



医学データ解析における 64bit版SASの有効性の検討

- Itaniumプラットフォーム (IA-64) を軸にして -

東京大学大学院 医学系研究科

土居 主尚

山口 拓洋



背景

■ 近年の医学研究


□ 解析対象が大きなデータ

- 変数の数や対象者数が大きい
- メモリ不足で解析自体が行えない

□ 解析手法の高度化

- 複雑なモデル
- 標本再抽出法(ブートストラップ法など)、マルコフ連鎖シミュレーション(MCMC)などの反復計算を前提とした解析手法

■ プログラムの効率化だけでは限界




SASが対応しているプラットフォーム

■ SAS Institute Japanのウェブサイト

<http://www.sas.com/offices/asiapacific/japan/software/sas9/sas9faq.html>

- SAS9が稼働するプラットフォーム(OS)について教えてください。
- SAS9対応プラットフォームは下記のとおりです。
Windows版: NT 4.0、2000、32bit版XP、32bit版2003、
64bit版2003
UNIX版: 64bit版 (AIX、Solaris、HP-UX、HP-IPF、Tru64)、
32bit版Linux



32bit ? 64bit ?

- Bitは情報の最小単位 (0か1のどちらか)
- コンピュータはbitの束をまとめて処理
 - 32個のbitを一度に扱えるコンピュータ
 - 64個のbitを一度に扱えるコンピュータ
- コンピュータのbit数とソフトウェアのbit数は一致
 - 互換性の問題から現状は混在
 - コンピュータのbit数の定義にも議論あり
- 通常目にするコンピュータ、ソフトウェアは32bit



64bit版SAS

- 64bit版SASとは？
 - 64bitのコンピュータ上で動くSAS
- EM64TやAMD64上で動くSASのこと？
 - SASが対応していない
- Itanium2上で動く64bit版SAS
 - 64bit Windows Server 2003上で動作

Itanium2とは

- 新しく開発された64bitのプロセッサ
IA-32 (Pentium4等) の延長線ではない
- 様々な高速化技術を実装
 - 豊富な内部リソース
 - 浮動小数点演算の高速化
 - 大規模なメモリを搭載可能
IA-32: 4GB、IA-64: 512GB以上
- コストは高い





解析の現場

- 個人のパソコンにてシミュレーション
 - 1回の計算ですら数日かかることも
 - さらに複数の設定でシミュレーションを行う場合
- RAM DISKを利用したSAS処理時間短縮の試み
 - Iwamoto and Hirata. (2004). 関西SASユーザー会.
 - 例数設計等でのシミュレーションの際、非常に時間がかかる
 - 夜間バッチで途中でダウンすることも、しばしばあった



Itanium2で解析の高速化

- 先行研究はほとんどなし
 - Itanium2 (IA-64) は特殊な分野で認知
- 様々な高速化技術を実装したItanium2
 - 時間のかかる解析が短時間で終わる
- コストは高いが導入する意義
 - SASはプラットフォームが異なっても同じプログラムが実行可能
- 一般にソフトウェアの実行速度は最適化にも依存
 - 実際に動かしてみないと分からない



目的

- 64bit版SASを用いた場合、従来の32bit版SASと比較して医学データ解析におけるパフォーマンスがどれだけ向上可能か
- どのような状況において64bit版SASを使用するのが有用か



コンピュータのスペック

- IA-64

CPU: Itanium 2 1.5GHz (L3キャッシュ6MB) 2基

メモリ: 10GByte

Windows Server 2003, Enterprise Edition
for Itanium-based systems

- IA-32

CPU: Pentium 4 2.6GHz (L3キャッシュなし) 1基

メモリ: 512MB

Windows XP Professional

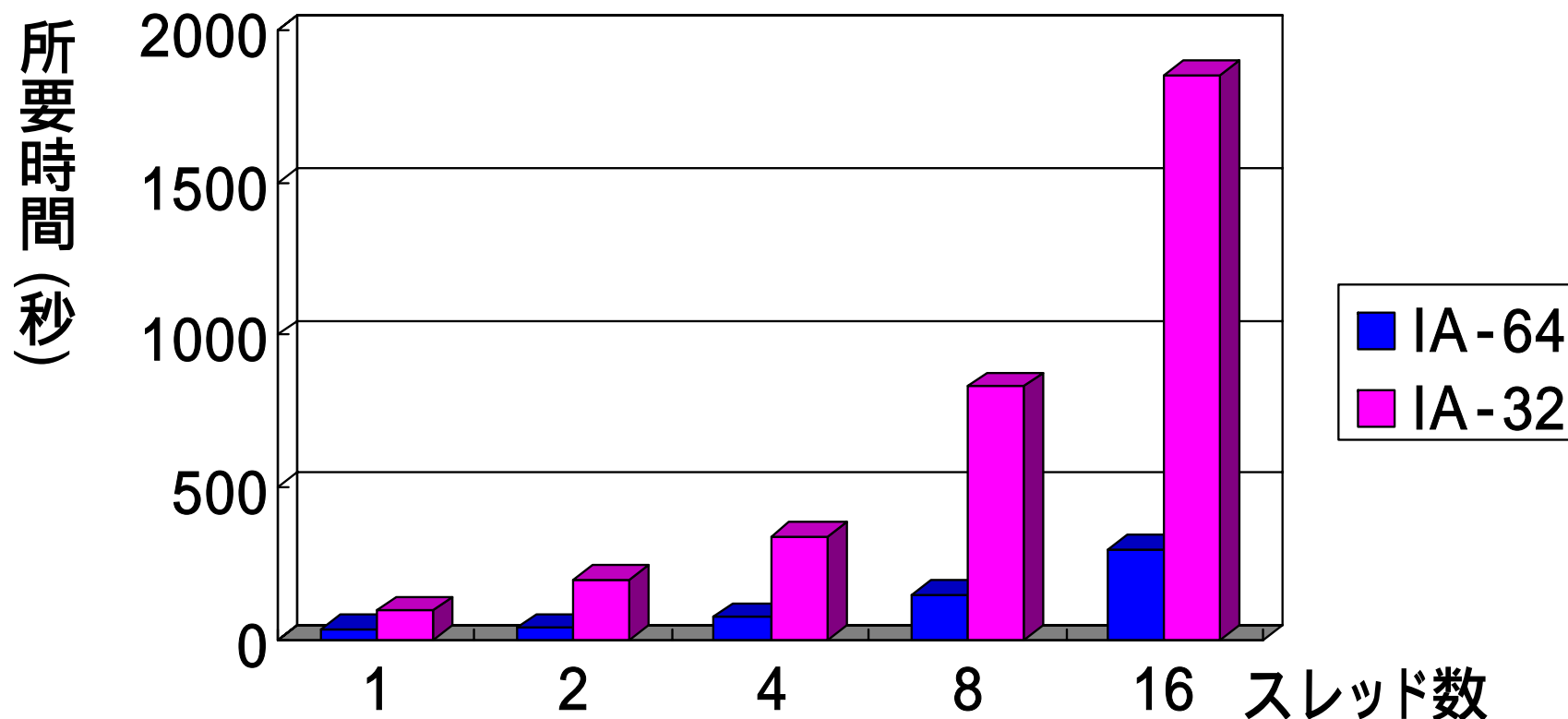
- それぞれに対応するWindows版SAS 9.1
(64bit版2003、32bit版XP) を用いた



比較した項目（一部）

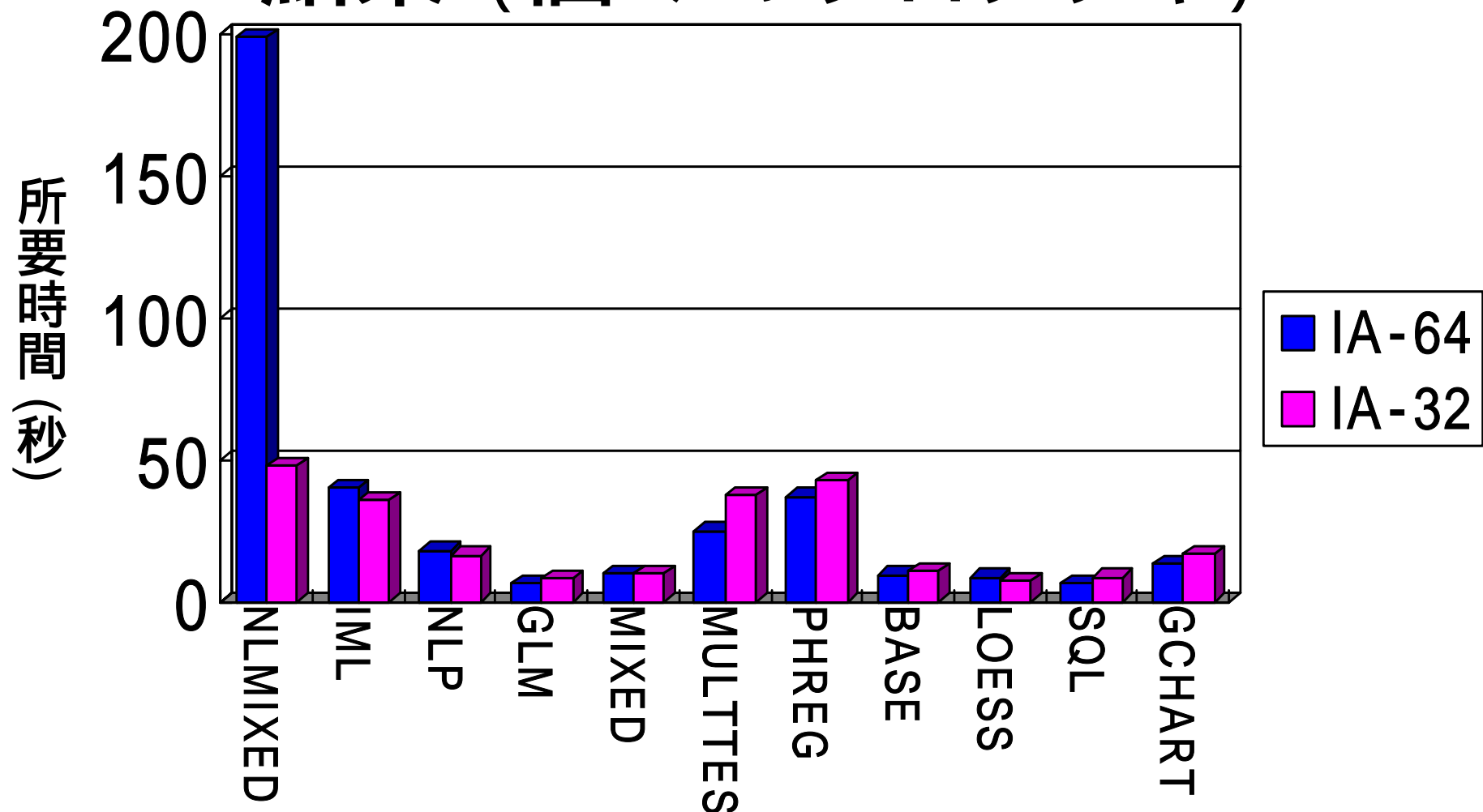
- マルチスレッド
 - 複数SASを起動させ、複数のプログラムが同時に走る状況の性能評価
 - (IMLでデータ発生、MIXEDプロシジャで解析の一連の流れの繰り返し)を複数同時平行
- 個々のプロシジャの速度比較
 - SASの代表的なプロシジャで比較
 - ヘルプのサンプルプログラムをベースに作成
- データステップ、プロシジャを組み合わせた手順
 - ブートストラップ法
- 大規模なメモリが生かされる例
 - IMLで作成可能な行列の最大サイズの比較

結果 (マルチスレッド)




- 1スレッドではIA-64がIA-32より倍以上速い
- IA-64は2基のプロセッサのため1スレッドと2スレッドが同じ
- IA-64は2スレッド以降、スレッド数の増加に比例して時間の増加
- IA-32はスレッド数の増加を上回る速度で時間の増加

結果（個々のプロシジャ）



- NLMIXEDにてIA-64が極端に遅い
- MULTTESTでIA-64が少し速い以外、大きく変わらず



結果（ブートストラップ）

- SASのウェブサイトに登録されているマクロ
<http://support.sas.com/ctx/samples/index.jsp?sid=479>
proc reg、ブートストラップサンプルは100000など
- 実行時間を計測
 - IA-64: 4分35秒
 - IA-32: 3分50秒



結果（大規模なメモリ）

- 作成できる行列の最大サイズ
 - IA-64 16300x16300の正方行列
 - IA-32 9400x9400の正方行列
- 最大サイズの行列の作成可能個数
 - IA-64 16300x16300の行列を5個
 - IA-32 9400x9400の行列を1個のみ



考察（マルチスレッド）

- IA-64は複数のプログラムが動いている状態でIA-32と比べて性能の低下が少ない
 - 複数のプログラムを動かす状況で有用
- IA-64は2基のプロセッサを搭載
 - 1スレッドと2スレッドの実行時間が等しい
 - 更なるプロセッサも搭載可能
 - IA-64の拡張可能性



考察（個々のプロシジャ、ブートストラップ）

■ 期待された結果は出ていない

- SASに比べてIA-64は新しいアーキテクチャ
- 様々なプラットフォームに対応しているSASにとって細かな最適化までは行っていない可能性
- 今後改善

■ マルチスレッド性能のさらなる向上

- 個々のプロシジャの高速化
- マルチスレッド性能のさらなる向上



考察（大規模なメモリ）

- IA-64はIA-32と比べて大規模な行列
 - メモリ搭載量 (20倍) と比べてその差は小さい
 - SASの仕様の可能性
- 最大サイズの行列の作成個数
 - IA-32では1個だが、IA-64は5個
 - IA-64の大規模なメモリが生かされている
- メモリを直接仕様するプロシジャ等には有用



考察 (IA-64の有用性)

- プログラムの書き方、プロシジャに依存
 - マルチスレッドの結果ではIA-64が倍以上速い
 - ブートストラップではIA-64が遅い
- マルチスレッドの状況では有効
 - 個々のPCで行っていたシミュレーションを一台のIA-64機で行う
 - コストはトータルではあまり変わらない
 - 管理の手間の削減
 - 信頼性の向上



結論

医学データ解析においてIA-64は

- 複数のプログラムを動かす状況で有効
- IMLにて大規模な行列を扱える
- 個々のプロシジャは期待された速度は出ない
- 今後最適化が進めばさらに高速化する可能性