

Windows上でのCaché共有メモリの割り当て

2011年9月7日

インターシステムズ・カスタマーサポート Ray Fucillo

Caché を起動すると、Caché は、データベース・キャッシュ (グローバル・バッファ)、ルーチン・キャッシュ、共有メモリ・ヒープ、ジャーナル・バッファなどの制御構造に使用される単一の大規模なメモリ・チャンクを割り当てます。このメモリは「共有メモリ」と呼ばれます。すべての Caché プロセスが同時にこのメモリにアクセスできるように、特別な方法で割り当てられるためです。Caché はすべてのプラットフォームおよびバージョンで共有メモリを使用しますが、Windows 特有とも言える特定の課題が存在します。

Windows の共有メモリは、オプションで「ラージページ」を使用して割り当てることができます。この機能を使用すると、Windows は共有メモリを 1 MB 以上のページ・コレクションとして編成します。ラージページの使用に当たっては利点と欠点があり、それらは基本的にシステムの構成や管理方法に影響を及ぼします。

このドキュメントではそれらの問題について説明し、最適なアプローチを選択するうえで役立つガイドラインを示します。ここで説明する問題は、Windows 上のあらゆるバージョンの Caché に関連するものですが、ラージページを利用するには Caché 2007.1 以降が必要となります。

ラージページを使用する理由

ラージページは、以下の 3 つの利点をもたらします。

1. 一部の環境では、ラージページを使用することで、Caché の何百何千もの同時プロセスを実行するマシン上にページ・テーブル・エントリを保存するのに必要とされる物理メモリ (仮想メモリではありません) の量を大幅に削減することができます。これに関する詳細はすぐ後で説明します。
2. ラージページを使用することで、メモリ・アクセスの効率がある程度向上します。厳密には、メモリへの参照が現在のトランスレーション・ルックアサイド・バッファ内に含まれる可能性を高めることで実現されます。
3. ラージページを使用することで、Windows 2003 64 ビット版で 2 GB を超える共有メモリを使用するとシステムが基本的に急停止する原因となる問題を回避でき

まず (Windows 2008 ではこの問題はありません)。これは、当社が 2010 年 3 月 15 日に発行した[推奨事項](#)「Performance Degradation on Windows Server 2003 64-bit (Windows Server 2003 64 ビット版でのパフォーマンス低下)」で取り上げた問題です。

上記の最初の利点には説明が必要ですが、内容はやや複雑です。Caché プロセスの開始時、Windows は、共有メモリを番地付けするページ・テーブル・エントリ (PTE) のためにプロセスの仮想アドレス空間のかたまりを割り当てます。64 ビット Windows 上の PTE は 8 バイトであり、共有メモリがラージページとして割り当てられるかどうかにかかわらず、4 KB の共有メモリごとに 1 つの PTE が必要になります。つまり、すべてのプロセスで 1 GB の共有メモリごとに 2 MB の仮想メモリが要求されることになり、システム上で 512 プロセスごとに共有メモリ自体と同じ量の仮想メモリが PTE に割り当てられます。

ここで、ラージページの主な利点をもたらされますが、それは必ずしもユーザが考えるものであるとは限りません。ラージページを使用する場合、Windows は各プロセス内の仮想メモリを、4 KB の「スモールページ」それぞれのアドレスを指定するのに必要なすべての PTE に割り当てますが、実際にはそのメモリを使用しないため、それらの PTE が物理メモリを消費することは一切ありません。ページファイル内に十分な空き領域が必要なだけです (ページファイル領域は、Windows 上での仮想メモリの割り当てごとに確保されます)。ただし、共有メモリがラージページとして割り当てられない場合、それらの PTE は物理メモリを消費しません。各プロセスは PTE を使用して共有メモリの増加部分を参照するので、PTE が物理メモリ内で次第に認識されるようになります。プロセスの実行時間がかかり長くかつビジー状態の場合、そのプロセスが (例えば、データベース・キャッシュ内のさまざまなブロックを参照するなどして) ほとんどの共有メモリを参照することもあり得るため、実メモリがその PTE のほとんどに割り当てられる可能性があります。多数のプロセスで同様のことが起こった場合、ラージページが使用されない限り、PTE に対して物理メモリが大量に使用されることになります。

ここで数字を当てはめてみましょう。2,000 のプロセスを実行し、16 GB の共有メモリを割り当てるシステムがあるとします。ラージページが使用されるかどうかにかかわらず、各プロセスは、仮想メモリ内で 32 MB を PTE に割り当てます (4 KB の共有メモリごとに 8 バイト)。2,000 のプロセスでは、それは PTE に対する 64 GB の仮想メモリと、共有メモリ自体に対する元の 16 GB の物理メモリとなります。つまり、共有メモリにおよそ 80 GB のページファイル領域と 16 GB の物理メモリが必要となるわけです。ラージページが使用された場合はそれで済みますが、使用されない場合はさらにその 64 GB

の PTE の一部が実メモリを消費することになります。残念ながら、実メモリとなる PTE の一部は予測することが難しく、アプリケーションや使用パターンによって異なります。2,000 のプロセスそれぞれの実行時間が長く、共有メモリのあらゆるページを使用する最悪のケースでは、64 GB の PTE すべてに物理メモリが必要となることも考えられます。多くの物理メモリが使用できる場合を除き、システムはそのページファイルにプロセス・メモリを切り替えることが必要となり、パフォーマンスに深刻な影響が及ぶ可能性があります。

一方、ラージページを使用する場合、PTE がラージページそれぞれのアドレスを指定するのに必要な実メモリが多少必要となるものの、その量は取るに足りないほどわずかです。2 MB のラージページごとに 8 バイトの「ラージページの PTE」が 1 つ必要となるので、上記の例で言うと、16 GB の共有メモリのアドレスを指定するうえで 2,000 のプロセスに対して必要な実メモリはたった 128 MB ということになります。

ラージページの問題点

ラージページの欠点として、Windows がかなりの量の処理を実行した後では取得が難しくなることが挙げられます。これはメモリの断片化が原因であり、Windows がしばらくの間実行された後で Caché が起動される場合や、Windows を再起動せずに Caché が再起動される場合に問題となる可能性があります。これらのケースでは、おそらくラージページの割り当ては単に失敗するでしょう。割り当てが失敗した場合、Caché はスモールページを使用したり、共有メモリのサイズを減らしたり、あるいはその両方を行います。Caché バージョンごとに使用されるアルゴリズムは多少異なりますが、Caché 2010.2.4 以降で使用されるアルゴリズムでは、共有メモリ・サイズの低減よりもスモールページの割り当てが優先されます。

その他の同様のケースでは、メモリを割り当てるための Windows システム・コール CreateFileMapping() で何分も時間がかかる場合がありますが、最終的には正常に完了します。ちなみに、この動作は Windows 2008 R2 でより一般的なようですが、以前のバージョンでは単に割り当てに失敗する可能性が高いようです。当社では、この割り当てが 1 分間に 2 GB の低速で進行することを確認済みです。

この低速な割り当てによって Caché の起動が遅くなりますが、それが原因で起動時間が 3 分を超える場合、Caché サービスのタイムアウトが過ぎて、Caché が完全には起動されなくなります (回復には、Windows の再起動が推奨されます)。

Microsoft 社では、メモリが断片化された場合にラージページの割り当てパフォーマンスを向上する方法として、[KB 2532917](#) に示されている手順を推奨しています。当社では、この設定を採用することで、ラージページの割り当て時間が 3 分の 1 に短縮されることを確認済みです。つまり、最悪のケースを想定した当社の現行テストにおいて、結果的にパフォーマンスが 1 分間につき 2 GB から最大 6 GB まで向上しました。

一方、スモールページにおけるメモリの割り当てでは、ほぼすべてのケースで問題なく短時間で処理が成功しています。

推奨事項

すべての Windows システムにおいて、ページファイルのサイズは、Caché 共有メモリとすべてのプロセスのページ・テーブル・エントリ (PTE) を合わせたサイズに対応するよう設定される必要があります。前述のとおり、PTE に必要なページファイルの容量はほぼ以下の値と等しくなります。

共有メモリのサイズ×プロセス数÷500

ラージページの使用は、ほとんどの実運用システムに対して推奨されます。また、2010 年 3 月 15 日に発行した[推奨事項](#)「Performance Degradation on Windows Server 2003 64-bit (Windows Server 2003 64 ビット版でのパフォーマンス低下)」で説明した問題を回避するためには、2 GB を超える共有メモリを使用する Windows 2003 サイトにもラージページが必要となります。

テストの実施は重要であり、特に共有メモリのサイズが数 GB になる場合には欠かせません。なぜなら、お客様の特定の環境がラージページの取得に問題が生じる傾向にあるかどうかを判断するうえで、テストが唯一の方法だからです。ラージページの割り当て速度が遅い場合、Caché の起動に時間がかかるので目立ちます。共有メモリが失敗した場合は、Caché は通知することなく割り当て要求を変更して再試行します。そのため、結果を確認するうえで cconsole.log ファイルを監視することが重要になります。ラージページが割り当てられた場合のメッセージは「Allocated <n>MB shared memory (large pages): ...」で、スモールページが割り当てられた場合にはメッセージから「(large pages)」の文字がなくなります。例については、付録を参照してください。

Windows 2008 R2 以降で起動時に時間がかかる場合は、Microsoft 社の推奨事項 [KB 2532917](#) を使用することで時間を大幅に短縮できます。

ラージページを使用する際に共有メモリの構成サイズが大きい場合、最も信頼できる方法として、Windows の起動時に Caché を起動すると共に、Caché の再起動時には常に Windows も再起動するよう計画する方法があります。[推奨されるバックアップ方法](#)ではバックアップのために Caché をシャットダウンする必要がないため、これがバックアップに負担をもたらすことはありません。ただし、この方法はフェイルオーバー・クラスタ環境には該当しません。フェイルオーバー・クラスタ環境では、アクティブ・ノードの失敗の結果として、既に実行中のパッシブ・クラスタ・ノードが Caché を起動することが求められるためです。フェイルオーバー・クラスタ環境でラージページを使用するには、場合によっては、フェイルオーバーが発生するまでパッシブ・ノードで重要な処理が行われないようにする必要があります。また、フェイルオーバー・クラスタリングに代わる手法として、[Caché の監視](#)を検討することもできます。

環境によっては、Caché の起動に時間がかかる可能性を避けるためにラージページの使用を抑制することが適している場合もあります。そのような環境は、通常以下の特性をすべて備えています。

- ギガバイト単位の多くの共有メモリを必要とする構成
- Windows を再起動しないで Caché を起動する必要がある (フェイルオーバー・クラスタなど)
- 実行時間の長いプロセスの数が比較的少ないか、それらのプロセスの PTE に対応できる十分な余裕が物理メモリにある

Caché がラージページの割り当てを試みるには、Caché サービスによって使用されるアカウントに「Lock Pages in Memory (メモリ内のページのロック)」という権限が付与されている必要があります。Windows 2008 以降では、ローカル・システム・アカウントがその権限を暗黙的に持ちますが、Caché がその他のアカウントで実行される場合には、この権限を手動で追加する必要があります。ラージページの使用を抑制する場合は、この権限を削除する必要があります。また、Caché サービスがシステム・アカウントとして実行されるように設定する場合は、別のアカウントを構成する必要があります。これに使用されるアカウントは、ローカルの Administrators グループのメンバーでなければならず、Caché プロセスがドライブやプリンタなどのドメイン・リソースにアクセスできるようにドメインのメンバーとなることが一般的です。

付録: ログおよびスクリーン・ショットの例

以下の cconsole.log の行は、Caché がラージページのメモリを取得した場合を示しています。

08/08/11-14:28:21:234 (0) 0 Allocated 8636MB shared memory (large pages):
8192MB global buffers, 30MB routine buffers

以下の cconsole.log の行は、Caché がラージページを使用しないでメモリを取得した
場合を示しています。「(large pages)」の文字列がないことに注目してください。

08/08/11-14:33:57:078 (0) 0 Allocated 8634MB shared memory: 8192MB global
buffers, 30MB routine buffers

以下の cconsole.log の抜粋は、ラージページの割り当ての最初の試みに約 6 分間か
かり (14:55:03 ~ 15:00:59)、最終的に失敗したことを示しています。Caché はその後、
スモールページを使用して再試行し、1 秒以内に成功しています (3 分の時点で
Caché サービス・コントローラがタイムアウトし、エラーがログに記録されています)。

```
*** Recovery started at Mon Aug 08 14:55:03 2011
Current default directory: c:\intersystems\cache\mgr
Log file directory: c:\intersystems\cache\mgr
WIJ file spec: c:\intersystems\cache\mgr\CACHE.WIJ
Recovering local (c:\intersystems\cache\mgr\CACHE.WIJ) image journal file...
Starting WIJ recovery for 'c:\intersystems\cache\mgr\CACHE.WIJ'.
0 blocks pending in this WIJ.
Exiting with status 3 (Success)
08/08-14:58:02:094 ( 2304) 3 cctrl.dll (error during startup):(231) Cache failed to
start:- Cache Control Process did not fully initialize- shared memory.-Call
InterSystems Technical Support if you need assistance.
08/08/11-15:00:59:578 (0) 1 Failed to allocate 19054MB shared memory using
large pages. Switching to small pages.
08/08/11-15:00:59:609 (0) 0 Allocated 19054MB shared memory: 18000MB global
buffers, 200MB routine buffers
```

以下の 2 つのスクリーンショットは、Caché サービスが「管理者」として動作するよう構
成されていることと、Caché がラージページの取得を試みることができるよう、そのユー
ザ・アカウントに「Lock Pages in Memory (メモリ内のページのロック)」権限が付与されて
いることを示しています。

