様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年 6 月 13 日現在

機関番号：３２６２９
研究種目：基盤研究（C）
研究期間：2009～2011
課題番号：２１５００００２
研究課題名（和文）　大規模データ処理のための高速仮想メモリシステムの研究

研究課題名（英文）　The Study of Fast Virtual Memory Systems for Massive data Processing

研究代表者
緑川　博子（MIDORIKAWA　HIROKO）
成蹊大学・理工学部・情報科学科
研究者番号：００１９０６８７

研究成果の概要（和文）：１つのコンピュータのメモリには収まらないような大規模データを処理するために，高速ネットワークで結ばれた複数の遠隔のコンピュータのメモリを利用して，計算するコンピュータにあたかも大容量のメモリがあるかのようにみせかける仮想メモリを構築した．従来，メモリが不足する場合にはハードディスクにデータを展開して処理するしかなかったが，本研究により，従来手法に比べ数十倍から百倍以上，高速に大容量データ処理を行うことができるようになった．

研究成果の概要（英文）：This research realized the fast and large virtual memory by using remote memory distributed over nodes in a cluster, which has a high-speed network such as 10GbE and Infiniband. It enables us to use a cluster as a huge memory resource for sequential and multi-threaded applications requiring a larger amount of memory beyond available local memory. It achieved more stable and much higher performance compared to the traditional page swap system incorporated in OS kernel using local hard disks.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	2,200,000	660,000	2,860,000
２０１０年度	600,000	180,000	780,000
２０１１年度	700,000	210,000	910,000
年度			
年度			
総　計	3,500,000	1,050,000	4,550,000

研究分野：　総合領域　情報学
科研費の分科・細目：　計算機システム・ネットワーク
キーワード：ハイパフォーマンスコンピューティング，クラスタ，仮想メモリ

１．研究開始当初の背景

64bit の OS や CPU の普及により，今までとは桁違いに大きなアドレス空間（x86_64 現実装でも 256TB）が利用可能となってきた．巨大なアドレス空間は，データベースやバイオインフォマティクスのような大規模データを扱う応用にとって，従来の 32bitOS では扱えなかった大きなサイズのデータをプログラムで記述してメモリ上で処理できるようになるため，大きな恩恵があるが，それだけでなく，今までの小さなアドレス空間では実用になり得なかった大規模データを扱う新

しい応用へ道を開く可能性も秘めている.

しかし，１台のコンピュータで提供できる物理メモリにはスロット数などのハードウエア制約もあり，大容量物理メモリ搭載マシンは非常に高価になる.通常の汎用 OS では，ユーザの使うメモリサイズがローカル物理メモリサイズを超えると，予めシステム構築時などに定めたサイズのスワップ領域（通常，ローカルディスクのファイル）との間でページの出し入れを行って仮想メモリを実現する．したがって，このローカルディスクのスワップファイル領域を大容量化することにより，大データを扱うプログラムを実行することは原理的には可能であるが，実メモリにデータが全て収まる場合に比べ非常に低速になるため，多くの場合の実用に耐えない.

最近では，ローカルハードディスクへの入出力性能をしのぐ性能を持つ 10GbEthernet や InfiniBand などの高速ネットワークが出現している．これに伴い，大アドレス空間が利用可能でありながらローカルメモリサイズが制限されているときに，ネットワークに接続され遠隔ホストにあるメモリを逐次処理に利用できないかと期待できる.

本研究課題申請時において，ローカルメモリサイズを越えた大きなメモリを逐次処理で利用可能するための研究のほとんどは，遠隔メモリへのページングを目指しており，遠隔メモリをアクセスするための新しいネットワークブロックデバイスドライバを構築し，OS カーネルが利用するスワップデバイス（通常，ローカルハードディスク）を，この新しい遠隔メモリアクセス用のネットワークブロックデバイスに取り替えようという手法をとっている.

ユーザに完全に透過的であるカーネルレベルで，遠隔メモリを記憶階層の一部として組み込むというのが一つの理想像であるが，それには，現状のハードディスク特性を前提とし，小さいアドレス空間を想定して設計されたOSスワップ機構に，遠隔メモリ用のチューニングを施し，大アドレス空間を前提にしたメモリ管理を含めたカーネル全体の再設計が望ましい．しかし現状では，カーネルの変更を最小限に抑え，単にOSのスワップデバイスを上記の遠隔メモリアクセス用のブロックデバイスに交換する研究がなされている．特に最近の研究の幾つかでは，高速通信のための専用NICや専用高速プロトコル，RDMA機能，データのメモリへの事前登録などを用いて，ブロックデバイス手法における高速化の工夫をしている.

しかし，上記の様々な高速化手法を用いているにも関わらず，期待ほど通信は高速化されずに性能は低く，さらに動作不安定性が観測されて正常に稼動しないという報告が多くなされている．事実，32bitOSのアドレス空間範囲を超える大きなメモリサイズ（4GB 以上）のデータで，メモリアクセス負荷の高い応用プログラムに対する評価を，高速ネットワークを用いたクラスタで行った研究は，本研究課題申請時点で，筆者らの研究以外にはほとんどなかった.

筆者は前述の他の手法とは全く異なり，OS スワップ機構やスワップデバイスを代替するものではなく，OS のスワップとは独立に，完全にユーザレベルソフトウエアで，クラスタ上の複数の計算ノードの遠隔メモリを集めて，仮想的に大容量メモリとして利用する分散型大容量メモリシステム（DLM）を設計，構築した．評価の結果，多くの研究で行っているブロックデバイスドライバや高速専用 NIC，高速通信プロトコルなどを遠隔メモリアクセスに利用し，カーネルスワップデバイスを交換する他の低レベル実装方式に比べ，ユーザレベルソフトウエアと汎用的な TCP で実装した DLM が，より高い性能（3 倍～11 倍）

で，かつ動作安定性が高い（実行時間変動率は，DLM が 1%未満，ローカルハードディスクをスワップデバイスとする OS スワップ利用時が 2%～60%）ことが明らかになった．

この手法による遠隔メモリ利用は，物理メモリ枯渇という緊急状況（カーネルによるスワップ処理が発動された状態）で行われるのではなく，通常，メモリに多少の余裕がある正常実行状態においてプログラムが実行される．このため，通常，スワップデーモンは起動されず，処理負荷の高いスワップデーモンによるユーザプログラム性能への悪影響がない上，他の研究で見られるメモリ枯渇状態における動作不安定性やネットワーク通信上の様々なトラブルが生じない．

またカーネルスワップ機構と独立であるため，カーネル自体の従来のスワップパラメタや特性に支配されず，DLM のシステムパラメタを自由に設定することが可能で，遠隔メモリとのデータ交換サイズ，メモリ管理の単位などを，用いるネットワークや CPU に応じて最大限に性能を引き出すような値に設定することが可能で，低レベル通信プロトコルによる他手法より高性能を得た．

２．研究の目的

本研究の目的は，高速ネットワーク結合クラスタ上で遠隔メモリを利用する高速仮想大容量メモリシステムの実用に向けての研究を行う．未だ誰も経験したことのない大アドレス空間，大規模実メモリを用いる実際の応用を稼動させて，新たな知見と問題点を明らかにし，今後ますます巨大化する大規模データ処理応用の高速化を目指す．科学研究費の交付を希望する期間内に，すでに他手法に比べ性能が高いことが実証された前述の大容量仮想メモリシステムを用い，10Gbps 以上の高速ネットワークで結ばれた専有クラスタにおいて，大規模データを扱う実応用（バイオ関連処理，データベースなど）を稼動させ，性能や問題点などを明らかにし，今後の OS カーネル，メモリ管理，高速通信方式における改良や，新方式の提案を行う．

２．研究の方法

本研究では 10Gbps 以上の高速ネットワーク結合クラスタにおける遠隔メモリを利用する仮想大容量メモリシステムの実用に向けての研究を行う．初期性能評価実験で得た知見をもとに，実際の大規模データ利用応用に対する評価，解析を進め，高速通信方式や OS カーネルにおけるメモリ管理に関する現状の問題点も明らかにし，高性能で安定したシステムを実現する．また、多くの人が利用できる MPI バッチシステムでの運用や、クラスタをメモリ資源として利用できるような環境を構築するためのメモリサーバの自動割付などの機構を構築する．

DLM（分散大容量メモリシステム）は，すでに図１の DLM－S（シングルクライアント用メモリサーバシステム）のほかに図２の DLM－M（マルチクライアント用デーモン型常駐メモリサーバシステム）を設計，構築し，10GbpsEthernet や Myrinet などの高速ネット

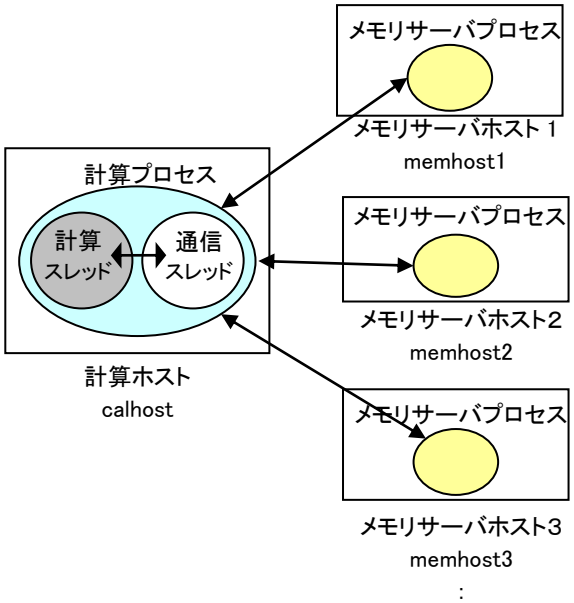

図１　DLM-S ランタイムシステム

ワークを用いたシステムでの性能評価を行う．さらに，複数のクラスタ群を接続した WAN 環境においても，計算負荷やメモリに余裕のあるクラスタ，計算ノード，メモリサーバノードを自動的に選定して，大容量メモリを用いることができるようにする．

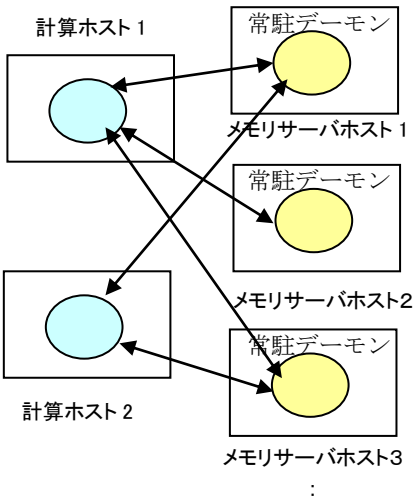

図２　DLM-M システム

３．研究成果

本研究では，遠隔メモリを用いた大容量メモリの実現にかかわる様々な方式設計，実装，評価を行った．以下に示す項目のような多くの成果を得ており，その成果論文は，国内外の多くの論文に引用されている．特に，本研究以前の国内の遠隔メモリ利用研究には大きな影響を与え，それまでの他組織における実装方式などは大きく変更されたと言っても過言ではない．さらに海外での評価も高く，最新内容で速報的に発表した自動適応型ページサイズ可変方式は，クラスタ・グリッド関連で権威ある国際会議においても賞を受賞している．

現在，ハイパフォーマンスシステムにおいては，多数ノード化，マルチコア化が進んでいるが，システム全体の消費電力の制限もあり，１ノードあたりのメモリ容量は，CPU 性能向上に比べむしろ相対的に減少傾向にある．したがってノード内のローカルメモリ容量の不足は依然，問題となり，今後，クラスタシステム全体のメモリの効率利用，従来のハードディスクに加え SSD などを代表とする不揮発性メモリ利用などを視野に入れた，多ノードにまたがる統合的なメモリ資源管理とメモリ階層管理がますます重要になってくると考えられる．そのような状況において，本研究成果は多くの方向性を示すものとなる．

(1) MPI バッチシステムクラスタにおける大規模データ逐次プログラム実行システム構築と評価

従来の TCP ソケットと 10GbpsEthernet を利用した DLM に加え，共同利用クラスタにおいて広く利用される MPI バッチシステムにおいて，一般ユーザがクラスタをメモリ資源として利用し，大規模データ処理応用を容易に実行できる DLM 利用環境を構築した．これに伴い，TCP ソケット通信による従来の DLM に加え、MPI 通信による DLM 実行方式を構築し，ページスワップにおける通信プロトコルも複数提案・実装して評価した．

(2) 高速ネットワーク接続クラスタにおける各種応用ベンチマークの性能評価

MPI の高移植性を生かし，最大 40Gbps（Myri10G ネットワーク４本のボンディング）のネットワーク性能で結ばれたクラスタ（東大 T2K，HA8000）において，各種応用（バイオ関連 cluster3.0，数値計算 Himeno ベンチマーク，NPB ベンチマークなど）の遠隔メモリ利用時の性能評価を行った．

(3) 分散仮想大容量メモリシステムにおける効率的なメモリ管理方式の構築

動的メモリ割り当て・解放機能を効率的に提供する，遠隔メモリ利用のための効率的なメモリ管理方式の設計・実装し，大容量のメモリを必要とする応用処理において，動的割り

当て（malloc）と解放（free）を利用する柔軟なプログラム環境を構築した．

(4) マルチクライアント向け・常駐型メモリサーバシステム DLM-M の構築・評価

常駐デーモンとして複数のメモリサーバを立ち上げ，このサーバに複数のユーザプログラムが随時接続して利用できる図２のような環境を構築した．さらに同一クラスタ内（LAN）で，負荷や提供メモリサイズを考慮したメモリサーバ自動割付システム（DLM-LAN）を設計・実装し，初期評価による有効性を確認した．

(5) LAN 接続ノードから適切なメモリサーバを選択する DLM-LAN の設計と構築

従来の DLM-M に自動メモリサーバ選定機能を導入した DLM-LAN では，１クラスタ内のメモリサーバノード群の中から，メモリ利用状況やクライアントサービス負荷などを考慮して，適切なメモリサーバを自動選択する LAN 管理プロセスを導入した．

(6) WAN 接続の複数クラスタ群から，適切なクラスタと計算・メモリサーバを自動選択して実行する DLM-WAN を設計，構築．

さらに，DLM-WAN では，WAN 接続の複数のクラスタ群において，WAN 管理プロセスを導入し，各クラスタの LAN 管理プロセスから情報を収集する．WAN 全体のクラスタの状況（計算負荷，メモリ利用状況）を考慮して，適切なクラスタと計算サーバ，メモリサーバを自動的に選択し，ユーザプログラムを実行するシステムを設計，構築した．また，利用者がどこからでも容易に大容量メモリを使うジョブを投入できるような web サーバと web インターフェースを構築し，ポータルサイトを利用してユーザホストからプログラムやデータなどを DLM へ投入することを可能とした．日本全国 17 か所以上の組織のクラスタを結ぶ Intrigger システムにおいて，DLM-WAN の稼働実験を行い，他のユーザの計算負荷，メモリ使用状況を考慮して，自動的にクラスタの計算サーバ，メモリサーバを選定して実行できることを確認した．

(7)遠隔メモリ利用可能な逐次 C プログラムのための DLM コンパイラの構築

DLM 用コンパイラは，既存の C 逐次プログラムで，大容量メモリを使う変数（配列など）の前に，図３に示すように，dlm という指定をユーザが行うだけで，ユーザには意識させずに，その変数を，ローカルメモリが不足する際に，遠隔メモリを利用して展開してくれるようにする．

従来の逐次プログラムでは，通常，大容量データは，スタックサイズの制限を回避するため，大域変数として定義することが多かった．DLM コンパイラでは，関数内部変数にも dlm 宣言を行うことを可能する新機能を追加した．これにより関数内部だけで大容量データを用いることも可能とした．

```
#include <stdio.h>
#define N 100000 //100k,total memory 10GB + 1.6MB

dlm double a[N][N], x[N], y[N];   // DLM 使用

int main(int argc, char *argv[])
{ int i,j;
   double temp;
   for(i = 0; i < N; i++)   //   行列 a を初期化
      for(j = 0; j <N; j++) a[i][j] = i;

   //    ベクトル x を初期化
   for(i = 0; i < N; i++)       x[i] = i;

   //   a[N][N]*x[N]=y[N]  計算
   for(i = 0; i < N; i++){
        temp = 0;
        for(j = 0; j <N; j++)    temp += a[i][j]*x[j];
        y[i] = temp;
   }
   return 0;
}
```

図３ DLM コンパイラ向けプログラム例
（行列・ベクタ積）
逐次プログラムとほぼ同等

(8) マルチスレッドプログラムに対応する遠隔メモリページングシステムの設計・構築・評価

従来のシングルスレッド逐次プログラムだけでなく、マルチコア・マルチCPUを有しているクラスタノードで実行されるマルチスレッドプログラムが，遠隔ノードの大容量メモリを使うことを可能にするマルチスレッド向け大容量分散メモリシステムを新たに設計構築し，既存のマルチスレッドライブラリへの対応，単一ノード内で実行される共有メモリ型並列プログラミング（単一メモリアドレス空間）にも対応するシステムを構築した．ユーザが書いたｐスレッドプログラムや，既存の数値計算ライブラリ（fftw）利用，NPBのOpenMP版のプログラム実行に対しても，遠隔メモリを透過的に利用できる環境を構築した．これによりマルチコアによるスレッド並列効果も得ることができた．

(9) MPIバッチシステムクラスタにおけるページスワッピングのための効率的通信実装方式の構築

MPI バッチシステムクラスタにおいて，遠隔メモリページ交換通信プロトコルを複数実装し，比較性能評価を行った．MPI スレッドサポートレベルに応じて，適切なものを利用できる．

(10) 適応型自動ページサイズ調整機構の設計と評価

計算ノードで利用可能なローカルメモリサイズと，応用プログラムのワーキングデータセットを考慮し，実行時に適切なページサイズを自動的に計測，調整する自動適応型ページサイズ制御方式を設計，実装した．繰り返しを含む応用に対して評価した結果，応用プログラムによるメモリアクセスローカリティの違いや，一つの処理含まれる各部分のメモリアクセス特性の違いにも対応して，性能に大きな効果をもたらすことを示した．最新情報をポスター論文で発表したところ，クラスタ・グリッド関連の主要国際会議であるCCGrid2012 でも評価されて賞を得ている．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 18　件）

① H.Midorikawa, J.Uchiyama:″Automatic Adaptive Page-size Control for Remote Memory Paging″, proc. of IEEE/ACM International Symp. on Cluster, Cloud and the Grid Computing CCGrid2012, pp.694-696, 2012-5, 査読有 The Best Poster Paper Award 受賞

② 鈴木悠一郎, 鷹見友博, 緑川博子：″マルチスレッドプログラムのための遠隔メモリ利用による仮想第容量メモリシステムの設計と初期評価″, 情報処理学会,ハイパフォーマンス研究会 Vol.2011-HPC-132, No.13, pp.1-6, アーキテクチャ研究会 Vol.2011-ARC-197, No.13, pp.1-6（2011.11）査読無

③ 古尾谷歩, 緑川博子,甲斐宗徳：″MPI を利用した分散大容量メモリシステムにおけるページスワッププロトコルの評価″,第 10 回情報科学技術フォーラム FIT 論文集(第一分冊),B-049, pp.361-364,（2011.9）査読無

④ 内山丞, 緑川博子,甲斐宗徳：″遠隔メモリアクセスのためのページスワップページサイズ動的変更機構の検討″, 第 10 回情報科学技術フォーラム FIT 論文集(第一分冊),B-050, pp365-367（2011.9）査読無

⑤ S. Yoshimura, H.Midorikawa, "A C Compiler for Large Data Sequential Processing using Remote Memory", proc. of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, pp.198-202, (2011.8) (DOI: 10.1109/ PACRIM.2011.6032892) 査読有

⑥ 吉村礎, 緑川博子："大容量データを扱うプログラムのための遠隔メモリ利用を容易にするＣコンパイラ", 成蹊大学理工学研究報告 ,第４８巻,第１号,pp.15-22 (2011.6) 査読無

⑦ 鈴木悠一郎, 緑川博子:"WAN接続クラスタ群をメモリ資源とする大容量メモリ提供システム", 成蹊大学 理工学研究報告,第４８巻,第１号,pp.23-30 (2011.6) 査読無

⑧ 鈴木悠一郎, 緑川博子:"WAN接続クラスタをメモリ資源として利用するためのメモリサーバ自動選定システム", 情報処理学会 第73回全国大会, 論文集,73-3 (2011.3) 査読無

⑨ 吉村 礎, 緑川博子, 甲斐宗徳："ローカルメモリを越える大容量データを扱う逐次処理のためのＣコンパイラ", 情報科学技術フォーラム FIT2010, FIT論文集, B-026 , pp.335 -336, (2010.9) 査読無

⑩ 齋藤和広, 緑川博子, 甲斐宗徳："ユーザレベル実装遠隔メモリページングシステムにおけるページ置換アルゴリズムの評価", 情報処理学会、ハイパフォーマンス研究会 Vol.2010-HPC-125, No.9, pp.1-6, (2010.6) 査読無

⑪ 鈴木悠一郎、緑川博子："分散大容量メモリDLMのWAN接続クラスタ群への適用 ー クラスタ・サーバ自動選定システムの提案 ー"、SACSIS2010, pp.173-174, (2010.5) 査読無

⑫ 緑川博子、齋藤和広、佐藤三久、朴 泰祐："クラスタをメモリ資源として利用するためのMPIによる高速大容量メモリ"、情報処理学会論文誌, コンピューティングシステム, Vol.2, No.4, pp.15-36, (2009.12) 査読有

⑬ 三浦望, 緑川博子, 甲斐宗徳:" クラスタをメモリ資源として利用するための動的メモリ提供システムの提案",情報科学技術フォーラム FIT2009, FIT論文集,B-029, pp.421-422, (2009.9) 査読無

⑭ H. Midorikawa, K.Saito, M.Sato, T.Boku: "Using a Cluster as a Memory Resource: A Fast and Large Virtual Memory on MPI", Proc. of IEEE cluster2009, (2009.9),pp.1-10 (DOI: 10.1109/CLUSTR.2009.5289180) 査読有

⑮ 緑川博子,斉藤和広,佐藤三久,朴泰祐:" クラスタをメモリ資源として利用するためのMPIに基づいた高速大容量仮想メモリ", 電子情報通信学会 CPSY研究会資料 CPSY2009-15,信学技報 Vol.109, No.168, pp.37-42 (2009.8) 査読無

⑯ K. Saito, H.Midorikawa, M.Kai,; "Page Replacement Algorithm using Swap-in History for Remote Memory Paging", proc. of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing,pp.533-538, (2009.8) (DOI: 10.1109/PACRIM.2009.5291315) 査読有

⑰ 齋藤和広, 緑川博子, 甲斐宗徳："遠隔メモリページングにおけるスワップイン履歴を用いたページ置換アルゴリズムの初期評価", 情報処理学会、ハイパ

フォーマンス研究会 Vol.2009-HPC-120, No.8, pp.1-6, (2009.6) 査読無

⑱ 齋藤和広, 緑川博子, 甲斐宗徳："遠隔メモリページングにおける各ページ固有のスワップ履歴を利用するページ置換アルゴリズム", SACSIS2009, pp.173-174, (2009.5) 査読無

〔学会発表〕(計 3 件)

① 内山丞, 緑川博子："遠隔メモリアクセスのためのスワップページサイズ自動調整機構の初期評価", ハイパフォーマンスコンピューティングと計算科学シンポジウム HPCS2012, HPCS2012 論文集, (2012.1/25) 愛知県, 名古屋大学

② 吉村 礎, 緑川博子："遠隔メモリ利用で大容量データ処理を可能にする逐次プログラムための C コンパイラ", ハイパフォーマンスコンピューティングと計算科学シンポジウム HPCS2011, HPCS2011 論文集, p.84 (2011.1/18) 茨城県, 産業技術総合研究所

③ 鈴木悠一郎, 緑川博子："WAN 接続クラスタをメモリ資源として利用するためのメモリサーバ自動選定システム －ウェブインターフェースによるユーザビリティの向上－", ハイパフォーマンスコンピューティングと計算科学シンポジウム HPCS2011, HPCS2011 論文集, p.85, (2011.1/18) 茨城県, 産業技術総合研究所

〔その他〕
ホームページ等
http://xserv0.ci.seikei.ac.jp/~midori/paper/index.html

６．研究組織
(1)研究代表者
緑川　博子 (MIDORIKAWA　HIROKO)
成蹊大学　理工学部　情報科学科
研究者番号：00190687