまり、シングル ソケット プラットフォームでは 128 の高速 PCIe Gen5 レーンが、デュアル ソケット プラットフォームでは最大 160 レーンです。

AMD EPYC 9004 シリーズ第 4世代 プロセッサは、表 1 に示す仕様で構築されています。

表 1　AMD EPYC 9004 シリーズ第 4世代 プロセッサの仕様

項目	仕様
Cores プロセス テクノロジー	5 ナノメートル（nm）Zen 4
コアの最大数	128
Maximum memory speed	毎秒 4800 メガ転送（MT/s）
最大メモリ チャネル数	ソケットあたり 12
最大メモリ容量	ソケットあたり 6 TB
PCI	1 ソケットに 128 レーン（最大） 2 ソケットの場合、160 レーン（最大） PCIe Gen 5

AMD EPYC 9004 シリーズ プロセッサのマイクロアーキテクチャの詳細については、「AMD EPYC 9004 シリーズ プロセッサのマイクロアーキテクチャの概要」を参照してください。

AMD EPYC 9005 シリーズ プロセッサ

第 5 世代 AMD EPYC プロセッサを使用したシステムは、データセンターの統合や最新化から要求の厳しいエンタープライズ アプリケーションのニーズに至るまで、IT イニシアチブをサポートできます。これらのシステムは、仮想化やクラウド環境の高密度サポートにより、エネルギー効率を向上させ、データセンターのスプロールに対処するためのビジネス上の課題をサポートしながら、企業内で AI を拡張することを可能にします。IT インフラストラクチャの近代化は、既存のデータセンターのフットプリント内で AI やその他の革新的なビジネスイニシアチブに対応するためのスペースとエネルギーを解放するための鍵となります。

AMD EPYC プロセッサは、新しい世代ごとに 1 クロックあたりの命令（IPC）パフォーマンスで常に 2 桁のゲインを達成しています。第 5 世代 AMD EPYC プロセッサの最新の Zen 5 コアは、ML、HPC、およびエンタープライズのワークロード。効率性のために最適化された Zen 5c コアは、x86 アーキテクチャプロセッサの中で最大のコア数で CPU に電力を供給し、仮想化およびクラウドワークロードに最大のコア密度を提供します。

第 5 世代 AMD EPYC プロセッサにより、継続的に拡大するワークロードの需要のユニバースに対応できます。Cisco のハイブリッド マルチチップ アーキテクチャにより、イノベーションパスを分離し、革新的で高性能な製品を一貫して提供することができます。Zen 5 および Zen 5c コアは、非常に複雑な機械学習および推論アプリケーションの新しいサポートにより、最新世代からさらに大幅に進化しています。

第 5 世代 AMD EPYC プロセッサでは、2 つの異なるコアを使用して、コアのタイプと数、およびパッケージ方法を変えることで、さまざまなワークロードのニーズに対応します。

Zen 5 コア

このコアはハイパフォーマンス向けに最適化されています。最大 8 つのコアが組み合わされて、32 MB の共有 L3 キャッシュを含むコアコンプレックス（CCX）になります。このコアコンプレックスはダイ（CCD）上に構築され、SP5 フォームファクタでは最大 128 コア用に最大 16 個を EPYC 9005 プロセッサに構成できます。前世代と比較して、高度な Zen 5 コアを搭載し、メモリの高速化やその他の主要な CPU の改善も行われた第 5 世代 AMD EPYC プロセッサは、同じ 360W TDP 範囲 9xx5-070、9xx5-073 内で動作する 64 コア プロセッサで、整数パフォーマンスが 20 パーセント、浮動小数点パフォーマンスが 34 パーセント向上しています。

Zen 5c コア

このコアは、密度と効率性が最適化されています。コアは Zen 5 コアと同じ登録転送ロジックを備えていますが、物理レイアウトに必要なスペースが少なくなり、1 ワットあたりのパフォーマンスが向上するように設計されています。Zen 5c コアコンプレックスには、最大 16 のコアと共有 32 MB L3 キャッシュが含まれています。これらの CCD を最大 12 個の I/O CCD と組み合わせることで、SP5 フォーム ファクタで最大 192 コアの CPU を実現できます。

AMD EPYC 9005 シリーズ第 5 世代 プロセッサは、表 2 に示す仕様で構築されています。

表 2　　AMD EPYC 9005 シリーズ第 5 世代 プロセッサの仕様

項目	仕様
Cores プロセス テクノロジー	4 ナノメートル（nm）Zen 5 および 3 ナノメートル Zen 5c
コアの最大数	192
最大 L3 キャッシュ	512 MB
最大メモリ速度	毎秒 6000 メガ転送（MT/s）
最大メモリ チャネル数	ソケットあたり 12
最大メモリ容量	ソケットあたり 6 TB
PCI	1 ソケットで 128 レーン（最大） 2 ソケットの場合、160 レーン（最大） PCIe Gen 5

注：　Cisco UCS M8 プラットフォームは、Zen 5c プロセッサの最大 160 コア 400W TDP のみをサポートします。

AMD EPYC 9005 シリーズ第 5 世代 プロセッサのマイクロアーキテクチャの詳細については、「AMD EPYC 9005 シリーズ プロセッサ マイクロアーキテクチャの概要」を参照してください。

Non-Uniform Memory Access（NUMA）トポロジ

AMD EPYC 9004 および 9005 シリーズ プロセッサは、Non-Uniform Memory Access（NUMA）アーキテクチャを使用しています。NUMA：プロセッサコアからメモリおよび I/O コントローラへの近接性に応じて、異なる遅延が発生する可能性があります。同じ NUMA ノード内でリソースを使用すると優れたパフォーマンスを実現できますが、異なるノードのリソースを使用すると遅延が増えます。

ユーザーはシステムの NUMA ノード/ソケット（NPS）BIOS 設定を調整して、特定の動作環境とワークロードにこの NUMA トポロジを最適化できます。たとえば、NPS = 4 を設定すると、プロセッサがクワドラントに分割され、各クワドラントには 3 つの CCD、3 つの UMC、および 1 つの I/O ハブが割り当てられます。プロセッサとメモリとの最も近い I/O 距離は、同じクワドラント内のコア、メモリ、および I/O 周辺機器間です。最も遠い距離は、交差クワドラントのコアおよびメモリ コントローラまたは I/O ハブ（または 2P 構成の他のプロセッサ）の間です。NUMA ベースのシステムにおけるコア、メモリ、および IO ハブ/デバイスの局所性は、パフォーマンスを調整する際の重要な要素になります。

第 4 世代 EPYC プロセッサでは、Infinity ファブリック インターコネクトへの最適化により、遅延の違いがさらに減少しました。EPYC 9004 シリーズ プロセッサを使用して、最後の 1 または 2% の遅延をメモリ参照以外に収める必要があるアプリケーションの場合、メモリ範囲と CPU ダイ（Zen 4 または Zen 4c）の間にアフィニティを作成することで、パフォーマンスを向上させることができます。図 1 はこれがどのように機能するかを示しています。NPS = 4 構成で I/O ダイを 4 つのクワドラントに分割すると、6 つの DIMM が 3 つのメモリコントローラに給電され、インフィニティ ファブリック（GMI）を介して最大 3 つの Zen 4 CPU のセットに密接に接続されます。または最大 24 個の CPU コア。

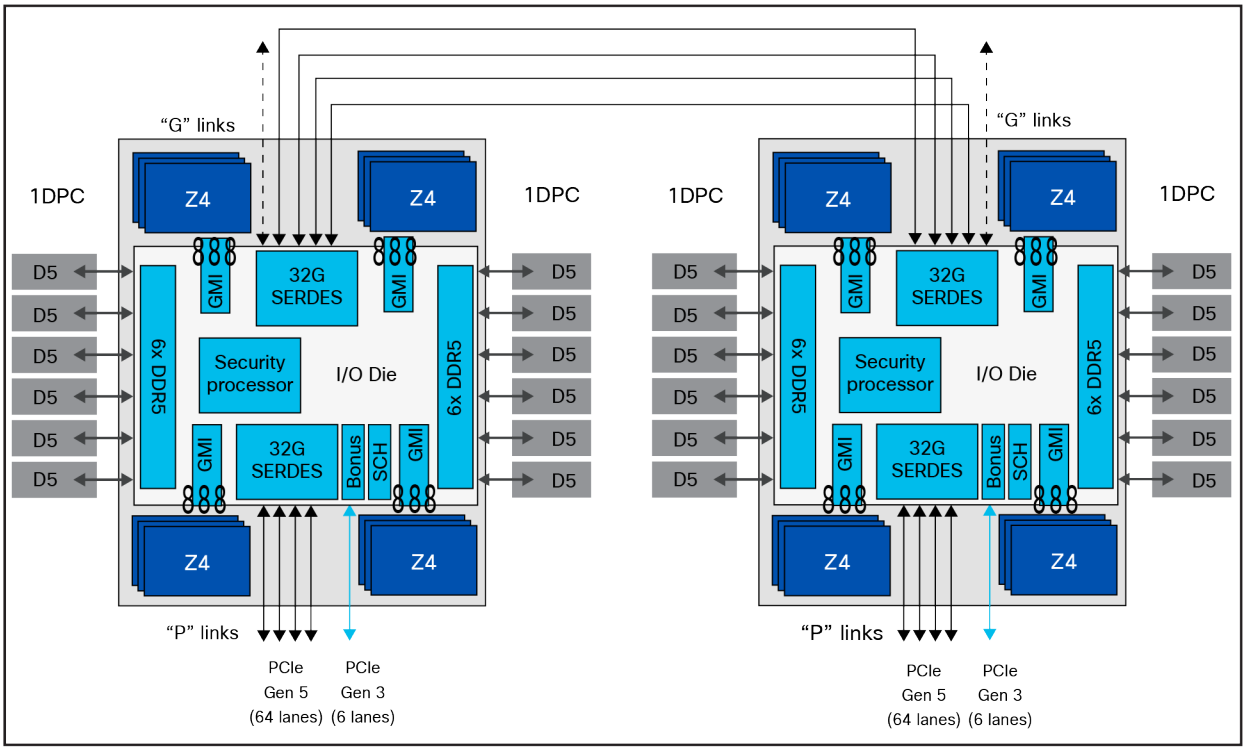

図 1.
NUMA ドメインを使用した AMD EPYC 第 4 世代プロセッサのブロック図

第 5 世代 EPYC プロセッサでは、AMD Infinity ファブリック インターコネクトに加えられた改善により、遅延の違いがさらに減少しました。最後の 1 または 2% の遅延をメモリ参照以外に収める必要があるアプリケーションの場合、EPYC 9005 シリーズ プロセッサを使用してメモリ範囲と CPU ダイ（Zen 5 または Zen 5c）の間にアフィニティを作成すると、パフォーマンスを向上できます。図 2 はこれがどのように機能するかを示しています。NPS=4 構成で I/O ダイを 4 つのクワドラントに分割すると、6 つの DIMM が 3 つのメモリ コントローラに供給され、Infinity Fabric (GMI) を介して最大 4 つの Zen 5 CPU ダイまたは最大 3 つの Zen 5c CPU ダイのセットに密接に接続されていることがわかります。

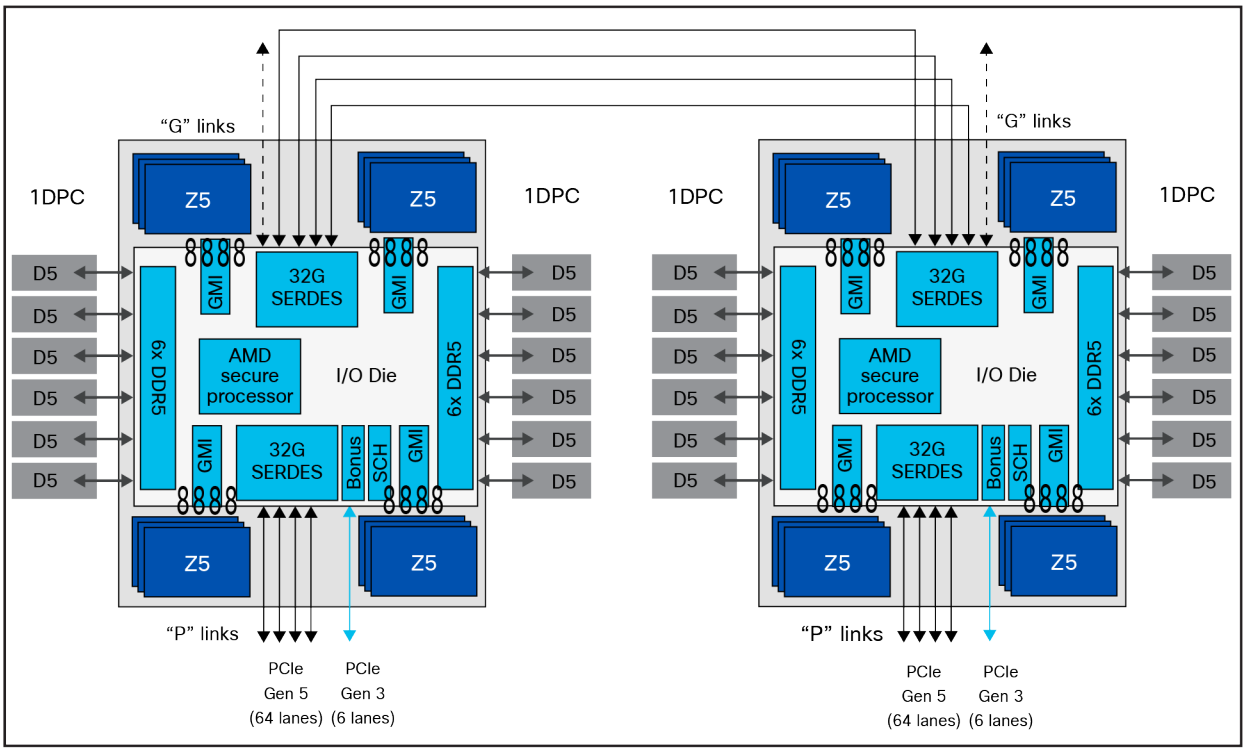

図 2.
NUMA ドメインを使用した AMD EPYC 第 5 世代プロセッサのブロック図

NPS1

NPS = 1 の設定は、ソケットごとに 1 つの NUMA ノードを示します。この設定により、プロセッサ上のすべてのメモリチャネルが単一の NUMA ノードに設定されます。すべてのプロセッサコア、すべての接続メモリ、および SoC に接続されたすべての PCIe デバイスが、その 1 つの NUMA ノードに存在します。メモリは、プロセッサ上のすべてのメモリ チャネルにわたって単一のアドレス空間にインターリーブされます。

NPS2

NPS = 2 の設定により、各プロセッサは 2 つの NUMA ドメインに構成されます。これにより、コアの半分とメモリチャネルの半分が 1 つの NUMA ドメインにグループ化され、残りのコアとメモリチャネルが 2 番目の NUMA ドメインにグループ化されます。メモリは、各 NUMA ドメイン内の 6 つのメモリチャネル間でインターリーブされます。PCIe デバイスは、そのデバイスの PCIe ルート コンプレックスを持つ半分に応じて、2 つの NUMA ノードのいずれかに対してローカルになります。

NPS4

NPS = 4 の設定では、プロセッサがソケットごとに 4 つの NUMA ノードに分割され、各論理クワドラントは独自の NUMA ドメインとして構成されます。メモリは、各クワドラントに関連付けられたメモリチャネル間でインターリーブされます。PCIe デバイスは、そのデバイスに対応する PCIe ルート コンプレックスを持つ IOD クワドラントに応じて、4 つのプロセッサ NUMA ドメインのいずれかに対してローカルになります。メモリ チャネルのすべてのペアがインターリーブされます。これは、HPC およびその他の高度に並列なワークロードに推奨されます。Windows は CPU グループのサイズを最大 64 論理コアに制限しているため、64 コアを超える AMD EPYC プロセッサに対して SMT が有効になっている Windows システムを起動する場合は、NPS4 を使用する必要があります。

注： Windows システムでは、デフォルトの NPS1 の代わりに NPS2 または NPS4 を使用して、NUMA ノードあたりの論理プロセッサ数が 64 未満であることを確認します。

NPS0（非推奨）

NPS = 0 の設定は、システム全体の単一の NUMA ドメインを示します（2 ソケット構成の両方のソケット上）。この設定により、システムのすべてのメモリチャネルが 1 つの NUMA ノードに設定されます。メモリは、システム上のすべてのメモリ チャネルにわたって単一のアドレス空間にインターリーブされます。すべてのソケットのすべてのプロセッサコア、すべての接続されたメモリ、およびいずれかのプロセッサに接続されたすべての PCIe デバイスは、その単一の NUMA ドメインにあります。

NUMA ドメインとしてのレイヤー 3 キャッシュ

NPS 設定に加えて、NUMA 設定を変更するための BIOS オプションがもう 1 つ使用できます。[NUMA としてのレイヤ 3 キャッシュ（L3CAN）] オプションを使用すると、各レイヤ 3 キャッシュ（CCD ごとに 1 つ）が独自の NUMA ノードとして公開されます。たとえば、8 つの CCD を備えたシングルプロセッサには、8 つの NUMA ノードがあり、各 CCD に 1 つずつあります。この場合、2 ソケットシステムには合計 16 の NUMA ノードが存在することになります。

プロセッサの設定

ここでは、構成可能なプロセッサ オプションについて説明します。

同時マルチスレッディング（SMT）モード

同時マルチスレッディング（SMT）オプションを設定すると、AMD SMT モード オプションをサポートするプロセッサ上の論理プロセッサ コアをイネーブルまたはディセーブルにすることができます。SMT モードが 自動（有効）に設定されている場合、各物理プロセッサ コアは 2 つの論理プロセッサ コアとして動作し、マルチスレッド ソフトウェア アプリケーションは各プロセッサ内でスレッドを並行して処理できます。

多くの HPC ワークロードを含む一部のワークロードでは、SMT が有効になっている場合、パフォーマンス ニュートラルな、またはパフォーマンス低下の結果が観察されます。物理コアだけでなく、一部のアプリケーションは、ハードウェア スレッドによって有効にライセンスされます。これらの理由から、EPYC 9004 シリーズプロセッサで SMT を無効にすることが望ましい場合があります。さらに、一部のオペレーティング システムでは、EPYC 9004 シリーズプロセッサで有効になっている x2APIC がサポートされていません。これは、255 スレッドを超えるサポートが必要です。AMD の x2APIC 実装をサポートしていないオペレーティング システムを実行し、2 つの 64 コア プロセッサがインストールされている場合は、SMT を無効にする必要があります。表 3 は設定をまとめたものです。

実際の環境で CPU ハイパースレッディング オプションを有効にした場合と無効にした場合の両方をテストしてください。単一スレッド アプリケーションを実行している場合は、ハイパースレッディングを無効にすべきです。

表 3　　SMT 設定

設定	オプション
SMT 制御	• 自動：コアごとに 2 つのハードウェア スレッドを使用します。 • 無効：コアごとに単一のハードウェア スレッドを使用します。 • 有効：コアごとにダブル ハードウェア スレッドを使用します。

セキュア仮想マシン（SVM）モード

セキュア仮想マシン (SVM) モードでは、プロセッサ仮想化機能が有効になり、プラットフォームは複数のオペレーティング システムとアプリケーションを独立したパーティション内で実行できます。AMD SVM モードは、次のいずれかの値に設定できます。

- 無効：プロセッサでの仮想化を禁止します。
- 有効：プロセッサで、複数のオペレーティング システムをそれぞれ独立したパーティション内で実行できます。

アプリケーションのシナリオに仮想化が不要な場合は、AMD 仮想化テクノロジを無効にします。仮想化を無効にした後、AMD IOMMU オプションも無効にします。これにより、メモリ アクセスの遅延に違いが生じる可能性があります。表 4 は設定をまとめたものです。

表 4　仮想化オプションの設定

設定	オプション
SVM	• 有効 • 無効

DF C-states

CPU コアと同様に、AMD Infinity ファブリックは、アイドル中に低電力状態に移行する可能性があります。ただし、完全電力モードに戻すときに遅延が発生し、遅延ジッターが発生します。低遅延のワークロードまたはバースト I/O が発生するワークロードでは、データファブリック（DF）の C ステート機能を無効にして、より高い電力消費でパフォーマンスを向上させることができます。表 5 は設定をまとめたものです。

表 5　DF C-states

設定	オプション
DF C-states	• 自動/有効：AMD Infinityファブリックが低電力状態になることができます • 無効：AMD Infinityファブリックが低電力状態になるのを防ぎます

NUMA ドメインとしての ACPI SRAT L3 キャッシュ

[NUMA ドメインとしての ACPI SRAT L3 キャッシュ（ACPI SRAT L3 Cache as NUMA Domain）] 設定が有効になっている場合、各レイヤ 3 キャッシュは NUMA ノードとして公開されます。[Layer 3 Cache as NUMA Domain (L3CAN)] 設定を使用すると、各レイヤ 3 キャッシュ（CCD ごとに 1 つ）が独自の NUMA ノードとして公開されます。たとえば、8 つの CCD を備えたシングルプロセッサには、8 つの NUMA ノードがあり、各 CCD に 1 つずつあります。デュアルプロセッサシステムには合計 16 の NUMA ノードがあります。

この設定により、ワークロードまたはワークロードのコンポーネントを CCX 内のコアにピン留めでき、レイヤ 3 キャッシュを共有することでメリットが得られる場合、高度に NUMA 最適化されたワークロードのパフォーマンスを向上させることができます。この設定を無効にすると、NUMA ドメインは NUMA NPS パラメータ設定に従って識別されます。

一部のオペレーティングシステムとハイパーバイザはレイヤ 3 対応のスケジューリングを実行せず、一部のワークロードはレイヤ 3 を NUMA ドメインとして宣言することでメリットを享受できます。表 6 は設定をまとめたものです。

表 6　NUMA ドメイン設定としての ACPI SRAT レイヤ 3 キャッシュ

設定	オプション
NUMAドメインとしてのACPI SRAT L3キャッシュ	• 自動（無効） • 無効：各レイヤ 3 キャッシュを NUMA ドメインとして OS に報告しません。 • 有効：各レイヤ 3 キャッシュを NUMA ドメインとして OS に報告

アルゴリズム パフォーマンス ブーストの無効化（APBDIS）

SMU の APBDIS（アルゴリズム パフォーマンス ブースト （APB）無効化）値を選択できます。デフォルト状態では、AMD Infinityファブリックは、ファブリックとメモリの使用に基づいて、フル電力および低電力のファブリック クロックとメモリ クロックを選択します。ただし、低帯域幅で遅延の影響を受けやすいトラフィック（およびメモリ遅延チェッカー）が含まれる特定のシナリオでは、低電力から全電力への移行が遅延に悪影響を与える可能性があります。APBDIS を 1（アルゴリズム パフォーマンス ブースト（APB）を無効にする）に設定し、固定の Infinity Fabric の P 状態を 0 に指定すると、Infinity ファブリックとメモリ コントローラが強制的にフルパワー モードになり、そのような遅延ジッターが排除されます。特定の CPU プロセッサとメモリ装着オプションを使用すると、固定の Infinity Fabric P-state を 1 に設定することで、メモリ帯域幅を犠牲にしてメモリ レイテンシを削減できます。この設定は、メモリ遅延の影響を受けやすいことが知られているアプリケーションに利点をもたらす場合があります。表 7 は設定をまとめたものです。

表 7　APBDIS 設定

設定	オプション
APBDIS	• 自動（0）：SMU の自動 APBDIS を設定します。これはデフォルトです。制限があります。 • 0：リンクの使用に基づいて Infinity Fabric P-state を動的に切り替えます。 • 1：固定 Infinity Fabric P-state 制御を有効にします。

固定 SOC P-State SP5F 19h

ACPI _PSD オブジェクトによって報告されるように、P-state を強制的に独立/従属にします。APBDIS が有効になっている場合、SOC P-State が変更されます。ここで、F はプロセッサ ファミリを指します。

設定	オプション
固定 SOC P-State SP5F 19h	• Auto • P0: 最高のパフォーマンスを発揮する Infinity Fabric P-stateです • P1：次にパフォーマンスの高い Infinity Fabric P-stateです • P2：P1 の後に次に高いパフォーマンスの Infinity Fabric P-state です

xGMI 設定: ソケット間の接続

2 ソケット システムでは、プロセッサはソケット間 xGMI リンクを介して相互接続されます。このリンクは、SoC のすべてのコンポーネントを接続する無限ファブリックの一部です。

NUMA 非認識ワークロードは、広範なクロスソケット通信のために最大の xGMI 帯域幅を必要とする場合があります。NUMA 対応ワークロードは、多くのクロスソケットトラフィックがなく、増加した CPU ブーストを使用することを避けるため、xGMI 電力を最小限に抑えることが必要になる場合があります。xGMI レーン幅を x16 から x8 または x2 に減らすことができます。または、電力消費が高すぎる場合は xGMI リンクを無効にすることができます。

xGMI リンク構成と 4 リンク xGMI 最大速度（Cisco xGMI 最大速度）

xGMI リンクの数と、xGMI リンクの最大速度を設定できます。この値を低い速度に設定すると、非コア電力を節約できます。これを使用して、コア周波数を上げたり、全体の電力を削減したりできます。また、クロスソケット帯域幅が減少し、クロスソケット遅延が増加します。Cisco UCS C245 M8 ラックサーバーは、最大速度 32 Gbps で 4 つの xGMI リンクをサポートしています。

Cisco xGMI 最大速度設定により、xGMI リンク設定と 4-Link/3-Link xGMI 最大速度を設定できます。Cisco xGMI 最大速度を有効にすると、xGMI リンク設定が 4 に設定され、4-リンクの xGMI 最大速度は 32 Gbps になります。Cisco xGMIの最大速度設定を無効にすると、デフォルト値が適用されます。

表 8 は設定をまとめたものです。

表 8 xGMI リンクの設定

設定	オプション
Cisco xGMIの最大速度	• [Disabled]（デフォルト） • Enabled
[xGMI リンク構成（xGMI Link Configuration）]	• Auto • 1 • 2 • 3 • 4
4 リンク xGMI 最大速度	• 自動（25 Gbps） • 20 Gbps • 25 Gbps • 32 Gbps
3 リンク xGMI 最大速度	• 自動（25 Gbps） • 20 Gbps • 25 Gbps • 32 Gbps

注： この BIOS 機能は、2 ソケット構成の Cisco UCS X215c M8 コンピューティング ノードおよび Cisco UCS C245 M8 ラック サーバーにのみ適用されます。

CPU 性能強化

この BIOS オプションは、拡張 CPU パフォーマンス設定を変更するのに役立ちます。有効にすると、このオプションによりプロセッサの設定が調整され、プロセッサが積極的に動作するようになります。これにより、CPU 全体のパフォーマンスが向上しますが、消費電力が増加する可能性があります。この BIOS オプションの値は [自動] または [無効] です。デフォルトでは、CPU パフォーマンス拡張オプションは無効になっています。

注： この BIOS 機能は、 Cisco UCS X215c M8 コンピューティング ノードおよび Cisco UCS C245 M8 ラック サーバーにのみ適用されます。このオプションを有効にすると、ファン ポリシーを最大電力に設定することを強く推奨します。

デフォルトでは、この BIOS 設定は [無効（Disabled）] になっています。

メモリの設定

このセクションで説明されているメモリ設定を構成できます。

ソケットごとの NUMA ノード（NPS）

この設定により、1 ソケットあたりの NUMA ノード（NPS）の数を指定でき、NUMA 対応または高度に並列化可能なワークロードのローカル メモリ遅延を削減し、NUMA に適していないワークロードのコアあたりのメモリ帯域幅を増やす間のトレードオフを有効にすることができます。ソケット インターリーブ（NPS0）は、2 つのソケットを 1 つの NUMA ノードに一緒にインターリーブしようとします。第 4 世代 AMD EPYC プロセッサは、プロセッサの内部 NUMA トポロジに応じて、さまざまな数の NUMA NPS 値をサポートします。特定のプロセッサまたは特定のメモリ構成では、NPS2 および NPS4 がオプションになっていない場合があります。

1 ソケット サーバーでは、ソケットあたりの NUMA ノードの数は 1、2、または 4 ですが、すべてのプロセッサがすべての値をサポートしているわけではありません。高度に NUMA 最適化されたアプリケーションのパフォーマンスは、ソケットあたりの NUMA ノード数をサポートされる値を 1 より大きく設定することで向上できます。

ほとんどのワークロードでは、デフォルト設定（ソケットごとに 1 つの NUMA ドメイン）が推奨されます。NPS4 は、ハイパフォーマンス コンピューティング（HPC）およびその他の高度なパラレル ワークロードに推奨されます。200 Gbps ネットワーク アダプタを使用する場合、ネットワーク インターフェイス カード（NIC）のメモリ遅延とメモリ帯域幅との間のバランスを取るために NPS2 が推奨される場合があります。この設定は、NUMA ドメイン設定としての Advanced Configuration and Power Interface（ACPI）Static Resource Affinity Table（SRAT）レイヤ 3（L3）キャッシュとは無関係です。NUMA ドメインとしての ACPI SRAT L3 キャッシュが有効になっている場合、この設定により、メモリのインターリーブの粒度が決まります。NPS1 では、8 つのメモリ チャネルすべてがインターリーブされます。NPS2 では、4 つのチャネルごとに相互にインターリーブされます。NPS4 では、チャネルのすべてのペアがインターリーブされます。表 9 は設定をまとめたものです。

表 9　　NUMA NPS 設定

設定	オプション
ソケットごとのNUMAノード	• 自動（NPS1） • NPS0：両方のソケットのすべてのチャネルでメモリアクセスをインターリーブします（非推奨）。 • NPS1：各ソケットの 8 つのチャネルすべてでメモリアクセスをインターリーブ。ソケットごとに 1 つの NUMA ノードを報告します（NUMA が有効になっているための L3 キャッシュを除く）。 • NPS2：各ソケットの 4 チャネル（ABCD と EFGH）のグループ間でメモリアクセスをインターリーブします。ソケットごとに 2 つの NUMA ノードを報告します（NUMA が有効になっているための L3 キャッシュの場合を除く）。 • NPS4：各ソケットのチャネルのペア（AB、CD、EF、および GH）間でメモリアクセスをインターリーブする。ソケットごとに 4 つの NUMA ノードを報告します（NUMA が有効になっているため L3 キャッシュを除く）。

I/O メモリ管理ユニット（IOMMU）

I/O メモリ管理ユニット（IOMMU）にはいくつかの利点があり、x2 プログラマブル割り込みコントローラ（x2APIC）を使用する場合に必要です。IOMMU を有効にすると、デバイス（EPYC 統合 SATA コントローラなど）は、サブシステムに対して 1 つの IRQ ではなく、接続されているデバイスごとに個別の割り込み要求（IRQ）を提示できます。また、IOMMU により、オペレーティング システムはダイレクト メモリ アクセス（DMA）対応の I/O デバイスに追加の保護を提供できます。IOMMU は、周辺機器からの割り込みのフィルタリングと再マッピングにも役立ちます。表 10 は設定をまとめたものです。

表 10 IOMMU 設定

設定	オプション
IOMMU	• 自動（有効） • 無効：IOMMU サポートを無効にします • 有効：IOMMU サポートを有効にします

Memory interleaving

メモリインターリーブは、アプリケーションで使用可能なメモリ帯域幅を増やすために CPU が使用する手法です。インターリーブを行わないと、連続するメモリ ブロック（多くの場合、キャッシュ ライン）が同じメモリ バンクから読み取られます。したがって、連続したメモリを読み取るソフトウェアは、次のメモリアクセスを開始する前に、メモリ転送処理が完了するのを待つ必要があります。メモリ インターリーブをイネーブルにすると、連続するメモリ ブロックが異なるバンクに存在するため、それらのすべてがプログラムが達成できる全体的なメモリ帯域幅に貢献できます。

AMD では、CPU ソケットあたり 8 つのメモリ チャネルすべてに、同じ容量のすべてのチャネルを装着することを推奨しています。この方法を使用すると、メモリ サブシステムを 8 方向インターリーブ モードで動作させることができ、ほとんどの場合に最高のパフォーマンスが得られます。表 11 は設定をまとめたものです。

表 11 メモリ インターリーブ設定

設定	オプション
AMD メモリ インターリーブ	• 自動：サポートされているメモリ DIMM 設定でインターリーブが有効になります。 • 無効：インターリーブは実行されません。

電力設定

このセクションで説明されている電力状態の設定を構成できます。

コア パフォーマンス ブースト

コア パフォーマンス ブースト機能により、プロセッサは、電力の可用性、温度ヘッドルーム、およびシステム内のアクティブ コアの数に基づいて、CPU の基本周波数よりも高い周波数に移行できます。コア パフォーマンス ブーストは、プロセッサ コアの周波数遷移によるジッターを引き起こす可能性があります。

一部のワークロードでは、許容可能なレベルのパフォーマンスを達成するために最大コア周波数で実行できる必要はありません。より良い電力効率を得るために、最大コア ブースト周波数を設定できます。この設定では、固定周波数を設定できません。最大ブースト周波数を制限するだけです。実際のブースト パフォーマンスは、本書で言及されている多くの要因およびその他の設定に依存します。表 12 は設定をまとめたものです。

表 12 コア パフォーマンス ブースト設定

設定	オプション
コア パフォーマンス ブースト	• 自動（有効）：プロセッサが CPU の基本周波数よりも高い周波数（ターボ周波数）に遷移できるようにします。 • 無効：CPU コアブースト周波数を無効にします。

グローバル C-State 制御

C 状態とは、プロセッサの CPU コアの非アクティブな電力状態を指します。C0 は、命令が処理される動作状態であり、より大きい番号の C-state（C1、C2 など）は、コアがアイドルである低電力状態です。グローバル C-state

設定を使用して、サーバ上で C-state を有効または無効にすることができます。デフォルトでは、グローバル C-state 制御は「自動」に設定されています。これにより、コアがより低い電力状態になることができます。これにより、プロセッサ コアの周波数遷移によるジッターが発生する可能性があります。この設定を無効にすると、CPU コアは C0 および C1 状態で動作します。表 13 は設定をまとめたものです。

C-state は ACPI オブジェクトを通じて公開され、ソフトウェアによって動的に要求できます。ソフトウェアは、HALT 命令を実行するか、特定の I/O アドレスから読み取ることで、C-state の変更を要求できます。低電力 C-state に入った時点でプロセッサにより実行されるアクションをソフトウェアで設定することもできます。第 4 世代 AMD EPYC プロセッサのコアは、最大 3 つの AMD 指定の C-state（I/O ベースの C0、C1、および C2）をサポートするように設計されています。

表 13　　グローバル C-state の設定

設定	オプション
グローバル C-state 制御	• 自動（有効）：I/O ベースの C-state を有効にします。 • 無効：I/O ベースの C-state を無効にします。

レイヤ 1 およびレイヤ 2 ストリーム ハードウェア プリフェッチャ

ほとんどのワークロードは、データを収集し、コアパイプラインをビジー状態に保つためのレイヤ 1 およびレイヤ 2 ストリーム ハードウェア プリフェッチャ（L1 Stream HW Prefetcher および L2 Stream HW Prefetcher）の使用によるメリットを受けます。ただし、一部のワークロードは本質的に非常にランダムであり、実際には、プリフェッチャの一方または両方が無効になっている場合に全体的なパフォーマンスが向上します。デフォルトでは、両方のプリフェッチャが有効になっています。表 14 は設定をまとめたものです。

表 14　　レイヤ 1 およびレイヤ 2 ストリーム ハードウェア プリフェッチャ設定

設定	オプション
L1 Stream HW Prefetcher	• 自動 （有効） • 無効：プリフェッチャを無効にします。 • 有効：プリフェッチャを有効にします。
L2 Stream HW Prefetcher	• 自動 （有効） • 無効：プリフェッチャを無効にします。 • 有効：プリフェッチャを有効にします。

デタミニズム スライダ（Determinism Slider）

デタミニズム スライダを使用すると、サーバーをパフォーマンス設定に設定してデータセンター内の同一構成のシステム間で均一なパフォーマンスを選択することも、サーバーを電力設定に設定してデータセンター全体でパフォーマンスを変えながら個々のシステムの最大パフォーマンスを選択することもできます. デタミニズム スライダが [パフォーマンス（Performance）] に設定されている場合は、設定可能な熱設計電力（cTDP）とパッケージ電力制限（PPL）が同じ値に設定されていることを確認してください。ほとんどのプロセッサのデフォルト（自動）設定はパフォーマンス決定モードであり、プロセッサは一貫したパフォーマンスでより低い電力レベルで動作できます。最大のパフォーマンスを実現するには、デタミニズム スライダを [電力（Power）] に設定します。表 15 は設定をまとめたものです。

表 15　　デタミニズム スライダ設定

設定	オプション
デタミニズム スライダ（Determinism Slider）	• 自動：この設定は [パフォーマンス（Performance）] オプションと同じです。 • 電力：同じように構成された多数の CPU の中で、各 CPU に最大のパフォーマンス レベルを保証しますが、同じ cTDP に到達した場合にのみ CPU をスロットリングします。 • パフォーマンス：一部の CPU をより低い電力レベルで動作するようにスロットルすることにより、同じように構成された多数の CPU で一貫したパフォーマンスレベルを保証します。

CPPC：コラボレーティブ プロセッサ パフォーマンス制御

コラボレーション プロセッサ パフォーマンス制御（CPPC）は、オペレーティング システムとハードウェア間のパフォーマンスを通信するモードとして ACPI 5.0 で導入されました。このモードを使用すると、エネルギー効率を維持するために、ターボ ブーストをいつ、どのくらい適用できるかを OS が制御できるようになります。すべてのオペレーティングシステムが CPPC をサポートしているわけではありませんが、Microsoft Windows 2016 以降でサポートが始まっています。表 16 は設定をまとめたものです。

表 16　　CPPC 設定

設定	オプション
CPPC	• 自動 • 無効：無効 • 有効：OS が ACPI CPPC を使用してパフォーマンスと電力最適化の要求を実行できるようにします。

電力プロファイル選択 F19h

プロファイル ポリシーでの DF P 状態の選択は、P 状態範囲、BIOS オプション、または APB_DIS BIOS オプションによってオーバーライドされます。この場合、F はプロセッサ ファミリ、M はモデルを指します。

設定	オプション
電力プロファイル選択 F19h	• 効率モード • ハイパフォーマンス モード • 最大 I/O パフォーマンス モード • バランス型メモリ パフォーマンス モード • バランス型コア パフォーマンス モード • バランス型コア メモリ パフォーマンス モード • Auto

ファン制御ポリシー

ファン ポリシーを使ってファンの速度を制御することにより、サーバーの消費電力を削減し、ノイズ レベルを下げることができます。ファン ポリシーが使用される前は、いずれかのサーバー コンポーネントの温度が設定済みしきい値を超過した場合に、ファン速度が自動的に増加しました。ファン速度を低く抑えるために、通常、コンポーネントのしきい値温度を高い値に設定しました。この動作はほとんどのサーバー構成に最適でしたが、次のような状況に対処できませんでした。

- 最大 CPU パフォーマンス：ハイパフォーマンスを得るには、いくつかの CPU を設定済みしきい値よりもかなり低い温度に冷却する必要があります。この冷却には、非常に高速なファン速度を必要とし、結果として電力消費とノイズ レベルが増大します。
- 低電力消費：電力消費を最も低く抑えるにはファンを非常に遅くする必要があり、場合によっては、この動作を許可するためにファン停止をサポートするサーバーで完全に停止する必要があります。ただし、ファンの速度を遅くすると、サーバーが過熱する可能性があります。この状況を回避するには、可能な最低速度よりもやや速くファンを作動させる必要があります。

次のファン ポリシーの中から選択できます。

- バランス型：これがデフォルト ポリシーです。この設定でほとんどのサーバー構成を冷却できますが、容易に加熱する PCIe カードを含むサーバーには適さない可能性があります。
- 低電力：この設定は、PCIe カードを含まない最小構成のサーバーに最適です。
- 高電力：この設定は、60 % ～ 85 % のファン速度を必要とするサーバー構成で使用できます。このポリシーは、容易に過熱して高温になる PCIe カードを含むサーバーに最適です。このポリシーで設定される最小ファン速度はサーバ プラットフォームごとに異なりますが、およそ 60 ～ 85 % の範囲内です。
- 最大電力：この設定は、70 ～ 100 % の範囲の非常に高いファン速度を必要とするサーバ構成に使用できます。このポリシーは、容易に過熱して極端に高温になる PCIe カードを含むサーバーに最適です。このポリシーで設定される最小ファン速度はサーバ プラットフォームごとに異なりますが、およそ 70 ～ 100 % の範囲内です。
- [音響（Acoustic）]：ファンの速度を遅くすることで、大きな音が問題になる環境でのノイズレベルを低減他のモードのように、電力消費を調整して、コンポーネントのスロットリングを防止するものでありません。[音響 (Acoustic)] オプションを使用すると、短時間のスロットリングが発生しますが、ノイズレベルも低くなります。このファン制御ポリシーを適用すると、短時間一時的なパフォーマンスの影響が発生する可能性があります。

注： このポリシーは、Cisco Integrated Management Controller (IMC) コンソールおよび Cisco IMC スーパーバイザを使用して、スタンドアロン Cisco UCS C シリーズ M8 サーバー用に設定できます。Cisco IMC Web コンソールから、[コンピューティング（Compute）] > [電力ポリシー（Power Policies）] > [構成済みのファン ポリシー（Configured Fan Policy）] > [ファン ポリシー（Fan Policy）] の順に選択します。

Cisco Intersight®の管理対象 C シリーズ M8 サーバーの場合、このポリシーはファン ポリシーを使用して設定できます。

Cisco UCS X215c M8 コンピューティングノード、Cisco UCS C245 M8 ラックサーバー、および Cisco UCS C225 M8 ラックサーバーの BIOS 設定

表 17 に、AMD EPYC 第 4 世代および第 5 世代プロセッサ ファミリを搭載した Cisco UCS M8 サーバーの BIOS トークン名、デフォルト、およびサポートされる値を示します。

表 17 BIOS トークンの名前と値

BIOS トークン名	[デフォルト値（Default value）]	サポートされる値
プロセッサ		
SMT モード	自動（有効）	自動、有効、無効
SVM モード	イネーブル	Enabled、Disabled
DF C-states	自動（有効）	自動、有効、無効
NUMA ドメインとしての ACPI SRAT L3 キャッシュ	自動（無効）	自動、有効、無効
APBDIS	自動（0）	自動、0、1
固定 SOC P-State SP5F 19h	自動（P0）	自動、P0、P1、P2
4 リンク xGMI 最大速度*	自動（32 Gbps）	自動、20Gbps、25Gbps、32Gbps
強力な CPU パフォーマンス*	無効化	自動、無効
メモリ		
ソケットごとの **NUMA** ノード	自動（NPS1）	自動、NPS0、NPS1、NPS2、NPS4
IOMMU	自動（有効）	自動、有効、無効
Memory interleaving	自動（有効）	自動、有効、無効
電力/パフォーマンス		
コア パフォーマンス ブースト	自動（有効）	自動、無効
グローバル C-State 制御	自動（有効）	自動、有効、無効
L1 Stream HW Prefetcher	自動（有効）	自動、有効、無効
L2 Stream HW Prefetcher	自動（有効）	自動、有効、無効
デタミニズム スライダ（Determinism Slider）	自動（電力）	Auto, Power, Performance
CPPC	自動（無効）	自動、無効、有効

BIOS トークン名	[デフォルト値（Default value）]	サポートされる値
電力プロファイル選択 F19h	ハイパフォーマンス モード	バランス型メモリ パフォーマンス モード、効率モード、ハイパフォーマンス モード、最大 I/O パフォーマンス モード、バランス型コア パフォーマンス モード、バランス型コア メモリ パフォーマンス モード

さまざまな汎用ワークロードに関する BIOS の推奨事項

このセクションでは、汎用ワークロードを最適化するために推奨される BIOS 設定の概要を説明します。

- 計算集約型
- I/O 集約型
- エネルギー効率
- 低遅延

次の節句ションでは、各ワークロードについて説明します。

CPU 集約型ワークロード

CPU 集約型のワークロードの場合、単一のジョブの作業を複数の CPU に分散して、処理時間をできるだけ減らすことが目標です。これを行うには、ジョブの一部を並行して実行する必要があります。各プロセス（スレッド）は作業の一部を処理し、同時に計算を実行します。通常、CPU は情報を迅速に交換する必要があり、特殊な通信ハードウェアが必要です。

通常、CPU 集約型のワークロードは、常に個々のコアの最大ターボ周波数を実現するプロセッサまたはメモリの恩恵を受けます。プロセッサの電力管理設定を適用して、コンポーネントの周波数の増加を簡単に実現できるようにします。CPU 集約型ワークロードは汎用ワークロードであるため、プロセッサ コアとメモリの速度を向上させるために汎用的な最適化が実行され、通常、計算時間の短縮によるメリットを得られるパフォーマンスの調整が使用されます。

I/O 集約型ワークロード

I/O 集約型の最適化は、I/O とメモリ間の最大スループットに依存する設定です。I/O とメモリ間のリンクのパフォーマンスに影響を与えるプロセッサ使用率ベースの電力管理機能は無効になっています。

エネルギー効率の高いワークロード

エネルギー効率の最適化は、最も一般的なバランスの取れたパフォーマンスの設定です。これにより、ほとんどのアプリケーション ワークロードにメリットを提供する一方で、全体的なパフォーマンスにほとんど影響を与えない電力管理設定が可能になります。エネルギー効率の高いワークロードに適用される設定では、電力効率ではなく、アプリケーションの全般的なパフォーマンスが向上します。仮想化オペレーティング システムを使用する場合、プロセッサの電力管理設定がパフォーマンスに影響することがあります。したがって、これらの設定は、通常、ワークロードのために BIOS を調整しないお客様に推奨されます。

低遅延のワークロード

金融取引やリアルタイム処理などの低遅延を必要とするワークロードでは、サーバーが一貫したシステム応答を提供する必要があります。低遅延ワークロードは、ワークロードの計算遅延を最小限に抑える必要があるお客様向けです。最大の速度とスループットは、多くの場合、全体的な計算遅延を低減するために犠牲になります。計算の遅延を引き起こす可能性のあるプロセッサの電力管理およびその他の管理機能は無効にしてください。

低遅延を実現するには、テスト対象のシステムのハードウェア構成を理解する必要があります。応答時間に影響する重要な要因には、コアの数、コアあたりの処理スレッド、NUMA ノードの数、NUMA トポロジの CPU とメモリの構成、NUMA ノードのキャッシュトポロジが含まれます。通常、BIOS オプションは OS に依存せず、確定的なパフォーマンスを実現するには、適切に調整された低遅延オペレーティング システムも必要です。

汎用ワークロード向けに最適化された BIOS 設定の概要

表 18 は、汎用ワークロード向けに最適化された BIOS 設定の概要です。

表 18　CPU 集約型、I/O 集約型、エネルギー効率の高い、低遅延のワークロードに対する BIOS の推奨事項

BIOS オプション	BIOS 値（プラットフォームのデフォルト）	CPU 集約型	I/O 高負荷	エネルギー効率	低遅延
プロセッサ					
SMT モード	自動（有効）	自動	自動	自動	無効化
SVM モード	有効	有効	有効	有効	無効
DF C-states	自動（有効）	Auto	無効	自動	無効化
NUMA ドメインとしての ACPI SRAT L3 キャッシュ	自動（無効）	Enabled	自動	自動	自動
APBDIS	自動（0）	1	1	自動	自動
固定 SOC P-State SP5F 19h	自動（P0）	自動	自動	P2	Auto
4リンクxGMI最大速度	自動（32 Gbps）	自動	自動	自動	自動
CPU 性能強化	無効	自動	無効	無効	無効
メモリ					
ソケットごとの **NUMA ノード**	自動（NPS1）	NPS4	NPS4	自動	自動
IOMMU	自動（有効）	自動*	自動	自動	無効*
Memory interleaving	自動（有効）	自動*	自動	自動	無効*

BIOS オプション	BIOS 値（プラットフォームのデフォルト）	CPU 集約型	I/O 高負荷	エネルギー効率	低遅延
電力/パフォーマンス					
コア パフォーマンス ブースト	自動（有効）	自動	自動	自動	無効化
グローバル C-State 制御	自動（有効）	自動	自動	自動	自動
L1 Stream HW Prefetcher	自動（有効）	自動	自動	無効	自動
L2 Stream HW Prefetcher	自動（有効）	自動	自動	無効	自動
デタミニズム スライダ（Determinism Slider）	自動（電力）	自動	自動	自動	パフォーマンス
CPPC	自動（無効）	自動	自動（Auto）	有効	Auto
電力プロファイル選択 F19h	ハイパフォーマンス モード	ハイパフォーマンス モード	最大 I/O パフォーマンス モード	効率モード	ハイパフォーマンス モード

注： * で強調表示されている BIOS トークンは、Cisco UCS X215c M8 コンピューティング ノードおよび Cisco UCS C245 M8 ラック サーバーにのみ適用されます。

- アプリケーションのシナリオに仮想化が不要な場合は、AMD 仮想化テクノロジを無効にします。仮想化を無効にした状態で、AMD IOMMU オプションも無効にします。これにより、メモリ アクセスの遅延に違いが生じる可能性があります。詳細については、「AMD パフォーマンス チューニング ガイド」を参照してください。

エンタープライズ ワークロードに関する追加の BIOS 推奨事項

このセクションでは、エンタープライズ ワークロードに最適な BIOS 設定の概要を示します。

- 仮想化
- コンテナ
- リレーショナル データベース（RDBMS）
- 分析データベース（Bigdata）
- HPC ワークロード

以下のセクションでは、各エンタープライズ ワークロードについて説明します。

仮想ワークロード

AMD 仮想化テクノロジーは、ソフトウェアベースの仮想化ソリューションを使用する IT 環境での管理性、セキュリティ、および柔軟性を提供します。このテクノロジーを使用すると、1 つのサーバーをパーティション化して複数の独立したサーバーとしてプロジェクトでき、サーバーはオペレーティングシステムで異なるアプリケーションを同時

に実行できます。仮想化ワークロードをサポートするには、BIOS で AMD 仮想テクノロジーを有効にすることが重要です。

ハードウェア仮想化をサポートする CPU を使用すると、プロセッサは仮想マシンで複数のオペレーティング システムを実行できます。仮想オペレーティング システムのパフォーマンスはネイティブ OS のパフォーマンスよりも比較的遅いため、この機能にはある程度のオーバーヘッドが伴います。

詳細については、AMD の 『VMware vSphere Tuning Guide』を参照してください。

コンテナワークロード

アプリケーション プラットフォームと関連する依存関係をコンテナ化することで、基盤となるインフラストラクチャと OS の違いが抽象化され、効率性が向上します。各コンテナは、すべての依存関係を持つアプリケーション、ライブラリ、その他のバイナリ、およびそのアプリケーションの実行に必要な構成ファイルを含む、ランタイム環境全体を含む 1 つのパッケージにバンドルされています。実稼働環境でアプリケーションを実行するコンテナは、一貫した稼働時間を確保するための管理が必要です。コンテナがダウンした場合、別のコンテナが自動的に起動する必要があります。

ベア メタルで適切にスケーリングされて実行されるワークロードは、パフォーマンス オーバーヘッドが最小限のコンテナ環境で同様のスケーリング曲線を示します。一部のコンテナ化されたワークロードでは、ベアメタルと比較して 0% に近いパフォーマンスのバリアンスが見られることもあります。通常、オーバーヘッドが大きいと、アプリケーション設定やコンテナ構成が最適に設定されていないことを意味します。これらのトピックは、このチューニング ガイドの範囲を超えています。ただし、Kubernetes または他のコンテナ オーケストレーション プラットフォーム スケジューラの CPU ロードバランシング動作は、ベアメタル環境とは異なる方法で、コンテナ化されたアプリケーションの割り当てまたはロードバランシングを行う場合があります。

詳細については、AMD の『Kubernetes Container Tuning Guide』を参照してください。

リレーショナル データベース ワークロード

Oracle、MySQL、PostgreSQL、Microsoft SQL Server などの RDBMS を AMD EPYC プロセッサと統合すると、特に高い同時実行性、迅速なクエリ処理、および効率的なリソース使用率を必要とする環境でデータベースのパフォーマンスを向上させることができます。AMD EPYC プロセッサのアーキテクチャにより、データベースは複数のコアとスレッドを効果的に活用できます。これは、トランザクション ワークロード、分析、および大規模なデータ処理に特に役立ちます。

要約すると、RDBMS 環境で AMD EPYC プロセッサを使用すると、パフォーマンス、スケーラビリティ、およびコスト効率が大幅に向上する可能性があり、エンタープライズ データベース ソリューションにとって強力な選択肢になります。

第 4 世代 AMD EPYC プロセッサは、すべてのデータベースで高い入出力操作/秒（IOPS）とスループットを提供します。データベース アプリケーションの最適なパフォーマンスを実現するには、適切な CPU を選択することが重要です。

詳細については、AMD の RDBMS チューニングガイド を参照してください。

ビッグ データ分析ワーックロード

ビッグ データ分析では、より良い意思決定を行うために使用できる、隠れたパターン、相関関係、およびその他のインサイトを明らかにするための膨大な量のデータの調査が行われます。これには、大量の計算能力、メモリ容量、および I/O 帯域幅が必要です（AMD EPYC プロセッサが優れている領域）。

AMD EPYC プロセッサは、ビッグデータ分析のための堅牢なプラットフォームを提供し、大規模なデータ処理の要求に対応するために必要な計算能力、メモリ容量、I/O 帯域幅を提供します。拡張性、コスト効率、エネルギー効率

が高いため、ビッグデータ分析インフラストラクチャの構築やアップグレードを試みている組織にとっては魅力的な選択肢となります。

HPC（ハイパフォーマンス コンピューティング）ワークロード

HPC とは、接続され、並行して動作する複数の個別のノードを使用するクラスタベースのコンピューティングを指し、それにより、1 つのシステムでの実行に大幅な時間がかかるような大規模なデータ セットの処理に必要な時間を削減します。HPC ワークロードは計算集約型であり、通常はネットワーク I/O 集約型です。HPC ワークロードでは、メッセージ パッシング インターフェイス（MPI）接続のために、高品質の CPU コンポーネントと高速で低遅延のネットワーク ファブリックが必要です。

コンピューティング クラスタには、 クラスタの管理、導入、モニタリング、および管理を行うための単一のポイントを提供するヘッド ノードが含まれています。クラスタには、スケジューラと呼ばれる内部ワークロード管理コンポーネントもあり、これは着信するすべての作業項目（ジョブと呼ばれます）を管理します。通常、HPC ワークロードには、拡張可能にするために、ノンブロッキング MPI ネットワークを持つ多数のノードが必要です。ノードのスケーラビリティは、クラスタの使用可能なパフォーマンスを判断する上で最も重要な唯一の要因です。

HPC には高帯域幅の I/O ネットワークが必要です。Direct Cache Access（DCA）のサポートを有効にすると、ネットワーク パケットはメイン メモリではなく、レイヤ 3 プロセッサ キャッシュに直接送られます。このアプローチにより、特定のイーサネットアダプタが使用されている場合に HPC ワークロードによって生成される HPC I/O サイクルの数が削減され、システムパフォーマンスが向上します。

詳細については、AMD の 『High-Performance Computing (HPC) Tuning Guide』を参照してください。

エンタープライズ ワークロードに推奨される BIOS 設定の概要

表 19 は、さまざまなエンタープライズ ワークロードに推奨される BIOS トークンと設定をまとめたものです。

表 19　仮想化、コンテナ、RDBMS、ビッグデータ分析、HPC エンタープライズ ワークロードに関する BIOS の推奨事項

BIOS オプション	BIOS 値（プラットフォームのデフォルト）	仮想/ コンテナ	RDBMS	ビッグ データ分析	HPC
プロセッサ					
SMT モード	有効	有効	有効	無効	無効
SVM モード	自動 （有効）	自動	自動	自動	自動
DF C-states	自動（有効）	Auto	無効	自動	自動
NUMA ドメインとしての ACPI SRAT L3 キャッシュ	自動（無効）	自動	自動	自動	自動
APBDIS	自動（0）	Auto	1	1	1
固定 SOC P-State SP5F 19h	自動（P0）	自動	自動	自動	自動
4 リンク xGMI 最大速度*	自動（32 Gbps）	自動	自動	自動	自動
強力な CPU パフォーマンス*	無効	無効	無効	無効	自動

BIOS オプション	BIOS 値（プラットフォームのデフォルト）	仮想/ コンテナ	RDBMS	ビッグ データ分析	HPC
メモリ					
ソケットごとの **NUMA** ノード	自動（NPS1）	Auto	NPS4	Auto	NPS4
IOMMU	自動（有効）	自動	自動	自動	自動
Memory interleaving	自動（有効）	自動	自動	自動	自動
電力/パフォーマンス					
コア パフォーマンス ブースト	自動（有効）	自動	自動	自動	自動
グローバル **C-State** 制御	自動（有効）	自動	自動	自動	自動
L1 Stream HW Prefetcher	自動（有効）	自動	自動	自動	自動
L2 Stream HW Prefetcher	自動（有効）	自動	自動	自動	自動
デタミニズム スライダ（**Determinism Slider**）	自動（電力）	自動	自動	自動	自動
CPPC	自動（無効）	Enabled	自動（Auto）	有効	Auto
電力プロファイル選択 F19h	ハイパフォーマンス モード	ハイパフォーマンス モード	最大 I/O パフォーマンス モード	ハイパフォーマンス モード	ハイパフォーマンス モード

注： *で強調表示されている BIOS トークンは、Cisco UCS C225 M8 1U ラックサーバーなどの単一ソケットに最適化されたプラットフォームには適用されません。

- ワークロードの仮想マシンごとの vCPU が少ない場合（つまり、ソケットごとのコア数の 4 分の 1 未満の場合）、次の設定で最高のパフォーマンスが得られる傾向があります。
  - NUMA NPS（ソケットごとのノード）= 4
  - LLC As NUMA がオン
- ワークロード仮想マシンに多数の vCPU がある場合（つまり、ソケットあたりのコア数の半分を超える場合）、次の設定によって最高のパフォーマンスが得られる傾向があります。
  - NUMA NPS（ソケットごとのノード）= 1
  - LLC As NUMA がオフ

詳細については、「VMware vSphere チューニング ガイド」を参照してください。

ハイパフォーマンスを実現するためのオペレーティング システムのチューニングに関するガイダンス

Microsoft Windows、VMware ESXi、Red Hat Enterprise Linux、および SUSE Linux オペレーティングシステムには、デフォルトで有効になる新しい電力管理機能が多数付属しています。したがって、最高のパフォーマンスを実現するには、オペレーティングシステムを調整する必要があります。

追加のパフォーマンス ドキュメントについては、 AMD EPYC パフォーマンス チューニング ガイドを参照してください。

Linux（Red Hat および SUSE）

CPU 周波数ガバウンタは、システム CPU の電力特性を定義します。これは、CPU パフォーマンスに影響します。各ガバナには、独自の独自の動作、目的、およびワークロードの観点における適合性があります。

パフォーマンス ガバナーは、CPU に使用可能な最高のクロック周波数を強制します。この頻度は静的に設定され、変更されません。したがって、この特定のガバには省電力のメリットはありません。これは、数時間の重いワークロードにのみ適しており、その後も、CPU がめったに（または決して）アイドルになっていない時間帯にのみ適しています。デフォルト設定は「オンデマンド」です。これにより、CPU はシステムの負荷が高いときに最大クロック周波数を実現し、システムがアイドル状態のときに最小クロック周波数を実現できます。この設定により、システムはシステム負荷に従って電力消費を調整できますが、周波数切り替えによる遅延は処理されません。

パフォーマンス ガバメントは、 **cpupower** コマンドを使用して設定できます。

```
CPU の frequency-set -g パフォーマンス
```

追加情報については、次のリンクを参照してください。

- Red Hat Enterprise Linux：パフォーマンス CPU Freq ガバナを設定します。
- SUSE Enterprise Linux Server：パフォーマンス CPU Freq ガバナを設定します。

Microsoft Windows Server 2019 および 2022

Microsoft Windows Server 2019 の場合、デフォルトでは、バランスの取れた（推奨）電力計画が使用されます。この設定は省電力を有効にしますが、遅延が増加し（一部のタスクの応答時間が遅く）、CPU 集約型アプリケーションのパフォーマンスの問題が発生する可能性があります。パフォーマンスを最大にするには、電力プランを [ハイパフォーマンス（High Performance）] に設定します。

追加情報については、次のリンクを参照してください。

- Microsoft Windows と Hyper-V：電力ポリシーをハイ パフォーマンスに設定します。

VMware ESXi

VMware ESXi の電力管理は、ESXi ホストの電力がオンになっているときの ESXi ホストの電力消費を削減するように設計されています。最大のパフォーマンスを実現するために、電力ポリシーを [高パフォーマンス（High Performance）] に設定します。

追加情報については、次のリンクを参照してください。

- VMware ESXi：電力ポリシーをハイパフォーマンスに設定します。

まとめ

パフォーマンスに関してシステム BIOS 設定を調整する際には、プロセッサおよびメモリのさまざまなオプションを考慮する必要があります。最高のパフォーマンスが目標である場合は、省電力よりも優先してパフォーマンスを最適化するオプションを選択してください。また、メモリインターリーブや CPU ハイパースレッディングなどの他のオプションも試してください。最も重要なのは、アプリケーションが必要とするパフォーマンスに対する設定の影響を評価することです。

詳細情報

AMD第 4 世代および第 5世代 プロセッサを搭載した Cisco UCS M8 サーバーの詳細については、次のリソースを参照してください。

- IMM BIOS トークン ガイド：
  - https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/Intersight/IMM_BIOS_Tokens_Guide/b_IMM_Server_BIOS_Tokens_Guide.pdf
- Cisco UCS X215c M8 コンピューティング ノード
  - https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-x-series-modular-system/ucs-x215c-m8-compute-node-aag.html
- Cisco UCS C245 M8 ラック サーバー：
  - https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c245-m8-rack-server-aag.html
- Cisco UCS C225 M8 ラック サーバー：
  - https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c225-m8-rack-server-aag.html
- AMD EPYC 調整ガイド：
  - https://developer.amd.com/resources/epyc-resources/epyc-tuning-guides/
  - https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/tuning-guides/58015-epyc-9004-tg-architecture-overview.pdf
  - https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/white-papers/58649_amd-epyc-tg-low-latency.pdf
  - https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/tuning-guides/57996-epyc-9004-tg-rdbms.pdf
  - https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/tuning-guides/58002_amd-epyc-9004-tg-hpc.pdf
  - https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/tuning-guides/58008-epyc-9004-tg-containers-on-kubernetes.pdf

- https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/tuning-guides/58013-epyc-9004-tg-hadoop.pdf
- https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/tuning-guides/58007-epyc-9004-tg-mssql-server.pdf
- https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/tuning-guides/58001_amd-epyc-9004-tg-vdi.pdf

米国本社
Cisco Systems, Inc.
カリフォルニア州サンノゼ

アジア太平洋本社
Cisco Systems (USA), Pte. Ltd.
シンガポール

ヨーロッパ本社
Cisco Systems International BV
Amsterdam, The Netherlands

2023 年 11 月発行

米国で印刷 C11-4692101-02 04/25