ホワイトペーパー
PRIMERGY & PRIMEQUEST
Xeon スケーラブル・プロセッサ（Sapphire Rapids / Emerald Rapids）搭載システムのメモリパフォーマンス

このホワイトペーパーでは、第 4 世代 Xeon スケーラブル・プロセッサ（Sapphire Rapids）、および、第 5 世代 Xeon スケーラブル・プロセッサ（Emerald Rapids）搭載 PRIMERGY 及び PRIMEQUEST でのメモリアーキテクチャーの重要な特徴と最近の改善点について説明し、それが商用アプリケーションのパフォーマンスに与える影響を定量化します。

バージョン
1.1 2024-07-30

目次

はじめに

第 4 世代 Xeon スケーラブル・プロセッサ（Sapphire Rapids）は、前世代の Xeon スケーラブル・プロセッサ（Ice Lake）世代の特長を受け継ぐとともに、Intel の最新製造プロセスを使用することで、前世代のプロセッサから大きな性能向上を実現しています。プロセッサの最上位モデルでは、多くのシナリオで前世代のプロセッサに比べて 50% 以上の性能向上を果たしています。この成果の大きな要因として、プロセッサあたりのコア数が最大 40 から最大 60 に増えたこと、より進化したマイクロアーキテクチャーを採用したことが挙げられます。さらに第 5 世代 Xeon スケーラブル・プロセッサ（Emerald Rapids）では、プロセッサあたり最大 64 コア、L3 キャッシュの大幅な増加（最大 320 MB）などの改善が行われています。
メモリアーキテクチャーの観点でも、プロセッサは大幅に進化しました。キャッシュの増量に加えて、最新の DDR5 メモリがサポートされました。最大メモリ転送速度は、前世代のシステムが 3,200 MT/s であったのに対して、Sapphire Rapids 世代では 4,800 MT/s をサポートします。これらにより、理論メモリ帯域はプロセッサあたり最大 307 GB/s となりました。続く Emerald Rapids 世代では 5,600 MT/s DDR5 のサポートにより、理論メモリ帯域はプロセッサあたり最大 358 GB/s に達します。また、大容量の 256 GB 3DS RDIMM　のサポートにより、1 プロセッサあたり 4 TB のメモリを搭載可能です。
このプロセッサが、隣接プロセッサのメモリ（リモートメモリ）の内容を要求するとき、Ultra Path Interconnect （UPI）リンクを使用します。リモートメモリへのアクセスのパフォーマンスは、ローカルメモリアクセスに比べるとさほど高くありません。ローカルメモリとリモートメモリのアクセスを区別するこのアーキテクチャーは、NUMA（Non-Uniform Memory Access：非均等型メモリアクセス）タイプのアーキテクチャーです。このプロセッサ間接続の速度は、Saphire Rapids 世代では、前世代の 11.2 GT/s から 16 GT/s へ、Emerald Rapids 世代では最大 20 GT/s へと大きく引き上げられました。また、リンクの個数は最大 3 リンクから最大 4 リンクに増えました。
Ice Lake 世代では、SNC（Sub-NUMA Clustering）と呼ばれるプロセッサ内のクラスタリングに関するオプションに加えて、新しく UMA-Based Clustering と呼ばれるオプションが追加されました。Sapphire Rapids 世代ではこれらのオプションが強化されています。なお、これらのオプションは、ローカルおよびリモートメモリアクセスに関するレイテンシと帯域幅のトレードオフの扱いが異なりますが、わずかなパフォーマンスの違いを気にする特殊な場合をのぞき、ほとんどのアプリケーションでは、デフォルト設定から変更する必要はありません。

このドキュメントでは、最新のサーバ世代の新しいメモリシステム機能について見ていきます。一方で、これまでのホワイトペーパーと同様に、強力なシステムを構成する上で不可欠な UPI ベースのメモリアーキテクチャーの基本的な知識について説明します。ここでは、次の点を取り上げます。

- NUMA アーキテクチャーであるため、各プロセッサのメモリを可能な限り同等の構成にする必要があります。これは、各プロセッサが原則としてそのローカルメモリ上で動作できるようにするためです。
- メモリアクセスを並列化し、さらに高速化するために、物理アドレス空間の隣接する領域をメモリシステムの複数のコンポーネントに分散させます。これは技術用語でインターリーブと呼ばれます。インターリーブは 2 つの次元で行われます。まず、プロセッサあたり 8 つのメモリチャネルが横方向に存在します。各プロセッサのメモリ搭載数を 8 の倍数とすることで、この方向への最適なインターリーブを実現します。また、個々のメモリチャネルの中でもインターリーブを実現しています。このための決定的なメモリリソースが、いわゆるランク数です。ランク数は、DIMM の下位構造で、ここに DRAM（Dynamic Random Access Memory：ダイナミックランダムアクセスメモリ）チップのグループが統合されています。個々のメモリアクセスでは、常にこのようなグループを参照します。
- メモリ転送速度はパフォーマンスに影響を与えます。プロセッサタイプ、DIMM タイプ、メモリ容量、チャネルあたりの DIMM の枚数、および BIOS 設定に応じて、5,600、5,200、4,800、4,400、4,000、3,200 MT/s のいずれかになります。

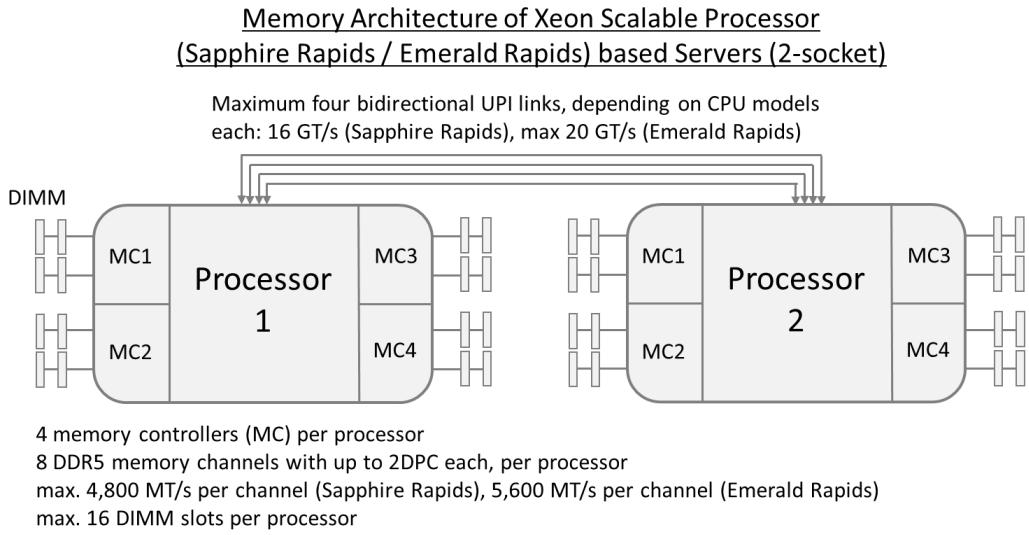
このホワイトペーパーでは、メモリ性能に影響を与える要因を取り上げ、定量化しています。定量化には、STREAM と SPECrate2017 Integer のベンチマークを使用します。STREAM でメモリ帯域幅を測定します。SPECrate2017 Integer は、商用アプリケーションのパフォーマンスのモデルとして使用されます。
測定結果では、プロセッサのパフォーマンスごとの影響を比で示します。構成プロセッサモデルが強力であるほど、本書で取り上げているメモリ構成の問題について十分に考慮する必要があります。
ミラーリングや ADDDC スペアリングなど、冗長性を考慮する場合のメモリパフォーマンスについては、本書の最後にまとめています。

メモリアーキテクチャー

ここでは、5 部構成でメモリシステムの概要を説明します。まずブロック図で、利用可能な DIMM スロットの配置を説明します。2 つ目のセクションでは、使用可能な DIMM タイプを示します。続く 3 つ目のセクションでは、有効なメモリ転送速度への影響について説明します。4 つ目のセクションでは、メモリシステムに影響を与える BIOS パラメーターについて説明します。最後のセクションでは、メモリパフォーマンスを最適化した DIMM 構成例のリストを示します。

DIMM スロットとメモリコントローラ

以下の図では、第 4 世代 Xeon スケーラブル・プロセッサ（Sapphire Rapids）、および、第 5 世代 Xeon スケーラブル・プロセッサ（Emerald Rapids）搭載システムでのメモリシステムの構造を例示しています。



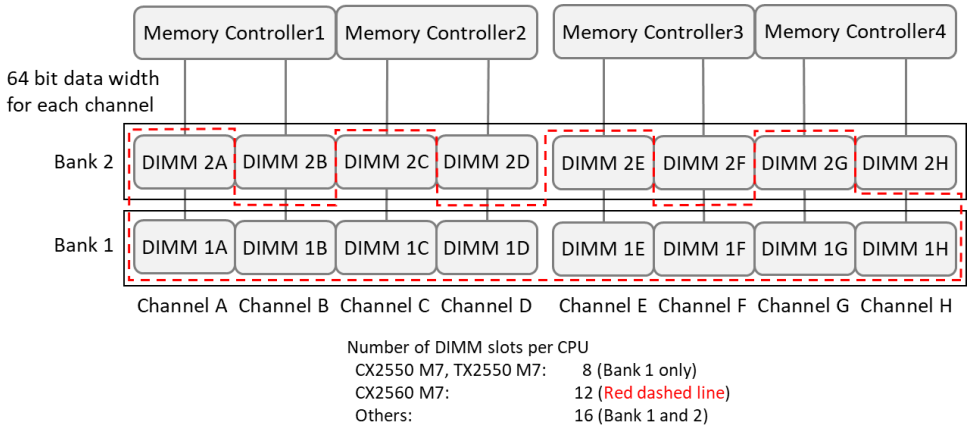
Sapphire Rapids 搭載 PRIMERGY 及び PRIMEQUEST サーバ、および、Emerald Rapids 搭載 PRIMERGY サーバは、プロセッサあたり最大 16 本の DIMM スロットを装備しています。データパス幅は、DDR4 同様に 64 ビットですが、DDR5 では、これは、2 つの 32 ビットサブチャネルとして独立して動作します。これにより、並列アクセス性能は DDR4 から大きく向上しました。

前世代のプロセッサと同様に、1 つのプロセッサには、4 つのメモリコントローラと 8 つのメモリチャネルが存在します。しかし、DDR5 メモリの採用により、4,800 MT/s DDR5 の場合、メモリ帯域幅は理論値で 50%、5,600 MT/s DDR5 の場合で 75% 向上しています。

Sapphire Rapids、および、Emerald Rapids プロセッサでは、プロセッサのモデルによっては、チャネルあたりの DIMM 数である DPC（以降、この用語を使用します）の値が変わると、メモリ転送速度に変化が生じ、メモリパフォーマンスに影響が生じます。これまでの Xeon スケーラブル・プロセッサ搭載 PRIMERGY 及び PRIMEQUEST サーバでは DPC によってメモリ転送速度は変化しなかったので、この点には注意が必要です。

前世代（Ice Lake）では、UPI リンクの速度は最大 11.2 GT/s でしたが、Sapphire Rapids では最大 16 GT/s、Emerald Rapids では最大 20 GT/s に向上しました。また、プロセッサ間の UPI リンクの本数は、Ice Lake 世代の RX サーバの場合、最大 3 本でしたが、Sapphire Rapids、および、Emerald Rapids 世代では最大 4 本に増加しました。これらにより、データベース処理のようなプロセッサ間にまたがるメモリアクセスが多いアプリケーションの性能向上が期待できます。

以降では、「メモリバンク」という用語も使用します。次の図では、複数のチャネルに分配されている 8 つの DIMM のグループが、1 つのバンクを形成しています。プロセッサあたりの利用可能なスロット経由で DIMM を分配する場合、バンク 1 から順に割り当てることにより、チャネル全体で最適なインターリーブが得られます。インターリーブは、メモリパフォーマンスに影響を与える主要因です。



64 ビットのデータ幅に対し、DIMM 上の個々の DRAM チップが 4 ビットまたは 8 ビットを受け持ちます（タイプ名の x4 または x8 を参照してください）。このようなチップのグループをランクと呼びます。1 ランク、2 ランク、4 ランク、または 8 ランクの DIMM タイプがあります。
DIMM スロットを使用するためには、対応するプロセッサを搭載する必要があります。プロセッサの搭載が最大構成でない場合、空の CPU ソケットに割り当てられたスロットは使用できません。

プロセッサの正確な分類については、次の表を参照してください。定量的なメモリパフォーマンステストは、トピックに応じて表の右から 2 番目の列に記載したサポートされているメモリ転送速度による分類を使用して、別々に実施しました。

プロセッサ（Sapphire Rapids）									
モデル	タイプ	コア数	スレッド数	L3 キャッシュ [MB]	UPI リンク数 スピード [GT/s]	定格 周波数 [GHz]	最大ターボ 周波数 [GHz]	最大メモリ 転送速度 [MT/s]	TDP [W]
Xeon Platinum 8490H	XCC	60	120	112.5	4x 16	1.90	3.50	4,800	350
Xeon Platinum 8480+	XCC	56	112	105.0	4x 16	2.00	3.80	4,800	350
Xeon Platinum 8470	XCC	52	104	105.0	4x 16	2.00	3.80	4,800	350
Xeon Platinum 8468V	XCC	48	96	97.5	3x 16	2.40	3.80	4,800	330
Xeon Platinum 8468H	XCC	48	96	105.0	4x 16	2.10	3.80	4,800	330
Xeon Platinum 8468	XCC	48	96	105.0	4x 16	2.10	3.80	4,800	350
Xeon Platinum 8462Y+	MCC	32	64	60.0	3x 16	2.80	4.10	4,800	300
Xeon Platinum 8460Y+	XCC	40	80	105.0	4x 16	2.00	3.70	4,800	300
Xeon Platinum 8460H	XCC	40	80	105.0	4x 16	2.20	3.80	4,800	330
Xeon Platinum 8458P	XCC	44	88	82.5	3x 16	2.70	3.80	4,800	350
Xeon Platinum 8454H	XCC	32	64	82.5	4x 16	2.10	3.40	4,800	270
Xeon Platinum 8452Y	XCC	36	72	67.5	4x 16	2.00	3.20	4,800	300
Xeon Platinum 8450H	XCC	28	56	75.0	4x 16	2.00	2.60	4,800	250
Xeon Platinum 8444H	XCC	16	32	45.0	4x 16	2.90	3.20	4,800	270
Xeon Gold 6454S	XCC	32	64	60.0	4x 16	2.10	3.40	4,800	270
Xeon Gold 6448Y	MCC	32	64	60.0	3x 16	2.10	4.10	4,800	225
Xeon Gold 6448H	MCC	32	64	60.0	3x 16	2.40	4.10	4,800	250
Xeon Gold 6444Y	MCC	16	32	45.0	3x 16	3.60	4.10	4,800	270
Xeon Gold 6442Y	MCC	24	48	60.0	3x 16	2.60	4.00	4,800	225
Xeon Gold 6438Y+	MCC	32	64	60.0	3x 16	2.00	4.00	4,800	205
Xeon Gold 6438N	MCC	32	64	60.0	3x 16	2.00	3.60	4,800	205
Xeon Gold 6438M	MCC	32	64	60.0	3x 16	2.20	3.90	4,800	205
Xeon Gold 6434H	MCC	8	16	22.5	3x 16	3.70	4.10	4,800	195
Xeon Gold 6434	MCC	8	16	22.5	3x 16	3.70	4.10	4,800	195
Xeon Gold 6430	XCC	32	64	60.0	3x 16	1.90	3.40	4,400	270
Xeon Gold 6428N	MCC	32	64	60.0	3x 16	1.80	3.80	4,000	185
Xeon Gold 6426Y	MCC	16	32	37.5	3x 16	2.50	4.10	4,800	185
Xeon Gold 6418H	MCC	24	48	60.0	3x 16	2.10	4.00	4,800	185
Xeon Gold 6416H	MCC	18	36	45.0	3x 16	2.20	4.20	4,800	165
Xeon Gold 5420+	MCC	28	56	52.5	3x 16	2.00	4.10	4,400	205
Xeon Gold 5418Y	MCC	24	48	45.0	3x 16	2.00	3.80	4,400	185
Xeon Gold 5418N	MCC	24	48	45.0	3x 16	1.80	3.80	4,000	165
Xeon Gold 5416S	MCC	16	32	60.0	3x 16	2.00	4.00	4,400	150
Xeon Gold 5415+	MCC	8	16	22.5	3x 16	2.90	4.10	4,400	150
Xeon Silver 4416+	MCC	20	40	37.5	2x 16	2.00	3.90	4,000	165
Xeon Silver 4410Y	MCC	12	24	30.0	2x 16	2.00	3.90	4,000	150
Xeon Silver 4410T	MCC	10	20	26.25	2x 16	2.70	4.00	4,000	150
Xeon Gold 6414U	XCC	32	64	60.0	-	2.00	3.40	4,800	250
Xeon Gold 5412U	MCC	24	48	45.0	-	2.10	3.90	4,400	185
Xeon Bronze 3408U	MCC	8	8	22.5	-	1.80	1.90	4,000	125

プロセッサ（Emerald Rapids）									
モデル	タイプ	コア数	スレッド数	L3 キャッシュ [MB]	UPI リンク数 スピード [GT/s]	定格 周波数 [GHz]	最大ターボ 周波数 [GHz]	最大メモリ 転送速度 [MT/s]	TDP [W]
Xeon Platinum 8592V	XCC	64	128	320.0	3x 16	2.00	3.90	4,800	330
Xeon Platinum 8592+	XCC	64	128	320.0	4x 20	1.90	3.90	5,600	350
Xeon Platinum 8580	XCC	60	120	300.0	4x 20	2.00	4.00	5,600	350
Xeon Platinum 8570	XCC	56	112	300.0	4x 20	2.10	4.00	5,600	350
Xeon Platinum 8568Y+	XCC	48	96	300.0	4x 20	2.30	4.00	5,600	350
Xeon Platinum 8562Y+	MCC	32	64	60.0	3x 20	2.80	4.10	5,600	300
Xeon Platinum 8558P	XCC	48	96	260.0	3x 20	2.70	4.00	5,600	350
Xeon Platinum 8558	XCC	48	96	260.0	3x 20	2.10	4.00	5,200	330
Xeon Gold 6558Q	MCC	32	64	60.0	3x 20	3.20	4.10	5,200	350
Xeon Gold 6554S	XCC	36	72	180.0	4x 20	2.20	4.00	5,200	270
Xeon Gold 6548Y+	MCC	32	64	60.0	3x 20	2.50	4.10	5,200	250
Xeon Gold 6548N	MCC	32	64	60.0	3x 20	2.80	4.10	5,200	250
Xeon Gold 6544Y	MCC	16	32	45.0	3x 20	3.60	4.10	5,200	270
Xeon Gold 6542Y	MCC	24	48	60.0	3x 20	2.90	4.10	5,200	250
Xeon Gold 6538Y+	MCC	32	64	60.0	3x 20	2.20	4.00	5,200	225
Xeon Gold 6538N	MCC	32	64	60.0	3x 20	2.10	4.10	5,200	205
Xeon Gold 6534	MCC	8	16	22.5	3x 20	3.90	4.20	4,800	195
Xeon Gold 6530	XCC	32	64	160.0	3x 20	2.10	4.00	4,800	270
Xeon Gold 6526Y	MCC	16	32	37.5	3x 20	2.80	3.90	5,200	195
Xeon Gold 5520+	MCC	28	56	52.5	3x 20	2.20	4.00	4,800	205
Xeon Gold 5515+	MCC	8	16	22.5	3x 20	3.20	4.10	4,800	165
Xeon Silver 4516Y+	MCC	24	48	45.0	2x 16	2.20	3.70	4,400	185
Xeon Silver 4514Y	MCC	16	32	30.0	2x 16	2.00	3.40	4,400	150
Xeon Silver 4510T	LCC	12	24	30.0	2x 16	2.00	3.70	4,400	115
Xeon Silver 4510	LCC	12	24	30.0	2x 16	2.40	4.10	4,400	150
Xeon Silver 4509Y	LCC	8	16	22.5	2x 16	2.60	4.10	4,400	125
Xeon Platinum 8581V	XCC	60	120	300.0	-	2.00	3.90	4,800	270
Xeon Platinum 8558U	XCC	48	96	260.0	-	2.00	4.00	4,800	300
Xeon Gold 5512U	MCC	28	56	52.5	-	2.10	3.70	4,800	185
Xeon Bronze 3508U	LCC	8	8	22.5	-	2.10	2.20	4,400	125

使用可能な *DIMM* タイプ

Sapphire Rapids 搭載の PRIMERGY 及び PRIMEQUEST サーバ、Emerald Rapids 搭載 PRIMERGY サーバには、これまでの Xeon スケーラブル・プロセッサ搭載サーバと異なり、DDR5 SDRAM メモリモジュールが使用されています。これらのシステムでは、以下の改善がなされました。

- Emerald Rapids プロセッサ搭載時に最大転送速度 5,600 MT/s までの DDR5 をサポートします。前世代の Ice Lake 搭載システムは、DDR4 SDRAM を使用し、最大 3,200 MT/s をサポートしていました。
- 前世代のシステムと同様に、256 GB 3DS RDIMM を使用することで、1 ソケットあたり最大 4 TB の DRAM を搭載可能です。

次の表に、これらのサーバでサポートされる DIMM を示します。DIMM には、Registered DIMM（RDIMM）、3DS Registered DIMM（3DS RDIMM） があります。RDIMM x4、RDIMM x8、および 3DS RDIMM の混在はできません。

DIMM タイプ[1]	制御	最大転送速度 (MT/s)	電圧 (V)	ランク数	容量
16GB (1x16GB) 1Rx8 DDR5-4800 R ECC	Registered	4,800	1.1	1	16 GB
32GB (1x32GB) 2Rx8 DDR5-4800 R ECC	Registered	4,800	1.1	2	32 GB
32GB (1x32GB) 1Rx4 DDR5-4800 R ECC	Registered	4,800	1.1	1	32 GB
64GB (1x64GB) 2Rx4 DDR5-4800 R ECC	Registered	4,800	1.1	2	64 GB
128GB (1x128GB) 4Rx4 DDR5-4800 3DS R ECC	3DS Registered	4,800	1.1	4	128 GB
256GB (1x256GB) 8Rx4 DDR5-4800 3DS R ECC	3DS Registered	4,800	1.1	8	256 GB
16GB (1x16GB) 1Rx8 DDR5-5600 R ECC	Registered	5,600	1.1	1	16 GB
32GB (1x32GB) 2Rx8 DDR5-5600 R ECC	Registered	5,600	1.1	2	32 GB
32GB (1x32GB) 1Rx4 DDR5-5600 R ECC	Registered	5,600	1.1	1	32 GB
64GB (1x64GB) 2Rx4 DDR5-5600 R ECC	Registered	5,600	1.1	2	64 GB
96GB (1x96GB) 2Rx4 DDR5-5600 R ECC	Registered	5,600	1.1	2	96 GB
128GB (1x128GB) 4Rx4 DDR5-5600 3DS R ECC	3DS Registered	5,600	1.1	4	128 GB
256GB (1x256GB) 8Rx4 DDR5-5600 3DS R ECC	3DS Registered	5,600	1.1	8	256 GB

[1] サポートされる DIMM のタイプは、サーバ、プロセッサにより異なります。

2 つの DIMM タイプの重要な特徴は、次のようになります。

- RDIMM：メモリコントローラの制御コマンドは、DIMM 上の独自のコンポーネントにあるレジスター内でバッファーされます（これが名前の由来です）。メモリチャネルの負荷が軽減されることで、最大 2 DPC（チャネルあたりの DIMM）での構成が可能になります。
- 3DS RDIMM：Three Dimensional Stack （3DS） という規格に基づき、シリコン貫通電極(Through Silicon Via) 技術により複数枚のシリコン・ダイを積層させた RDIMM です。マスターと呼ばれる 1 枚のダイだけが外部と信号をやり取りし、それ以外のダイはスレーブとしてマスターとだけ信号をやり取りするアーキテクチャーを採用しており、大容量化や高速化が可能になります。

RDIMM または 3DS RDIMM のうち、どのタイプが望ましいかは、通常、必要なメモリ容量によって決まります。ただし、3DS RDIMM には、若干の性能オーバーヘッドがあります。
なお、販売地域によっては、利用できない DIMM タイプがあります。

メモリ転送速度

Sapphire Rapids 搭載 PRIMERGY 及び PRIMEQUEST サーバのメモリ転送速度には、4,800、4,400、4,000 および 3,200 MT/s の 4 種類が、Emerald Rapids 搭載 PRIMERGY サーバには 5,600、5,200、4,800、4,400 および 3,200 MT/s の 5 種類があります。システムに電源が入ると、転送速度が BIOS によって設定され、プロセッサごとではなくシステムごとに適用されます。

メモリ転送速度は、プロセッサモデルの最大メモリ転送速度、メモリ構成の DPC 値、そして、BIOS 設定が影響します。プロセッサの最大メモリ転送速度は、DIMM スロットとメモリコントローラの表にあるとおり、モデルによって異なります。また、2DPC 構成では、最大メモリ転送速度が 4,800 MT/s 以上の CPU モデルでもメモリ転送速度は 4,400 MT/s に低下します。これを BIOS で無効にすることはできません。

BIOS パラメーターの DDR Performance を使用することで、限定的ですがパフォーマンスと消費電力のどちらを優先させるかを選択できます（詳細は後述）。Performance optimized（性能に最適化）を選択した場合、有効なメモリ転送速度は次の表のようになります。Performance optimized はデフォルトの BIOS 設定です。

DDR Performance = Performance optimized （性能に最適化、デフォルト）				
プロセッサ最大メモリ転送速度	RDIMM		3DS RDIMM	
	1DPC	2DPC	1DPC	2DPC
5,600 MT/s	5,600	4,400	5,600	4,400
5,200 MT/s	5,200	4,400	5,200	4,400
4,800 MT/s	4,800	4,400	4,800	4,400
4,400 MT/s	4,400	4,400	4,400	4,400
4,000 MT/s	4,000	4,000	4,000	4,000

Energy optimized（消費電力に最適化）を選択した場合、有効なメモリ転送速度は次の表のようになります。前述のように、現在のところ DDR5 メモリモジュールに低電圧版はありません。DDR5 モジュールは常に 1.1 V 電圧で動作します。
メモリ転送速度を下げることでわずかに消費電力を節約できますが、メモリモジュールの消費電力は主に電圧の影響を受ける点に注意してください。メモリ転送速度を下げるとシステムパフォーマンスも低下するため（本ドキュメントの第 2 部で説明）、次の表に従って設定を行う際は、ある程度の注意を払うことをお勧めします。注意を払うとは、本稼働の前に影響をテストするということです。

DDR Performance = Energy optimized （消費電力に最適化）				
プロセッサ最大メモリ転送速度	RDIMM		3DS RDIMM	
	1DPC	2DPC	1DPC	2DPC
5,600 MT/s	3,200	3,200	3,200	3,200
5,200 MT/s	3,200	3,200	3,200	3,200
4,800 MT/s	3,200	3,200	3,200	3,200
4,400 MT/s	3,200	3,200	3,200	3,200
4,000 MT/s	3,200	3,200	3,200	3,200

BIOS パラメーター

前のセクションでは、BIOS パラメーター DDR Performance を見ましたが、ここでは、メモリシステムに影響を与える他の BIOS オプションを見ていきます。このパラメーターは、Advanced（詳細）の下のサブメニュー、Memory Configuration（メモリ構成）にあります。

Memory Configuration（メモリ構成）のメモリパラメーター

次の 10 個のパラメーターがあります。それぞれ下線付きのオプションがデフォルトです。

- Memory Mode[2,3]：Independent／Mirroring／Address Range Mirroring
- ADDDC Sparing[2,3]：Disabled／Enabled
- DDR5 ECS：Disabled／Enabled
- NUMA[2]：Disabled／Enabled
- Virtual NUMA：Disabled／Enabled
- DDR Performance：Performance optimized／Energy optimized
- PPR Type：Hard PPR／Soft PPR／PPR Disabled
- Patrol Scrub：Disabled／Enabled
- SNC(Sub NUMA)：Disabled／Enable SNC2／Enable SNC4
- UMA-Based Clustering：Hemisphere (2-clusters)／Quadrant (4-cluster)

最初の 3 つのパラメーター、Memory Mode、ADDDC Sparing（ADDDC：Adaptive Double Device Data Correction）、DDR5 ECS（Error Correcting Code Scrubbing）は冗長性機能を扱います。これらは、RAS（Reliability：信頼性、Availability：可用性、Serviceability：サービス性）機能の一部です。

Memory Mode は、メモリのデータを複製するか（ミラーリング）を指定します。Mirroring を指定するとミラーリングが有効となりますが、メモリ容量は半分になります。Address Range Mirroring は、システムメモリの一部をミラーリングします。これには、オペレーティングシステムのサポートが必要です。

ADDDC Sparing は、メモリエラーが頻繁に発生する場合に DIMM ランクまたはバンクのレベルで予備領域を有効化すること（スペアリング）で、耐故障能力を向上させます。2 つの DRAM デバイスのエラーに対してエラー訂正を行うことができます。ミラーリング有効化時には、ADDDC Sparing は無効となります。

DDR5 ECS（Error Check and Scrub）は、信頼性とエラー訂正機能を改善する DDR5 の機能です。

DDR5 では、データ読み込み時にデバイス内部でエラー訂正が可能となりました（On Die ECC 機能）。ECS はこの機能を利用し、DRAM 内でデータの読み込みとエラー時のデータ修正と書き戻しを行います。有効にすると定期的にチェックが行われます。

Memory Mode、ADDDC Sparing が利用できる構成については制限があります。これらについては、システム構成図を参照してください。

これらの機能が要求される場合、工場での出荷時には適切なデフォルト設定が行われます。それ以外の場合、これらのパラメーターは Independent（通常の冗長性なし）、および、 Disabled（無効化）に設定されます。これらの冗長性機能がシステムパフォーマンスに与える影響に関する数値を後で示します。

---

[2] このパラメーターは PRIMEQUEST サーバにはありません。

[3] PRIMEQUEST サーバでは iRMC で同様な機能を設定できます。

4 番目のパラメーター NUMA は、有効にすることでローカルメモリのセグメントを構築し、オペレーティングシステムに構造を通知するか、または、無効にすることでソケットレベルでのメモリインターリーブを行うかを定義します。デフォルト設定は Enabled です。明確な理由がないこれを限り変更しないでください。このトピックの数量的な面については、後述します。
5 番目のパラメーター Virtual NUMA は、64 個を超える論理 CPU を持つプロセッサを Windows で使用する場合に使用します。Windows が論理 CPU を管理するために用いるプロセッサ・グループは、64 論理 CPU が上限のため、それを超える論理 CPU は別のプロセッサ・グループとして管理されます。その結果、プロセッサ・グループの大きさが不均一となることで、性能面で不利となります。Virtual NUMA を有効にすることで、プロセッサは 2 つの同サイズの仮想的な NUMA ノードに分割して使用されます。後述の SNC と似ていますが、SNC の持つ性能向上の効果はありません。
6 番目のパラメーター DDR Performance は、メモリ転送速度に関係しています。これについては、直前のセクションで説明しました。
7 番目の PPR Type は、DDR5 の機能である Post Package Repair（PPR、ポストパッケージリペア）を扱います。PPR は、システム起動時に故障したメモリセルを DRAM チップ内の予備領域に置き換えます。Soft PPR を設定すると、この置き換えはシステムの電源オフやリセットで失われます。Hard PPR を設定すると、置き換えは恒久的に保持されます。PPR Disabled の場合、置き換えは行われません。
8 番目のパラメーターPatrol Scrub パラメーターは、Enabled に設定すると、メインメモリに対し修正可能なエラーの検索が定期的に実行され、必要に応じて修正が開始されます。これにより、自動修正が不可能になるようなメモリエラーの累積を防ぎます（対応するレジスターでカウントされます）。感度の高いパフォーマンス指標がある場合は、この機能に影響をうける可能性があります。ただし、パフォーマンスに及ぶ影響を実証するのは難しい場合があります。

最後の 2 つのパラメーターは、プロセッサ内のクラスタリングに関する設定です。
SNC（Sub NUMA）は、プロセッサ内でコア、L3 キャッシュ、メモリコントローラをクラスタに分割するためのパラメーターです。XCC タイプの Sapphire Rapids プロセッサでは Enable SNC4、Enable SNC2、Disabled の 3 つのオプションが、それ以外のプロセッサでは Enable SNC2、Disabled の 2 つのオプションが選択できます。デフォルトでは Disabled が設定されます。
Enable SNC4 に設定すると、これらのリソースは、4 つに分割されたプロセッサ内のクラスタのどれか一つにくくりつけられます。Enable SNC2 に設定すると、2 つに分割されたプロセッサ内のクラスタのどちらか一つにくくりつけられます。このクラスタは、オペレーティングシステムからは一つの NUMA ドメインとして扱われます。Disabled の場合、プロセッサは UMA（Uniform Memory Access：均等型メモリアクセス）である 1 つのクラスタとして扱われます。
SNC により、NUMA ノード内のコアから L3 キャッシュやメモリへのアクセスは、そのレイテンシが改善します。ローカルメモリレイテンシを最小化、ローカルメモリ帯域を最大化することができるため、NUMA 最適化されたアプリケーションにおいて特に推奨されます。
UMA-Based Clustering パラメーターは、XCC タイプの Sapphire Rapids プロセッサのみで利用可能です。UMA 構成でのキャッシュコヒーレンシーの動作を変えます。デフォルトの Quadrant (4-clusters) では、L3 キャッシュとメモリコントローラは、お互いの近さに基づいて 4 つの領域に分割されます。コアは分割されません。Hemisphere (2-clusters) では、2 つの領域に分割されます。分割された領域が小さいほど、L3 キャッシュとメモリ間の距離が短くなり、レイテンシが改善します。
SNC や UMA-Based Clustering を利用できる構成については制限があります。DIMM をシステム構成図に記載の通りに搭載した場合、SNC2 は DIMM の枚数が 2 の倍数の時、SNC4 と Quadrant は DIMM の枚数が 4 の倍数の時、そして Hemisphere は DIMM の枚数が 2 の倍数（ただし 6 枚を除く）の時に利用できます。

パフォーマンスを考慮したメモリ構成

メモリパフォーマンスにはメモリ転送速度と使用するメモリチャネル数が大きく影響します。メモリ転送速度は、搭載するプロセッサの種類と DPC に依存します。また、Xeon スケーラブル・プロセッサはプロセッサあたり全部で 8 本のメモリチャネルがあり、高いメモリパフォーマンスを実現するためには、可能な限り多くのメモリチャネルに DIMM を配置する必要があります。さらに、メモリパフォーマンスに影響する構成機能がいくつかあります。ランク数、冗長機能の有効化、NUMA 機能の無効化などの機能です。本ドキュメントの第 2 部では、これらのトピックのテスト結果を報告します。

パフォーマンスモード構成

常に注意すべき 2 つ目の要因は、DIMM 配置の影響です。最小構成（構成プロセッサあたり 16 GB DIMM 1 枚）から最大構成（複数の 256 GB DIMM からなるフル構成）の間には、いくつかのメモリパフォーマンスについての理想的な構成があります。次の表に、特に興味深い構成を挙げています（すべての構成を網羅している訳ではありません）。
これらの構成で、プロセッサあたり全部で 8 本のメモリチャネルという点は同じです。バンク単位の構成では、同タイプの DIMM 8 枚セットを使用しています。これにより、メモリアクセスは、これらのメモリシステムリソースに均等に分散されます。技術的に言えば、メモリチャネル経由で最適な 8 WAY インターリーブが実現します。本書では、これをパフォーマンスモード構成と呼んでいます。

Xeon スケーラブル・プロセッサ搭載 PRIMERGY 及び PRIMEQUEST サーバのパフォーマンスモード構成（最大メモリ転送速度 4,800 MT/s 以下の Sapphire Rapids および Emerald Rapids の場合）						
1 CPU システム	2 CPU システム	DIMM タイプ	DIMM サイズ（GB）バンク 1	DIMM サイズ（GB）バンク 2	最大メモリ速度 MT/s	注
128 GB	256 GB	DDR5-4800 R	16		4,800	
192 GB	384 GB	DDR5-4800 R	16	8	4,400	混在構成
256 GB	512 GB	DDR5-4800 R	16	16	4,400	
256 GB	512 GB	DDR5-4800 R	32		4,800	
384 GB	768 GB	DDR5-4800 R	32	16	4,400	混在構成
512 GB	1024 GB	DDR5-4800 R	32	32	4,400	
512 GB	1024 GB	DDR5-4800 R	64		4,800	
768 GB	1536 GB	DDR5-4800 R	64	32	4,400	混在構成
1024 GB	2048 GB	DDR5-4800 R	64	64	4,400	
1024 GB	2048 GB	DDR5-4800 3DS R	128		4,800	
2048 GB	4096 GB	DDR5-4800 3DS R	128	128	4,400	
2048 GB	4096 GB	DDR5-4800 3DS R	256		4,800	メモリ速度 4,800 MT/s での最大構成
4096 GB	8192 GB	DDR5-4800 3DS R	256	256	4,400	最大構成

Xeon スケーラブル・プロセッサ搭載 PRIMERGY サーバのパフォーマンスモード構成（最大メモリ転送速度 4,800 MT/s 超の Emerald Rapids の場合）						
1 CPU システム	2 CPU システム	DIMM タイプ	DIMM サイズ（GB）バンク 1	DIMM サイズ（GB）バンク 2	最大メモリ速度 MT/s	注
128 GB	256 GB	DDR5-5600 R	16		5,600	
192 GB	384 GB	DDR5-5600 R	16	8	4,400	混在構成
256 GB	512 GB	DDR5-5600 R	16	16	4,400	
256 GB	512 GB	DDR5-5600 R	32		5,600	
384 GB	768 GB	DDR5-5600 R	32	16	4,400	混在構成
512 GB	1,024 GB	DDR5-5600 R	32	32	4,400	
512 GB	1,024 GB	DDR5-5600 R	64		5,600	
768 GB	1,536 GB	DDR5-5600 R	64	32	4,400	混在構成
1,024 GB	2,048 GB	DDR5-5600 R	64	64	4,400	
1,024 GB	2,048 GB	DDR5-5600 3DS R	128		5,600	
2,048 GB	4,096 GB	DDR5-5600 3DS R	128	128	4,400	
2,048 GB	4,096 GB	DDR5-5600 3DS R	256		5,600	メモリ速度 5,600 MT/s での最大構成
4,096 GB	8,192 GB	DDR5-5600 3DS R	256	256	4,400	最大構成

表は左端の総メモリ容量に従って構成されています。総容量は、1 つまたは 2 つのプロセッサ構成で定義されています。どのプロセッサについてもメモリ構成は同じであるという想定です。次の列は、使用した DIMM タイプです。RDIMM または 3DS RDIMM テクノロジーが決定要因です。その次の列で DIMM サイズがバンク単位で表記されているのは、パフォーマンスモード構成を使用するため、DIMM を 8 枚 1 組で（バンク単位で）構成するからです。
表の最小構成で 1 つのプロセッサに対して 128 GB となっているのは、各プロセッサに 16 GB DIMM を 8 枚（128 GB）使用しているためです。パフォーマンスモード構成では、バンクあたり 8 枚の同一の DIMM をセットで使用する必要がありますが、次の制限を満たしていれば、別のバンクで異なる DIMM サイズを使用することもできます。

- RDIMM、3DS RDIMM を併用することはできません。
- タイプ x4 と x8 の RDIMM を併用することはできません。
- 構成はバンク 1 から 2 に DIMM サイズを小さくしながら進めます。大きいモジュールほど先に取り付けます。

表の右から 2 列目は、それぞれの構成で達成可能な最大メモリ転送速度を示しています。ただし、その値に達するかどうかは使用するプロセッサモデルに依存します。

インディペンデントモード構成

これには、パフォーマンスモードには含まれない構成がすべて含まれます。以下のルールがありますが、詳細については、各機種のシステム構成図を参照ください。

- RDIMM、3DS RDIMM を併用することはできません。
- DIMM タイプが x4 と x8 の RDIMM を併用することはできません。
- 構成はバンク 1 から 2 に DIMM サイズを小さくしながら進めます。大きいモジュールほど先に取り付けます。
- 1 プロセッサ に搭載できる DIMM 枚数は 1、2、4、6、8、12、または 16 枚に制限されます。

プロセッサあたりの DIMM 枚数 が 8 の倍数にならない構成、つまりパフォーマンスモード構成に必要な最小数未満の構成にも注意する必要があります。省電力や、必要なメモリ容量が少ないといった理由のためにこのような構成にすることがあります。DIMM 数の最小化で費用削減が実現できる場合もあります。この後で紹介する、メモリチャネルへのインターリーブ構成がシステムパフォーマンスに及ぼす影響を示した定量的評価からは、1、2DIMM 構成での動作はお勧めしません。

対称型メモリ構成

最後のこのセクションでは、すべてのプロセッサのメモリを可能な限り同等に構成し、BIOS のデフォルト設定を確たる理由なく変更するべきではないということに再度焦点を当てます。
工場でのプレインストールでは、このような状況が当然考慮されています。要求されたメモリモジュールは、プロセッサ全体に可能な限り均等に分散されます。
こうした手法と、関連するオペレーティングシステムによって、ローカルのハイパフォーマンスメモリで可能な限りアプリケーションを実行する前提条件が整備されます。プロセッサコアのメモリアクセスは、通常、各プロセッサに直接割り当てられた DIMM モジュールに対して行われます。
これにどのようなパフォーマンス上のメリットがあるのかを見積もるため、2 WAY サーバのメモリが対称型に構成されているものの、BIOS オプションが NUMA = disabled に設定されている場合の測定結果を後述しています。統計上、メモリアクセスの 2 回に 1 回は、リモートメモリに対して行われます。アプリケーションが 100 %リモートメモリによって実行される非対称型メモリ構成、または片側メモリ構成では、ローカルメモリとリモートメモリが 50 %／50 %の割合で実行される場合の 2 倍パフォーマンスが低下するものとして見積もる必要があります。
なお、1 つ目のプロセッサに 16 枚の DIMM、2 つ目のプロセッサに 8 枚の DIMM の構成は、パフォーマンスモードの基準を満たします。なぜなら、プロセッサごとのメモリチャネルが同じように処理されるからです。ただし、このような構成は推奨されません。

メモリパフォーマンスに対する定量的影響

メモリシステムの機能とその定性的情報を説明した後は、メモリ構成に関係するパフォーマンスの向上と低下について説明します。その準備として、最初のセクションでは、メモリパフォーマンスの特徴を表すための使用する 2 つのベンチマークについて説明します。

その後、すでに説明した特徴であるメモリチャネルのインターリーブ、メモリ転送速度、DIMM タイプの影響、およびキャッシュコヒーレンスプロトコルについて、その影響の大きさの順に説明します。最後に、NUMA = disabled の場合と冗長性を考慮する場合のメモリパフォーマンスについて測定します。

Xeon スケーラブル・プロセッサでは、プロセッサタイプによりサポートする最大メモリ転送速度が異なります。そのため、一部の例外を除いて、定量的テストは、プロセッサがサポートする最大メモリ転送速度を基準にプロセッサを選択し実施しました。

測定は、Linux オペレーティングシステムが動作する 2 つのプロセッサを搭載した PRIMERGY RX2540 M7、または PRIMERGY RX2530 M7 で行いました。次の表は、定量化テストの構成、特にプロセッサクラスの代表的な構成を示します。

SUT（System Under Test：テスト対象システム）	
ハードウェア	
モデル	PRIMERGY RX2540 M7 (Sapphire Rapids 使用時) PRIMERGY RX2530 M7 (Emerald Rapids 使用時)
プロセッサ	Xeon Platinum 8480+ (56 コア、2.0 GHz、最大メモリ転送速度 4,800 MT/s) × 2 Xeon Gold 6430 (32 コア、2.1 GHz、最大メモリ転送速度 4,400 MT/s) × 2 Xeon Silver 4416+ (20 コア、2.0 GHz、最大メモリ転送速度 4,000 MT/s) × 2 Xeon Platinum 8570 (56 コア、2.1 GHz、最大メモリ転送速度 5,600 MT/s) × 2 Xeon Gold 6548N (32 コア、2.8 GHz、最大メモリ転送速度 5,200 MT/s) × 2
メモリタイプ	16GB (1x16GB) 1Rx8 DDR5-4800 R ECC 32GB (1x16GB) 2Rx8 DDR5-4800 R ECC 32GB (1x32GB) 1Rx4 DDR5-4800 R ECC 64GB (1x64GB) 2Rx4 DDR5-4800 R ECC 128GB (1x128GB) 4Rx4 DDR5-4800 3DS R ECC 256GB (1x256GB) 8Rx4 DDR5-4800 3DS R ECC 16GB (1x16GB) 1Rx8 DDR5-5600 R ECC 64GB (1x64GB) 2Rx4 DDR5-5600 R ECC 128GB (1x128GB) 4Rx4 DDR5-5600 3DS R ECC 256GB (1x256GB) 8Rx4 DDR5-5600 3DS R ECC
ディスクサブシステム	SATA 6G SSD 1 台（オンボード SATA コントローラを介して）
ソフトウェア	
BIOS	R1.6.0 (Sapphire Rapids 使用時) R2.4.0 (Emerald Rapids 使用時)
オペレーティングシステム	SUSE Linux Enterprise Server 15 SP4 (Sapphire Rapids 使用時) SUSE Linux Enterprise Server 15 SP5 (Emerald Rapids 使用時)

以下に説明するテストセットには、後述の例外を除き、4,800 MT/s（Sapphire Rapids 使用時）、または、5,600 MT/s（Emerald Rapids 使用時）の 64 GB 2Rx4 RDIMM を使用しました。メモリチャネルへのインターリーブの影響の評価を除き、表に示したその他のすべての DIMM は、DIMM タイプの影響についてのテストセットのみで使用しました。

以降の表は、相対的なパフォーマンスを示します。理想的なメモリ条件下での STREAM および SPECrate2017 Integer ベンチマークの絶対測定値は（通常、表の 1.0 の値に相当）、Xeon スケーラブル・プロセッサ搭載 PRIMERGY サーバの個別のパフォーマンスレポートに記載されています。

測定ツール

測定は、STREAM および SPECrate2017 Integer ベンチマークを使用して行いました。

STREAM ベンチマーク

STREAM ベンチマーク（開発者：John McCalpin 氏）は、メモリのスループットを測定するツールです。このベンチマークは、double 型データの大規模な配列でコピーおよび算術演算を実行して、Copy、Scale、Add、Triad の 4 種類のアクセスの結果を提供します。Copy 以外のアクセスタイプには、算術演算が含まれています。結果は、常に GB/s 単位のスループットで示されます。一般に、Triad の値が最もよく引用されます。以降、STREAM のベンチマークの測定値は、Triad アクセスでの値であり、単位は GB/s です。

STREAM は、サーバのメモリ帯域幅を測定するための業界標準で、シンプルな方法を使用してメモリシステムに大規模な負荷を与えることができます。特にこのベンチマークは、複雑な構成でのメモリパフォーマンスに対する影響を調査する場合に適しています。STREAM は、構成によるメモリへの影響とそれによって生じるパフォーマンスへの影響（低下または向上）を示します。後述する STREAM ベンチマークに関する値は、パフォーマンスへの影響度を示しています。

アプリケーションのパフォーマンスに対するメモリの影響は、各アクセスの遅延時間とアプリケーションが必要とする帯域幅に区別されます。メモリ帯域幅が増加すると遅延時間は増加するため、両者は関連しています。並列メモリアクセスによって遅延時間が相殺される度合いは、アプリケーションや、コンパイラーによって作成されたマシンコードの質にも依存します。このため、すべてのアプリケーションシナリオでの全般的な予測を立てることは非常に困難です。

SPECrate2017 Integer ベンチマーク

SPECrate2017 Integer ベンチマークは、商用アプリケーションパフォーマンスのモデルとして追加されました。これは、Standard Performance Evaluation Corporation（SPEC）の SPEC CPU2017 の一部です。SPEC CPU2017 は、システムのプロセッサ、メモリおよびコンパイラーを評価するための業界標準です。大量の測定結果が公開され、販売プロジェクトおよび技術調査に使用されているため、サーバ分野で最も重要なベンチマークとなっています。

SPEC CPU2017 は、大量の整数演算および浮動小数点演算を使用する独立した 2 つのテストセットで構成されています。整数演算部分は商用アプリケーションに相当し、10 種類のベンチマークから構成されます。浮動小数点演算部分は科学アプリケーションに相当し、10 または 13 種類のベンチマークで構成されます。いずれの場合も、ベンチマークの実行結果は、個々の結果の幾何平均です。

さらに、それぞれのテストセットには、単体実行時の処理性能を評価する速度測定と、並行処理の性能を評価するスループット測定があります。多数のプロセッサコアとハードウェアスレッドを持つサーバにとっては、後者が重要です。

また、測定の種類により、コンパイラーに許可される最適化が異なります。ピーク値の測定では、各ベンチマークを個別に最適化できますが、ベース値の測定では、コンパイラーフラグがすべてのベンチマークで同一である必要があり、特定の最適化は許可されません。

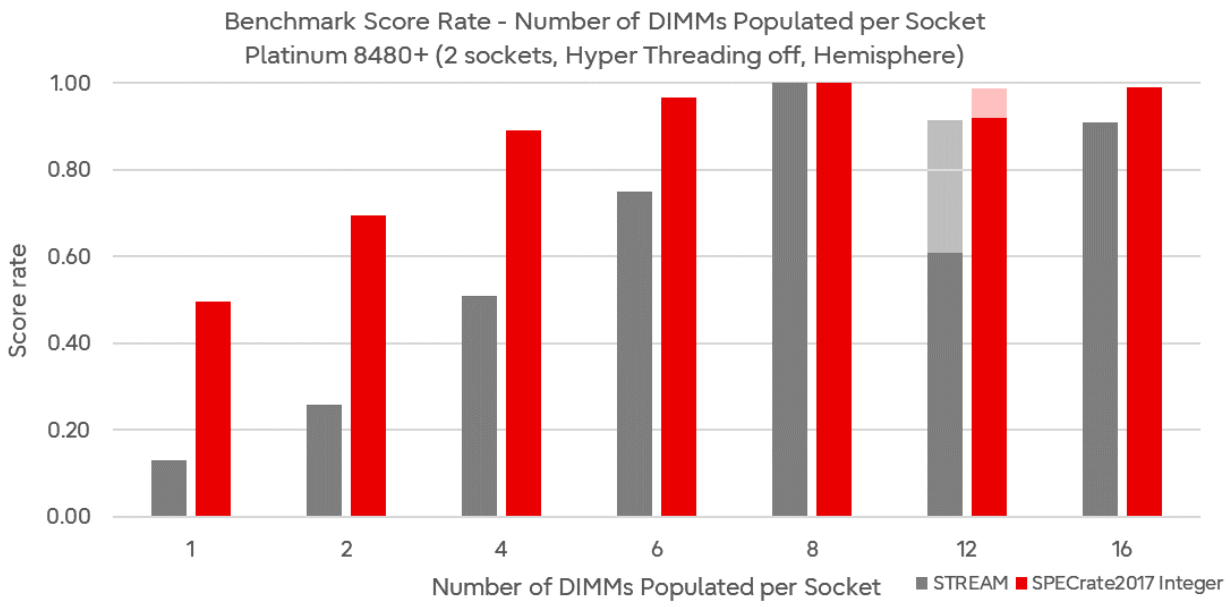
以上が SPEC CPU2017 概要です。PRIMERGY サーバでは商用アプリケーションの使用が主流であるため、整数演算を使用するテストセットである SPECrate2017 Integer でスループットを測定しました。

メモリチャネルへのインターリーブ

メモリチャネルへのインターリーブは、最初のブロックは最初のチャネルに、2 番目のブロックは 2 番目のチャネルにという具合に、プロセッサ単位で最大 8 つのメモリチャネルを交互に利用するように物理アドレス領域を設定する手法です。メモリアクセスは、局所性原理により主に隣接するメモリ領域に行われ、結果としてすべてのチャネルに分散されます。このようなパフォーマンスの向上は、並列化によるものです。

次の図は、同種類の DIMM をプロセッサあたり 8 枚単位で搭載せず理想的な 8 WAY インターリーブを行わない場合に、DIMM 枚数が 8 の場合を 1 としたパフォーマンスの比率を示しています。Sapphire Rapids 搭載 PRIMERGY 及び PRIMEQUEST サーバでは、1 プロセッサに搭載できる DIMM の枚数は 1、2、4、6、8、12、16 枚のみに制限されます。ここでは、SNC(Sub NUMA) が Disabled かつ UMA-Based Clustering が Hemisphere の設定での結果だけを記載しています。SNC(Sub NUMA) が Enable SNC2 または Enable SNC4 の場合、および、SNC(Sub NUMA) が Disabled かつ UMA-Based Clustering が Quadrant の場合の結果は、設定可能な DIMM の構成が限られること、および、性能比の傾向はほぼ同じであったことから、省略しています。

このテストに使用したシステムは、Xeon Platinum 8480+ 搭載 PRIMERGY RX2540 M7 です。使用した DIMM タイプは 4,800 MT/s 128 GB 4Rx4 3DS RDIMM です。これは、ワーキングセットを満たすメモリ容量を確保するために選択されました。



特にメモリのスループットを測定する STREAM の指標において、顕著な変化が見られます。DIMM 枚数が 8 枚以下の場合には、DIMM 枚数の増加に応じてパフォーマンスが向上しています。12 枚以上の場合、2DPC 構成によるメモリ転送速度の低下により、STREAM では 10% 程度の性能低下が見られます。

1 プロセッサに搭載する DIMM の枚数が 12 枚の場合は注意が必要です。1 番目のバンクに 8 枚、2 番目のバンクに 4 枚の DIMM を搭載する場合、前者の DIMM が構成する物理アドレス領域は 8 チャネルインターリーブ構成となり、後者の DIMM が構成する物理アドレス領域は 4 チャネルインターリーブとなります。そのため、アプリケーションがどの領域の上で動作するかで性能が変わります。上記グラフの 12 DIMM では、薄い色と濃い色の 2 つの棒グラフで示される 2 種類の結果が得られました。

SPECrate2017 Integer に関する評価は、商用アプリケーションのパフォーマンスに関するものです。STREAM で示されているように、メモリ帯域幅の関係は、特に HPC（High-Performance Computing：高性能コンピューティング）環境では、特定のア

プリケーション領域において除外できない極端なケースとして理解する必要があります。ただしこうした動作は、ほとんどの商用のワークロードでは見られません。STREAM および SPECrate2017 Integer に関する解釈の質は、このセクションで取り上げているパフォーマンス面だけでなく、以降のすべてのセクションにも当てはまります。
SPECrate2017 Integer パフォーマンスの低下が穏やかな 4-WAY、6-WAY インターリーブを選ぶ場合は、それなりの理由があるかもしれません。つまり、必要となるメモリ容量が少ないか、または低消費電力のために DIMM 数が最小限に抑えられるような場合です。1 WAY インターリーブは推奨できません。これは厳密に言うとインターリーブではなく、分類上そのように呼ばれているだけです。この場合、プロセッサとメモリシステムのパフォーマンスのバランスが取れていません。
また、12DIMM 構成での測定結果が示すように、不均等な DIMM の構成は、安定して性能を最大限に引き出す観点からは推奨できません。

メモリ転送速度

Sapphire Rapids 搭載 PRIMERGY 及び PRIMEQUEST サーバ、および、Emerald Rapids 搭載 PRIMERGY サーバでは、DPC によってメモリ転送速度が変化する場合があります。また、プロセッサの種類と BIOS パラメーターがメモリ転送速度に影響を与えます。節電設定（BIOS パラメーター DDR Performance で管理）は、実効メモリ転送速度が、そのプロセッサモデルでサポートされる最大メモリ転送速度より低くなってしまう原因になります。

次の表は、影響を比較し、バランスを取る際に役立ちます。表の数値は、理想的なケース、つまりそのプロセッサクラスで最大限の周波数に基づいています。

ベンチマーク	プロセッサ種類	最大メモリ転送速度	3,200 MT/s	4,000 MT/s	4,400 MT/s	4,800 MT/s	5,200 MT/s	5,600 MT/s
STREAM	Platinum 8570	5,600 MT/s	0.65					1.00
	Gold 6548N	5,200 MT/s	0.70				1.00	
	Platinum 8480+	4,800 MT/s	0.72			1.00		
	Gold 6430	4,400 MT/s	0.78		1.00			
	Silver 4416+	4,000 MT/s	0.89	1.00				
SPECrate2017 Integer	Platinum 8570	5,600 MT/s	0.92					1.00
	Gold 6548N	5,200 MT/s	0.95				1.00	
	Platinum 8480+	4,800 MT/s	0.92			1.00		
	Gold 6430	4,400 MT/s	0.96		1.00			
	Silver 4416+	4,000 MT/s	0.98	1.00				

このテストで使用したプロセッサモデルは、Xeon Platinum 8570（最大メモリ転送速度 5,600 MT/s）、Xeon Gold 6548N（最大メモリ転送速度 5,200 MT/s）、Xeon Platinum 8480+（最大メモリ転送速度 4,800 MT/s）、Xeon Gold 6430（最大メモリ転送速度 4,400 MT/s）、Xeon Silver 4416+（最大メモリ転送速度 4,000 MT/s）です。使用した DIMM タイプは 64 GB 2Rx4 RDIMM です。1DPC 構成で使用しています。

BIOS で「DDR Performance = Energy optimized（電力に最適化）」と設定すると、メモリ転送速度は常に 3,200 MT/s になります。

DIMM タイプの影響

Sapphire Rapids 搭載 PRIMERGY 及び PRIMEQUEST サーバ、および、Emerald Rapids 搭載 PRIMERGY サーバでは、最大 13 タイプの DIMM がサポートされています。ただし、特定のサーバで利用可能な構成については、それぞれのシステム構成図を参照してください。

次の表に、これらの DIMM タイプの同一条件下でのパフォーマンスの違いを示します。

- 測定は、Xeon Platinum 8480+（最大メモリ転送速度 4,800 MT/s）、Xeon Gold 6430（最大メモリ転送速度 4,400 MT/s）、Xeon Platinum 8570（最大メモリ転送速度 5,600 MT/s）、Xeon Gold 6548N（最大メモリ転送速度 5,200 MT/s） を使用して実施しました。
- これらの測定では、すべてのメモリチャンネルが同じ構成、すなわちパフォーマンスモード構成での比較が行われました。取り付けられた DIMM の数は、1DPC の測定では 16、2DPC の測定では 32 でした。
- すべての測定は、プロセッサの最大メモリ転送速度で実施されました。例えば、Xeon Platinum 8480+ の 1DPC 構成では 4,800 MT/s、2DPC 構成では 4,400 MT/s で動作し、Xeon Gold 6430 の場合は、DPC に関係なく 4,400 MT/s で動作しました。
- 測定に要するメモリ容量の制限により、Xeon Platinum 8480+、Xeon Platinum 8570 の測定は Hyper Threading = disabled の設定で実施しました。
- この表では、現時点で最善のパフォーマンスを提供するであろう 64 GB 2Rx4 RDIMM を使用した 1DPC 構成（太字で強調表示）を基準としています。実現されるメモリ容量が十分である場合に限り、ベンチマークではこの DIMM が推奨されます。

Xeon スケーラブル・プロセッサ搭載 PRIMERGY 及び PRIMEQUEST サーバ 最大メモリ転送速度 4,800 MT/s 以下の Sapphire Rapids および Emerald Rapids での性能比						
			Platinum 8480+ (最大 4,800 MT/s)		Gold 6430 (最大 4,400 MT/s)	
DIMM タイプ	構成	チャネルあたりランク数	STREAM	SPECrate 2017 Integer	STREAM	SPECrate 2017 Integer
16GB (1x16GB) 1Rx8 DDR5-4800 R ECC[4]	1DPC	1	0.86	0.96	0.87	0.98
	2DPC	2	0.91	0.99	0.97	1.00
32GB (1x32GB) 2Rx8 DDR5-4800 R ECC[4]	1DPC	2	1.00	1.00	1.00	1.00
	2DPC	4	0.85	0.98	0.93	1.00
32GB (1x32GB) 1Rx4 DDR5-4800 R ECC	1DPC	1	0.85	0.97	0.87	0.98
	2DPC	2	0.91	0.99	0.97	1.00
64GB (1x64GB) 2Rx4 DDR5-4800 R ECC	**1DPC**	**2**	**1.00**	**1.00**	**1.00**	**1.00**
	2DPC	4	0.85	0.98	0.92	1.00
128GB (1x128GB) 4Rx4 DDR5-4800 3DS R ECC	1DPC	4	0.94	0.99	1.00	0.99
	2DPC	8	0.86	0.97	0.91	0.98
256GB (1x256GB) 8Rx4 DDR5-4800 3DS R ECC	1DPC	8	0.99	0.99	0.99	0.99
	2DPC	16	0.73	0.96	0.81	0.98

[4] PRIMEQUEST サーバではサポートされていません。

Xeon スケーラブル・プロセッサ搭載 PRIMERGY サーバ 最大メモリ転送速度 4,800 MT/s 超の Emerald Rapids での性能比						
			Platinum 8570 (最大 5,600 MT/s)		Gold 6548N (最大 5,200 MT/s)	
DIMM タイプ	構成	チャネルあたりランク数	STREAM	SPECrate 2017 Integer	STREAM	SPECrate 2017 Integer
16GB (1x16GB) 1Rx8 DDR5-5600 R ECC	1DPC	1	0.83	0.99	0.84	0.98
	2DPC	2	0.83*	0.99*	0.86*	0.99*
32GB (1x32GB) 2Rx8 DDR5-5600 R ECC	1DPC	2	1.00*	1.00*	1.00*	1.00*
	2DPC	4	0.78*	0.98*	0.83*	0.98*
32GB (1x32GB) 1Rx4 DDR5-5600 R ECC	1DPC	1	0.82*	1.00*	0.84*	0.98*
	2DPC	2	0.83*	0.99*	0.85*	0.99*
64GB (1x64GB) 2Rx4 DDR5-5600 R ECC	**1DPC**	**2**	**1.00**	**1.00**	**1.00**	**1.00**
	2DPC	4	0.77*	0.97*	0.82*	0.98*
128GB (1x128GB) 4Rx4 DDR5-5600 3DS R ECC	1DPC	4	0.92	1.00	0.93	0.99
	2DPC	8	0.77	0.97	0.83	0.97
256GB (1x256GB) 8Rx4 DDR5-5600 3DS R ECC	1DPC	8	0.94	1.00	0.95	0.99
	2DPC	16	0.72*	0.97*	0.76*	0.97*

（*：推測値）

ここで示されているパフォーマンスの違いは、ランクインターリーブ数が異なることが主な原因です。ランクインターリーブ数はメモリチャネルあたりのランク数に等しく、DIMM タイプおよび DPC 値に従います。たとえば、表に示されているデュアルランク DIMM を使用した 1DPC 構成の場合は、2 WAY のランクインターリーブ、2DPC 構成の場合は 4 WAY のインターリーブが許容されます。

上の表で示されている通り、メモリチャネルあたりのランク数が 1 ランクの場合よりも 2 ランクの場合の方が、性能が高くなることが確認できます。16 GB 1Rx8 RDIMM や 32 GB 1Rx4 RDIMM の 1DPC 構成、すなわち、1-WAY ランクインターリーブでは、パフォーマンス低下が目立ちます。

また、2DPC 構成では、1DPC 構成時よりも性能が低下する場合が見られます。これは、2DPC 構成での最大メモリ転送速度が 1DPC 構成時よりも低下することによるものです。STREAM の性能低下は、この影響を大きく受けています。

プロセッサ内クラスタリングの設定

Sapphire Rapids プロセッサでは、プロセッサ内のクラスタリング設定として、SNC（Sub NUMA）に加えて、UMA-Based Clustering と呼ばれる設定があります。Sapphire Rapids 搭載 PRIMERGY サーバでは、これらの設定により、4 つのクラスタリングのモード、SNC4、SNC2、Quadrant、Hemisphere を選択できます。詳細については、メモリシステムの BIOS オプションに関するセクションを参照してください。

次の表は、本ドキュメントで実施した 2 つの負荷（ベンチマーク）の効果を示しています。

測定は、64 GB 2Rx4 RDIMM 1DPC 構成の PRIMERGY RX2540 M7 で行われました。

表には、1～4 パーセントの範囲でパフォーマンスに影響が出ることが示されています。この表を評価する際は、テストのセットアップ中の注意深いプロセスバインディングにより、両ベンチマークが極めて NUMA と相性が良いテストになっている点を考慮してください。したがって、一般の商用アプリケーションでは、この結果のような効果が得られない可能性があります。

ベンチマーク	プロセッサ種類	SNC4	SNC2	Quadrant	Hemisphere
STREAM	Platinum 8480+	1.03	1.01	1.00	1.00
	Gold 6430	1.04	1.02	1.00	1.00
	Silver 4416+	-[5]	1.03	-[5]	1.00
SPECrate2017 Integer	Platinum 8480+	1.03	1.02	1.01	1.00
	Gold 6430	1.03	1.01	1.00	1.00
	Silver 4416+	-[5]	1.01	-[5]	1.00

これらのクラスタリングモードを選択するための BIOS 設定値は以下の通りです。

クラスタリングモード	SNC(Sub NUMA)	UMA-Based Clustering
SNC4	Enable SNC4	-[6]
SNC2	Enable SNC2	-[6]
Quadrant	Disabled	Quadrant (4-clusters)
Hemisphere	Disabled	Hemisphere (2-clusters)

[5] XCC モデルの Sapphire Rapids プロセッサの場合のみ、選択できます

[6] SNC(Sub NUMA)を Enable SNC4 または Enable SNC2 に設定した場合、選択できません

リモートメモリへのアクセス

前述の STREAM および SPECrate2017 Integer ベンチマークを使ったテストでは、ローカルメモリのみが対象になっていました（プロセッサが自身のメモリチャネルの DIMM モジュールにアクセスする）。隣接するプロセッサのモジュールはまったくアクセスされないか、まれに UPI リンクを経由してアクセスされるのみです。実際のアプリケーションにおいて、オペレーティングシステムやシステムソフトウェアの NUMA サポートによってアクセスされるメモリの大半がローカルメモリである限り、この状況は代表的なものであると言えます。

次の表は、最大メモリ転送速度で動作する 64 GB 2Rx4 RDIMM 1DPC 構成という理想的なメモリ構成でありながら、BIOS 設定が「NUMA = Disabled」に設定されている場合の影響を示しています。統計的にメモリアクセスの半数がリモート DIMM、つまり隣接プロセッサに割り当てられた DIMM に対して行われ、データが UPI リンク経由で迂回する必要があるために、パフォーマンスの低下が発生します。特に、UPI リンクが 2 本の Silver 4416+ プロセッサでは、STREAM のような帯域幅が重視されるベンチマークで顕著に性能が低下しています。

ベンチマーク	プロセッサ種類	UPI リンク数	NUMA = Enabled	NUMA = Disabled
STREAM	Platinum 8480+	4	1.00	0.68
	Gold 6430	3	1.00	0.66
	Silver 4416+	2	1.00	0.58
SPECrate2017 Integer	Platinum 8480+	4	1.00	0.88
	Gold 6430	3	1.00	0.90
	Silver 4416+	2	1.00	0.92

NUMA = Disabled では、プロセッサの詳細メッシュ切り替えによって物理アドレス空間を設定しています。この切り替えは、両プロセッサが同一のメモリ容量であることを前提にしています。こうした一般的な状態が存在しない場合、アドレス空間はソケット間インターリーブが許可される主要部分と、プロセッサ - ローカルの残りの部分に分割されます。

NUMA = Disabled に関する測定は、システムソフトウェアまたはシステム関連ソフトウェアで NUMA がサポートされていないか、または十分でないために設定が推奨される例外的なケースとして、狭い範囲で実施しました。上記の測定はすべて、大部分あるいはすべてのアクセスがリモートメモリに対して行われる場合の影響を見積もる場合に役立ちます。この状況は、プロセッサごとの構成メモリ容量が大幅に異なる場合に発生します。ローカルアクセスと比較したパフォーマンスの低下は、表に示した低下分の最大 2 倍になることがあります。

冗長性、信頼性を考慮した際のメモリパフォーマンス

ここでは、システムが正常時の冗長性、信頼性のオプションの性能への影響を評価します。

ミラーリングではプロセッサの 1 つのメモリコントローラ内で　2 つのメモリチャネルの間でミラーが構成されます。オペレーティングシステムは、実際に構成されているメモリの 50 %を利用できます。

ADDDC スペアリングの場合、DIMM の予備領域を使用して 故障 DRAM セルを置きかえるため、容量の減少はありません。

次の表では、最大メモリ転送速度で動作する 64 GB 2Rx4 RDIMM 1DPC 構成という理想的なメモリ構成で、冗長オプションが有効化されている場合の影響を示しています。表の各列は、デフォルト設定、および、BIOS パラメーター Memory Mode、ADDDC Sparing のオプション設定に対応しています。

ミラーリング下で発生する性能低下は、デフォルト時の性能の 50% よりも小さくなります。これは、ミラーの半分がどちらも読み取りアクセスで使用できるためです。ADDDC スペアリングでは、機能を有効にすることによる性能低下はみられません。

ベンチマーク	プロセッサ種類	デフォルト	ミラーリング	ADDDC スペアリング
STREAM	Platinum 8480+	1.00	0.72	1.00
	Gold 6430	1.00	0.73	1.00
	Silver 4416+	1.00	0.82	1.00
SPECrate2017 Integer	Platinum 8480+	1.00	0.95	1.00
	Gold 6430	1.00	0.97	1.00
	Silver 4416+	1.00	0.98	1.00

関連資料

PRIMERGY サーバ

https://www.fsastech.com/products/pcserver/

PRIMEQUEST サーバ

https://www.fsastech.com/products/mission-critical/

メモリパフォーマンス

このホワイトペーパー

https://docs.ts.fujitsu.com/dl.aspx?id=fec08359-b897-435c-96ff-b2bd0daabbfc

https://docs.ts.fujitsu.com/dl.aspx?id=c2f30fb8-a486-4934-a773-b76b18c5d407

過去のホワイトペーパー

Xeon スケーラブル・プロセッサ（Ice Lake）搭載システムのメモリパフォーマンス
https://docs.ts.fujitsu.com/dl.aspx?id=74e595de-344f-47a6-8995-fe340733dfbb

Xeon スケーラブル・プロセッサ（Cascade Lake-SP）搭載システムのメモリパフォーマンス
https://docs.ts.fujitsu.com/dl.aspx?id=ade521ff-45c7-408d-9b36-a88b248497ca

ベンチマーク

STREAM
https://www.cs.virginia.edu/stream/
SPECcpu2017
https://docs.ts.fujitsu.com/dl.aspx?id=0f641c7e-bb5e-45e4-854f-cdd31faf5343

BIOS 設定

4th および 5th Generation Xeon スケーラブル・プロセッサ搭載システムのための BIOS 最適化
https://docs.ts.fujitsu.com/dl.aspx?id=cbcdbe16-0a73-49ce-8765-355602fb16d1

PRIMERGY のパフォーマンス

https://jp.fujitsu.com/platform/server/primergy/performance/

文書変更履歴

版数	日付	説明
1.1	2024-07-30	第 5 世代 Xeon スケーラブルプロセッサ（Emerald Rapids）の記載を追加 対象機種を追加
1.0	2023-07-04	初版

お問い合わせ先

エフサステクノロジーズ株式会社

Web サイト: https://www.fsastech.com

PRIMERGY のパフォーマンスとベンチマーク

mailto:fj-benchmark@dl.jp.fujitsu.com

