小口研究室　研究紹介　(2007年度)

(お茶の水女子大学理学部情報科学科)

IP-SAN統合型PCクラスタにおける並列相関関係抽出実行時のシステム特性解析(研究担当：原　明日香)

研究背景

情報量が爆発的に増大

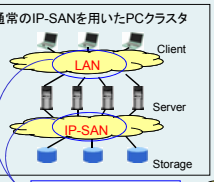
大容量のデータを格納し、高速にユーザにとって有益な情報を抽出することが必要

IP-SANの利用
サーバとストレージを結ぶ高速なネットワークであるSANをIPネットワークで接続

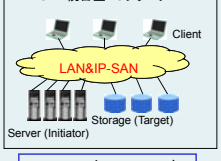
PCクラスタの利用
分散メモリ型並列計算機に汎用の計算機と汎用のネットワークを用いて接続

データマイニング
最も処理が重い相関関係抽出
Aprioriアルゴリズム、FP-growthアルゴリズム

通常のIP-SANを用いたPCクラスタ



IP-SAN統合型PCクラスタ



Front-endもBack-endもIPネットワークを使用

Front-endとBack-endを同じIPネットワークに統合

同じネットワークを使用するため、性能劣化の懸念

Ethernet等のコモディティなネットワークによる統一的な管理
ネットワーク構築コストと管理コストの削減

既存研究

ローカルデバイスを用いたPCクラスタ
IP-SAN統合型PCクラスタ上
・HPA
・PFP
をそれぞれ実行
どちらのアルゴリズムにおいても
同程度の性能

原因を解析
ネットワークに比べてローカルストレージの帯域幅が低いため、ネットワークに余裕があり、
IP-SAN統合型PCクラスタは有効

ストレージのディスクを強化した新しいクラスタの構築

実験内容と実験結果

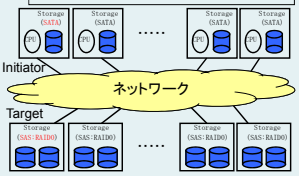
Aprioriアルゴリズムの並列化→HPA
FP-growthアルゴリズムの並列化→PFP

これまでよりストレージの性能を強化したIP-SAN統合型PCクラスタ上で実行

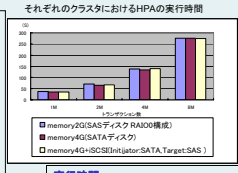
実行時間、ネットワークトラフィック、メモリ使用率をモニタリング

実験環境

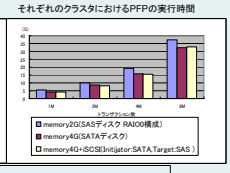
・8台
OS:CentOS4.5(linux2.6.18)
CPU:クアッドコア Intel Xeon 1.6GHz
Memory:2GB
HDD:73GB SAS×2
RAIDコントローラ:SAS5iR(RAID0)
接続ネットワーク:Gigabit Ethernet
・8台
OS:CentOS4.5(linux2.6.18)
CPU:Intel Xeon 1.6GHz
Memory:4GB
HDD:250GB SATA
接続ネットワーク:Gigabit Ethernet
・iSCSI
Initiator:open-iscsi-2.0.865
Target:iSCSI Enterprise Target(IET)-0.4.15



それぞれのクラスタにおけるHPAの実行時間



それぞれのクラスタにおけるPFPの実行時間



実行時間
HPA:どのクラスタも同程度の性能
PFP:基本性能測定で一番性能の良かったクラスタの性能がやや劣った

ネットワークトラフィック
どのクラスタにおいてもネットワークの帯域にはまだ余裕があり、iSCSIのトラフィックはネットワークにあまり大きな影響を与えていない

メモリ使用率
4GBのメモリを搭載しているノードがメモリの多くを消費している

HPA:本研究で用いたプログラムはディスクアクセスの性能やメインメモリの容量の影響を受けにくい
PFP:本研究で用いたプログラムはディスクアクセスの性能は受けにくいがメインメモリの影響は受けやすい

IP-SAN統合型PCクラスタは有効

VPN環境を用いたユビキタスデータ分散処理に関する検討(研究担当：甲山　絵梨奈)

研究背景と研究目的

ネットワークの高速化が進む中
複数の拠点に映像データを配信することは容易に実現可能

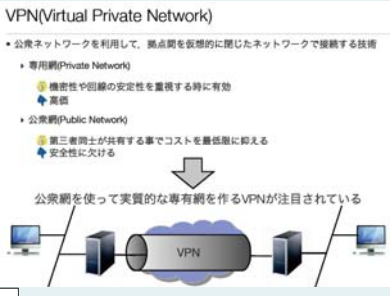
ビデオカメラなどのマルチメディアデバイスの普及が進展している中
複雑な動画処理のニーズも高まっている

・撮影された映像ストリームを加工し
そのデータをVPN経由で他ノードに送信したり
必要なDBへリモートアクセスするといった
システム環境を構築
・地理的に分散した複数の拠点でどこで
どのような加工処理を行う形が
最適であるかについて検討

VPN(Virtual Private Network)

・公衆ネットワークを利用して、拠点間を仮想的に閉じたネットワークで接続する技術
・専用網(Private Network)
機密性や回線の安定性を重視する時に有効
高価
・公衆網(Public Network)
第三者同士が共有する事でコストを最低限に抑える
安全性に欠ける

公衆網を使って実質的な専有網を作るVPNが注目されている



実験内容と実験結果

PC環境

hp	Dell
OS:Fedora Core5	OS:Fedora Core5
kernel:2.6.20-1.2320	kernel:2.6.20-1.2320
CPU:Intel Xeon 3.8GHz	CPU:Intel PentiumIII 796MHz
Main Memory: 4GB	Main Memory: 640MB

	ローカル	リモート
パターン1	動画キャプチャ・DBアクセス 計算処理・結果表示	—
パターン2	動画キャプチャ 計算処理・結果表示	DBアクセス
パターン3	動画キャプチャ 計算処理	DBアクセス 結果表示

左をローカル、右をリモートとし顔認識アプリfacedetectを実行

・facedetect
人の顔を認識して検出し円で囲うことにより示す

	左がdell・右がhpの場合 1フレームを表示にあたり		左がhp・右がdellの場合 1フレーム表示するにあたり	
	最大(ms)	最小(ms)	最大(ms)	最小(ms)
全ての処理をローカルで行う場合	1070.3	846.6	245.2	213.7
DBのみリモートに置く場合	989.5	809.1	227.7	200.4
DBと結果表示をリモートで行う場合	863.0	815.0	224.5	205.8

結果表示
DB照会
動画キャプチャ　→　ネットワークの影響は無視できる
計算処理　→　最も高負荷

PCクラスタを利用したバイオインフォデータマイニング(研究担当：島本　真衣)

研究背景

データマイニング
大規模なデータから思いがけないパターン、規則等を発見すること

相関関係抽出
大規模なデータから、あるパターンが現れる頻度を調べ、その頻度によって有意義なデータを抽出すること

Aprioriアルゴリズム
候補アイテムセットから頻出アイテムセットを生成し、繰り返し数え上げを行う

FP-growthアルゴリズム
巨大なデータベースから相関関係抽出に必要な情報をコンパクトに圧縮したデータ構造であるFP-treeを利用し、候補アイテムセットを生成せず再帰的に頻出アイテムセットを生成する

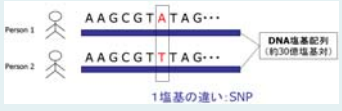
扱うデータは大規模であるため
並列化が不可欠

分散メモリ型並列計算機である
PCクラスタの利用

SNP (Single Nucleotide Polymorphism：スニップ)

DNA上で、それぞれの染色体が1塩基異なっている部をSNP(一塩基多型)と呼ぶ

人によって薬の効き方や副作用の出方が違うのはこのSNP(スニップ・個人差)が原因ではないかと考えられている



SNPデータのマイニング

被験者番号をトランザクションID、
SNPとなる塩基番号をアイテムとする

SNPデータをPCクラスタ上でマイニングし、人工的に作成されたトランザクションデータを用いてマイニングを行った際とのシステムの振舞の違いを考察

実験内容と実験結果

Aprioriアルゴリズムの並列化アルゴリズムHPA
FP-growthアルゴリズムの並列化アルゴリズムPFP

SNPデータの出力変換

SNPデータを人工的に作成されたトランザクションデータと同じ形式に変換

データの表示方法の変換出力されたテキストデータをバイナリデータに変換

並列処理を行うノード台数分にデータを分割

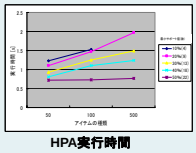
被験者番号(ID)
塩基番号(アイテム)
塩基番号(アイテム)
塩基番号(アイテム)
・
・
・
0
被験者番号(ID)
塩基番号(アイテム)
塩基番号(アイテム)
塩基番号(アイテム)
・
・
・

テストデータを作成し、実行時間を比較

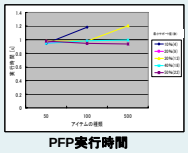
テストデータ実行結果

ノード数　4
トランザクション数　45
アイテム総数　65000

一台あたりのメインメモリが384MBのクラスタ上での2つのアルゴリズムの実行時間



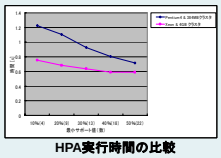
HPA実行時間



PFP実行時間

実行環境での比較

一台あたりのメインメモリが384MBと4GBのクラスタ上で実行した際の実行時間の比較



HPA実行時間の比較

oguchi@computer.org