

MIRACLE

【c-3】

Oracle Linux  
Summit2003 in Spring

Unbreakable Linux

～Linuxカーネルの拡張機能～

ミラクル・リナックス株式会社

製品本部技術部 チーフアーキテクト

伊東達雄

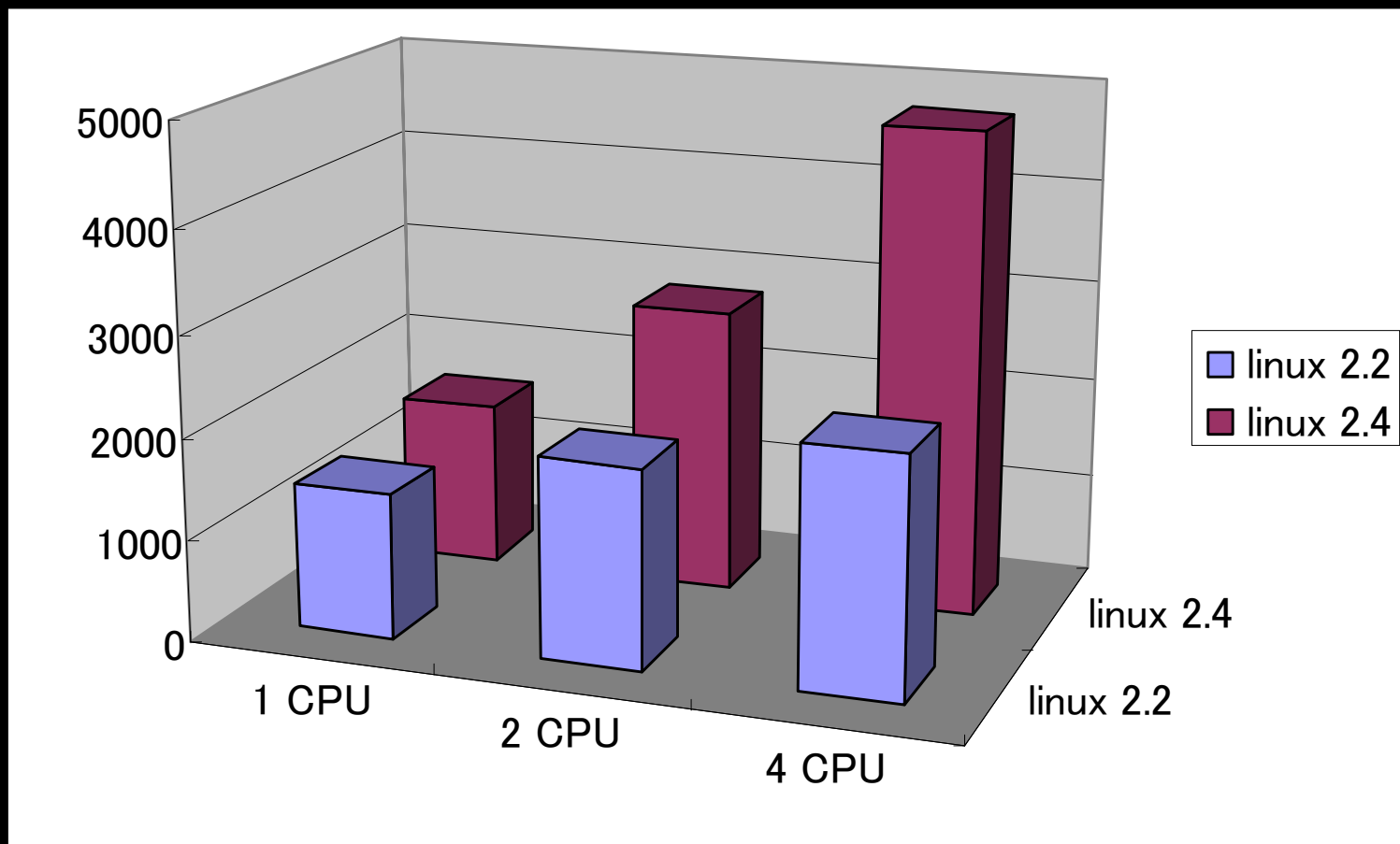
# 目次

- Linux の性能
- Linux カーネル概要
- 強化されたカーネル機能
  - OS のスケーラビリティを向上させる機能
  - 大規模 Oracle DB 向け機能
  - 保守管理性を向上させる機能

# Linuxの性能

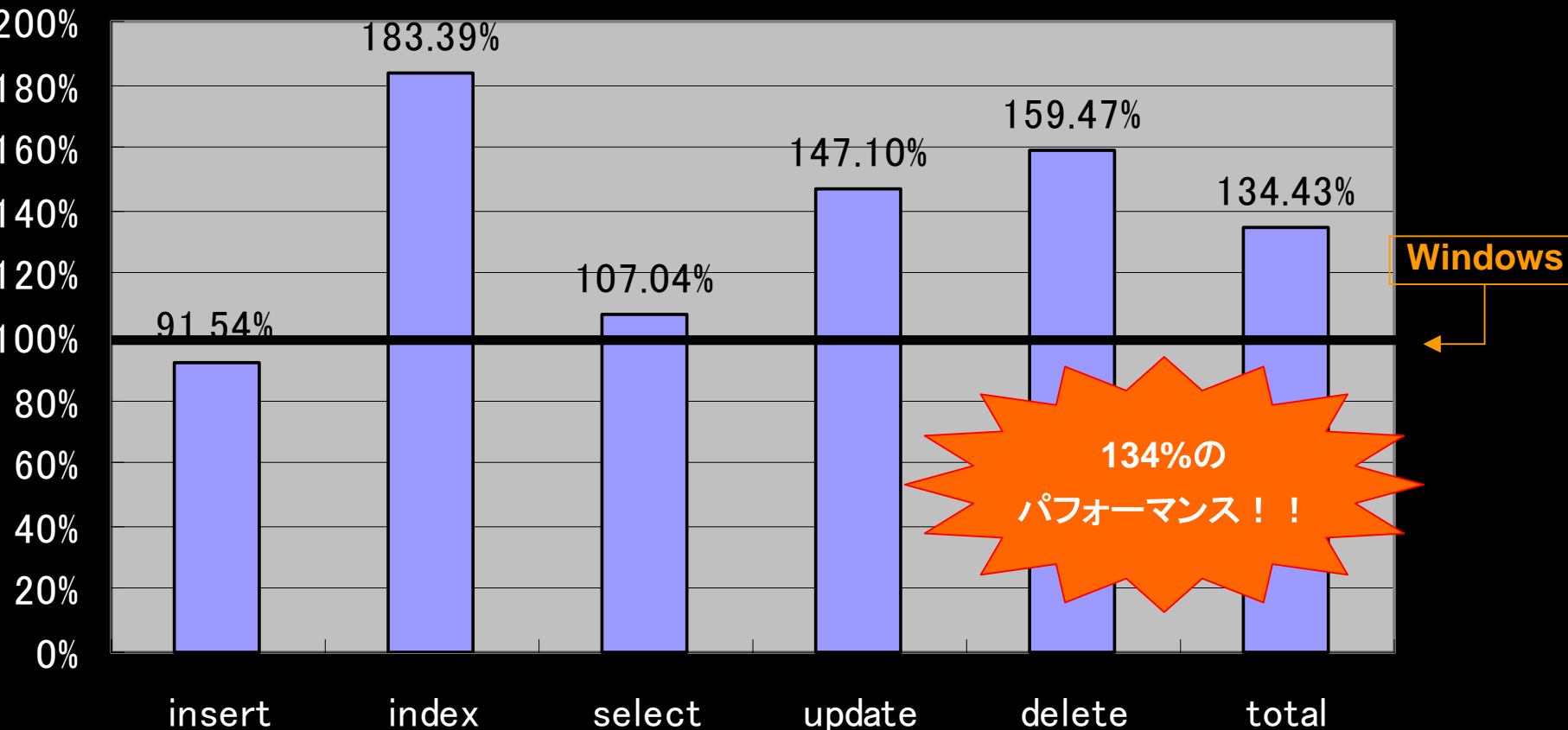
# linux 2.2 と 2.4 のスケーラビリティ

- 2.4でスケーラビリティが強化された
- WebBench で測定



# Oracle9/DB性能比較(Linux vs Windows)

Windowsの処理速度に対する  
MIRACLE LINUXの処理速度の割合



MIRACLE

強化されたカーネル機能

# Oracle9i DB性能比較(Linux vs. Solaris)

Dell PowerEdge6300

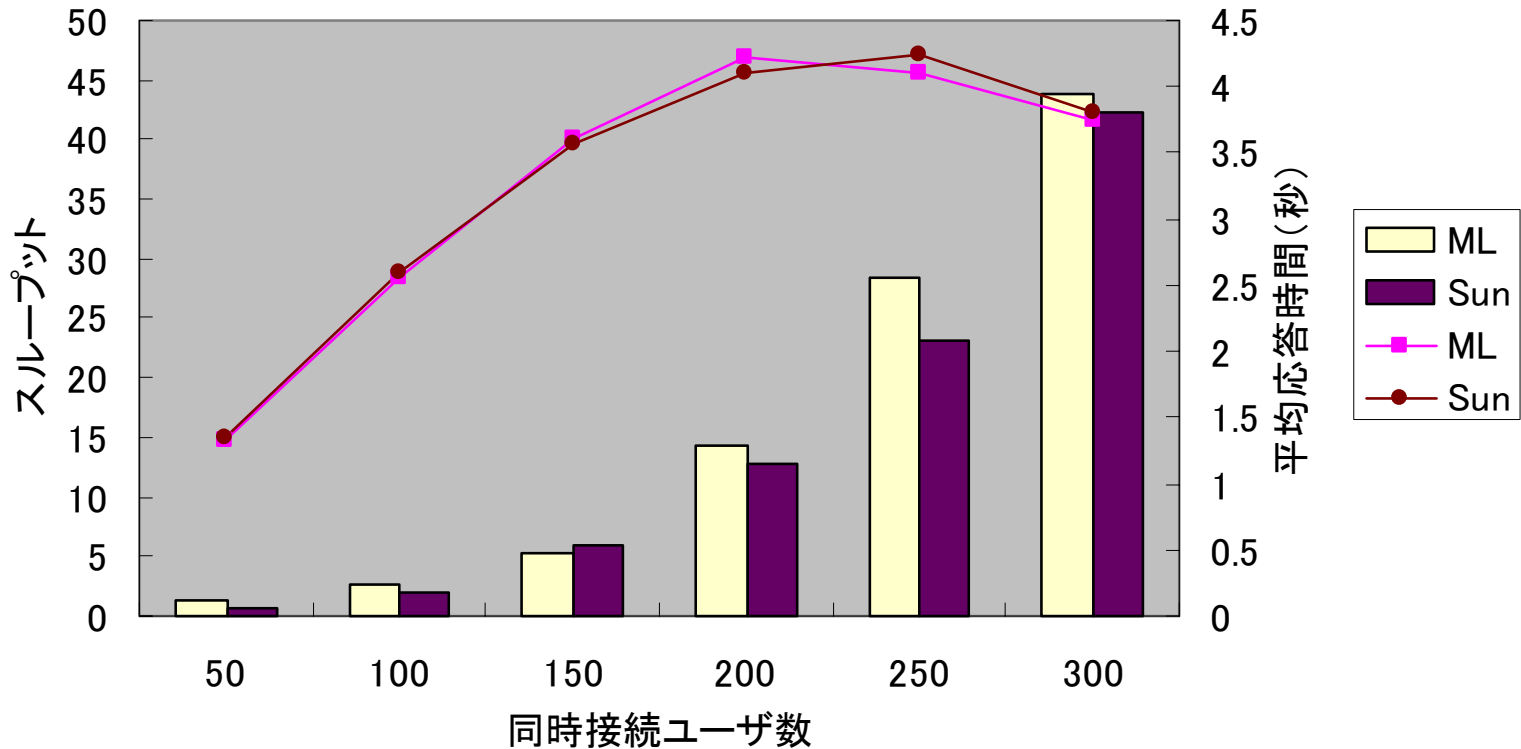
CPU:Pentium II Xeon450MHz × 2



SUN Enterprise 420/R

CPU:UltraSPARC- II × 2

TPC-C

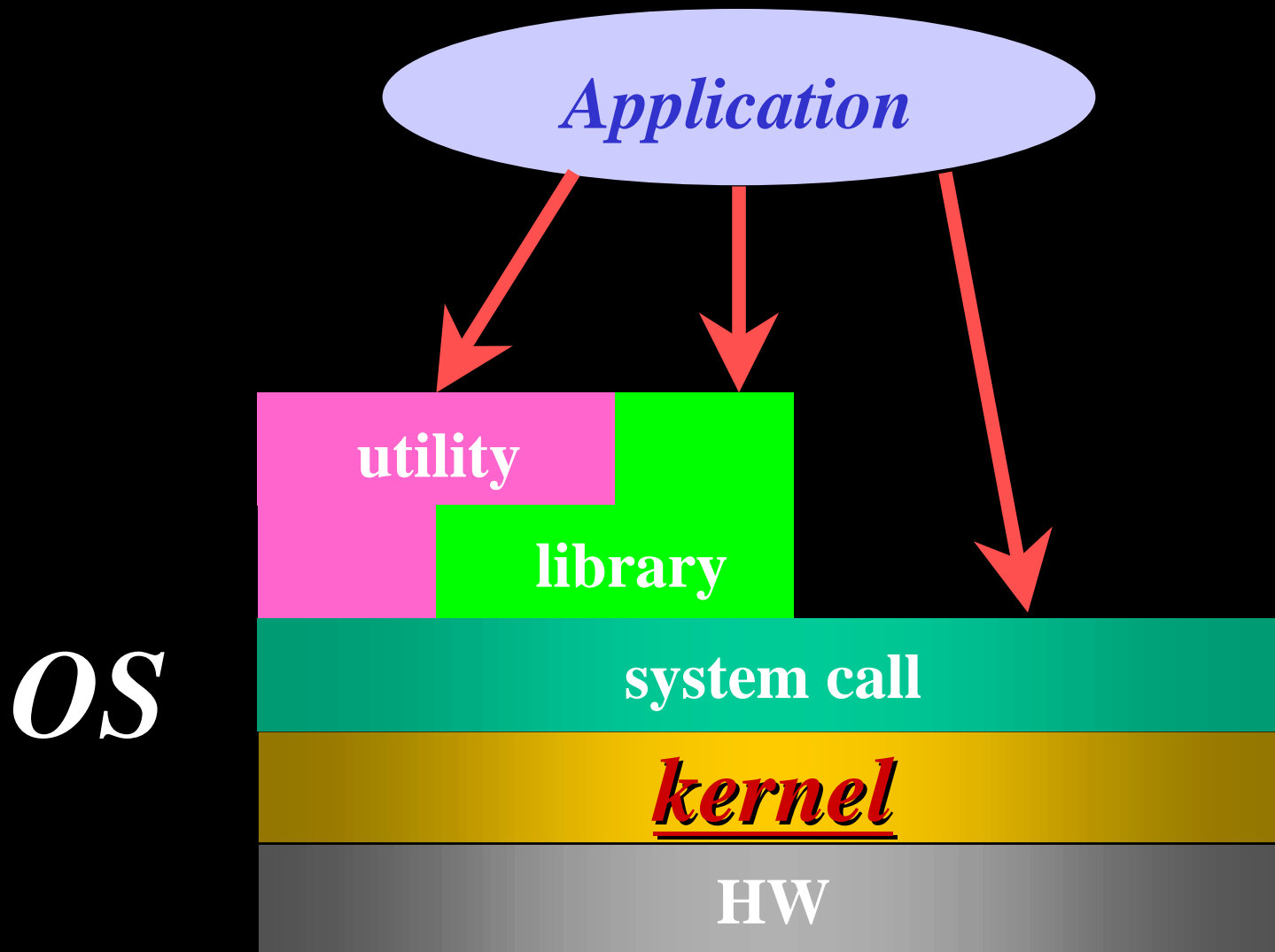


MIRACLE

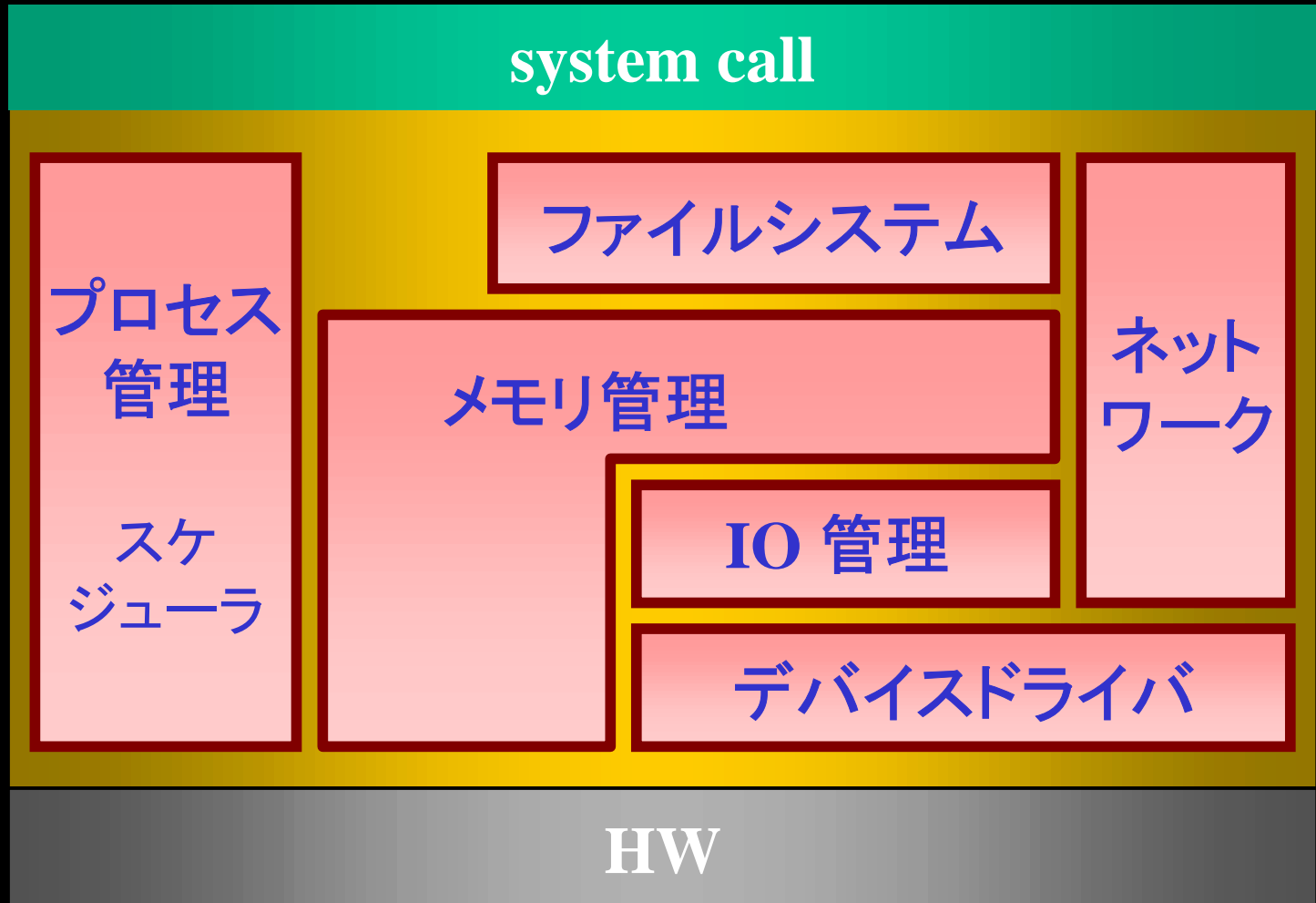
# Linux カーネル 概要



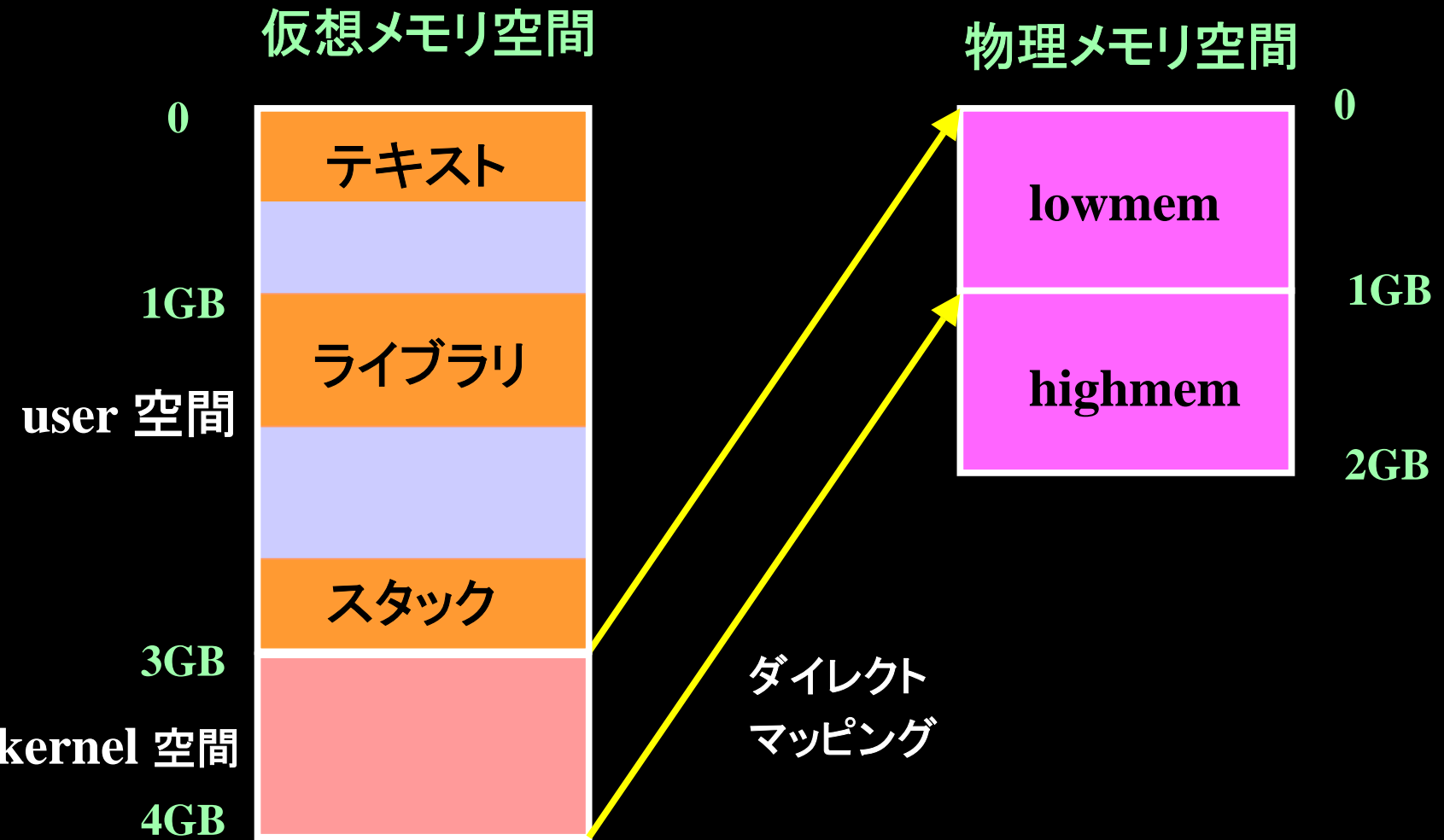
# What is kernel ? (1)



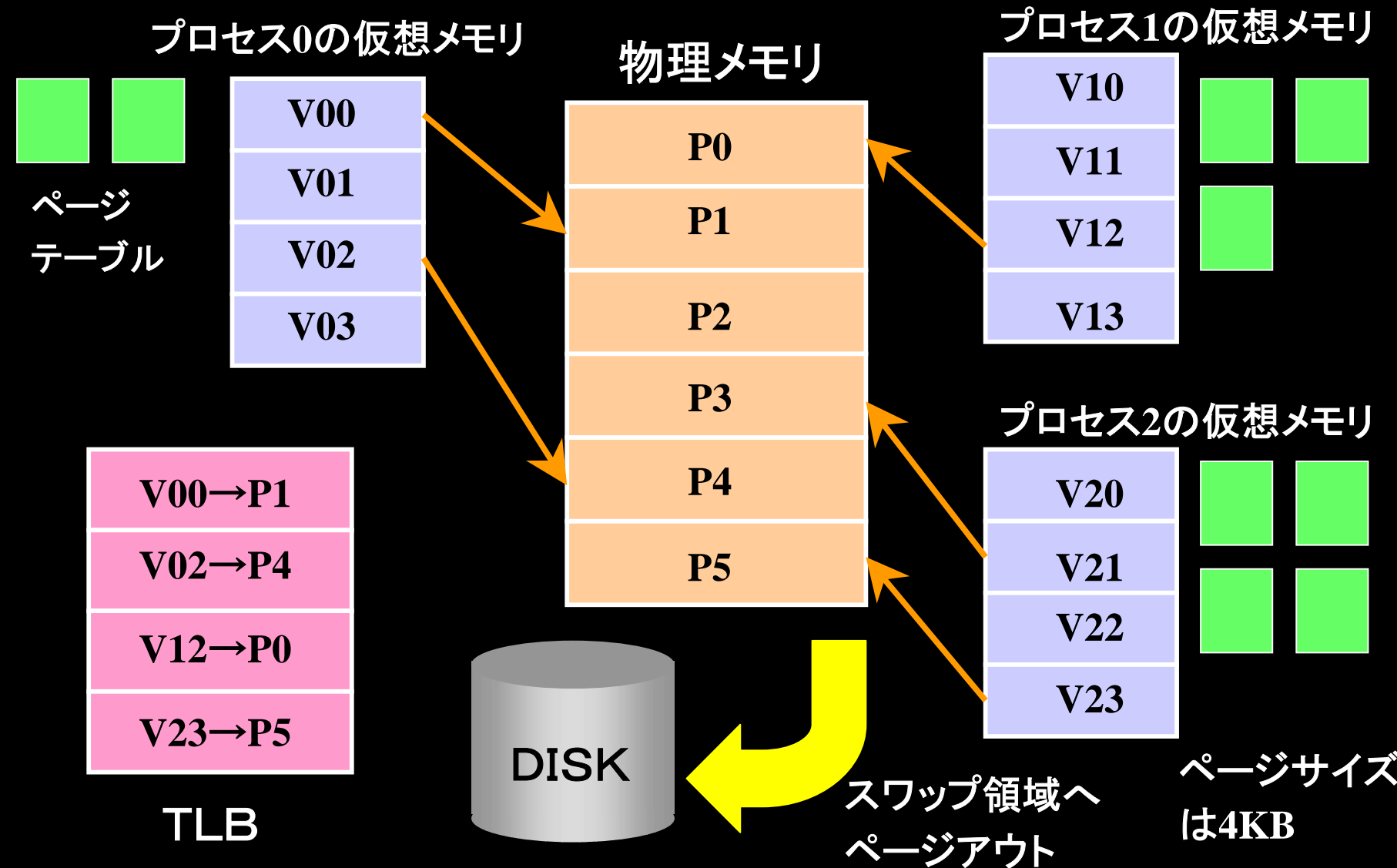
# What is kernel ? (2)



# Linux のメモリ空間レイアウト (IA32)



# 仮想アドレスから物理アドレスへのマッピング (IA32)



MIRACLE

# 強化されたカーネル機能

# 機能強化されたカーネルの特長

- エンタープライズ領域で要求されるスケーラビリティを実現
- Oracle 9iR2 と組み合わせることによって、より大規模な DBシステムを構築可能
- LKCD/LKSTの導入により障害解析が容易となり、保守管理性が向上

# OS のスケーラビリティを 向上させる機能

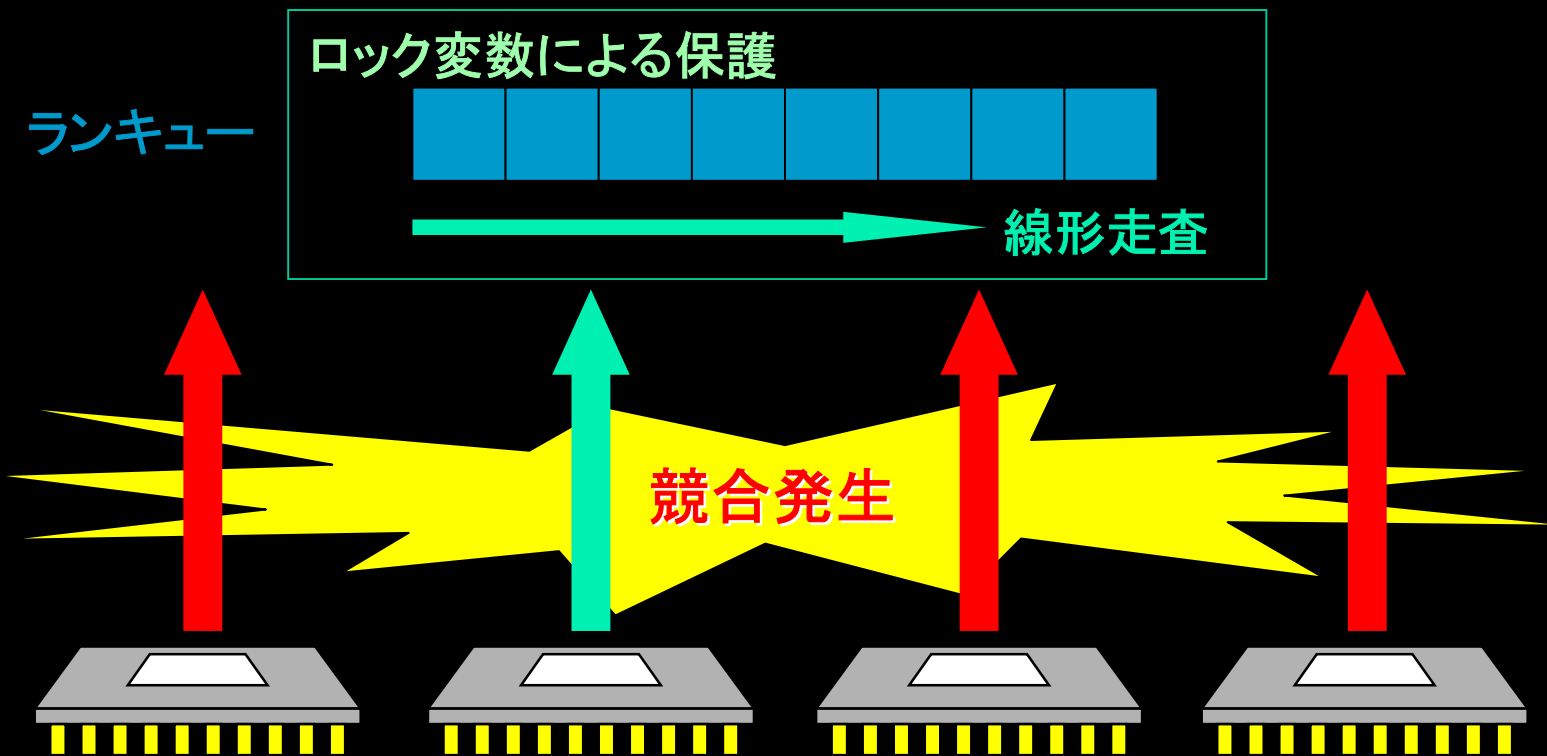
# 強化された OS 機能一覧

- プロセススケジューラの改善
  - O(1) スケジューラの導入
- 非同期I/O のサポート
- I/O リクエストロックの細分化
- バウンズバッファ処理の改善
- ページテーブルの highmem 領域利用



# O(1)スケジューラ

- 従来スケジューラ

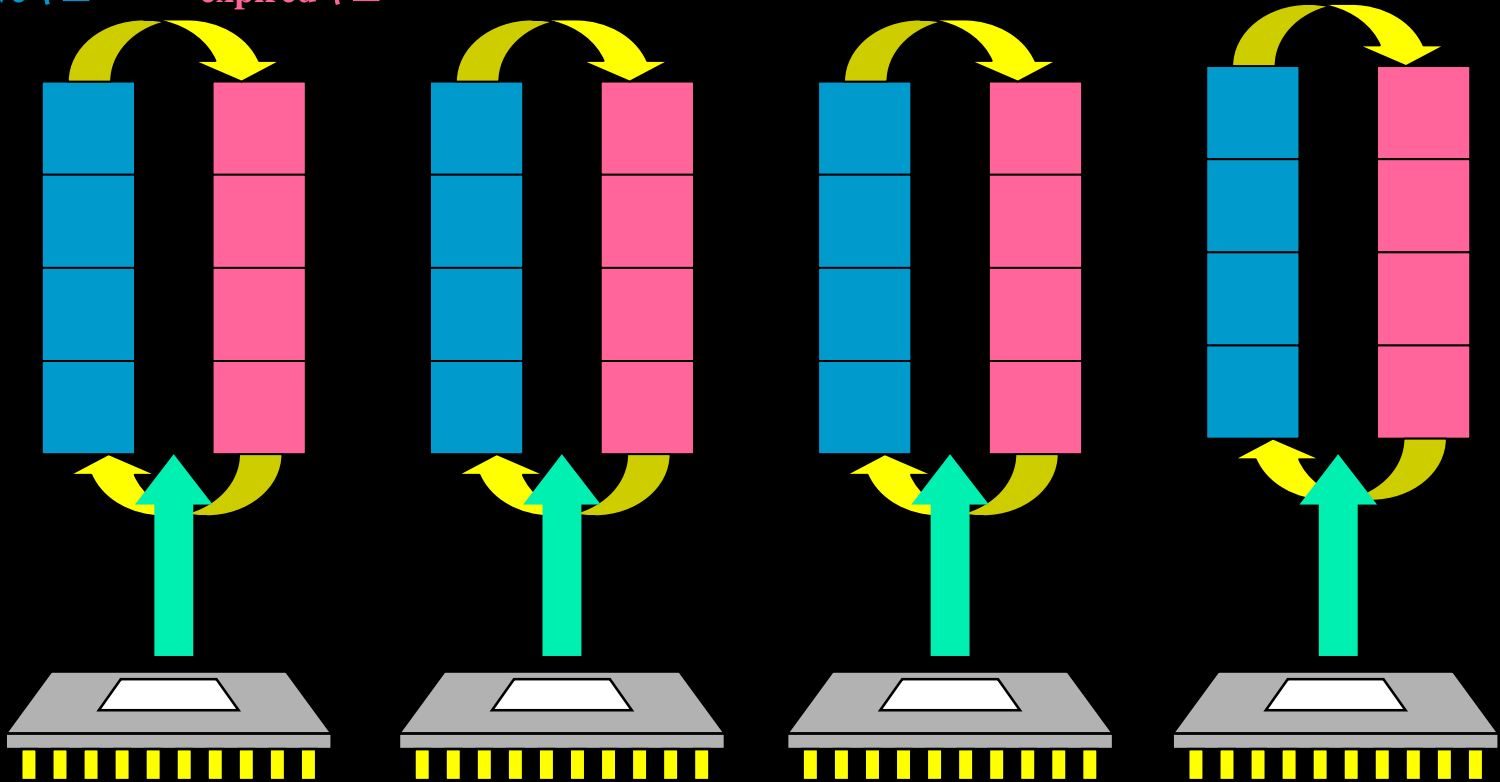


# O(1)スケジューラ

- O(1) スケジューラ

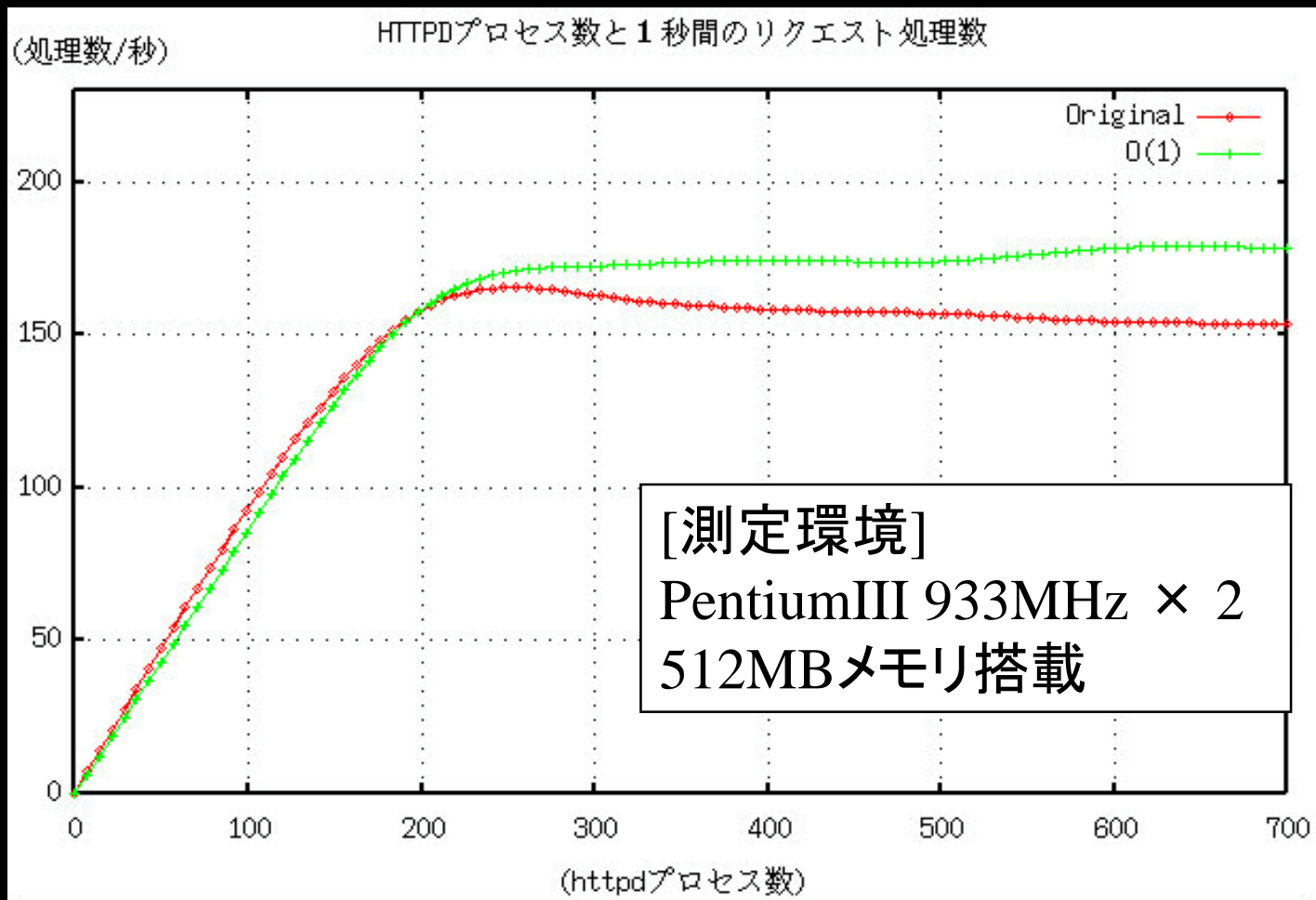
activeキュー

expiredキュー

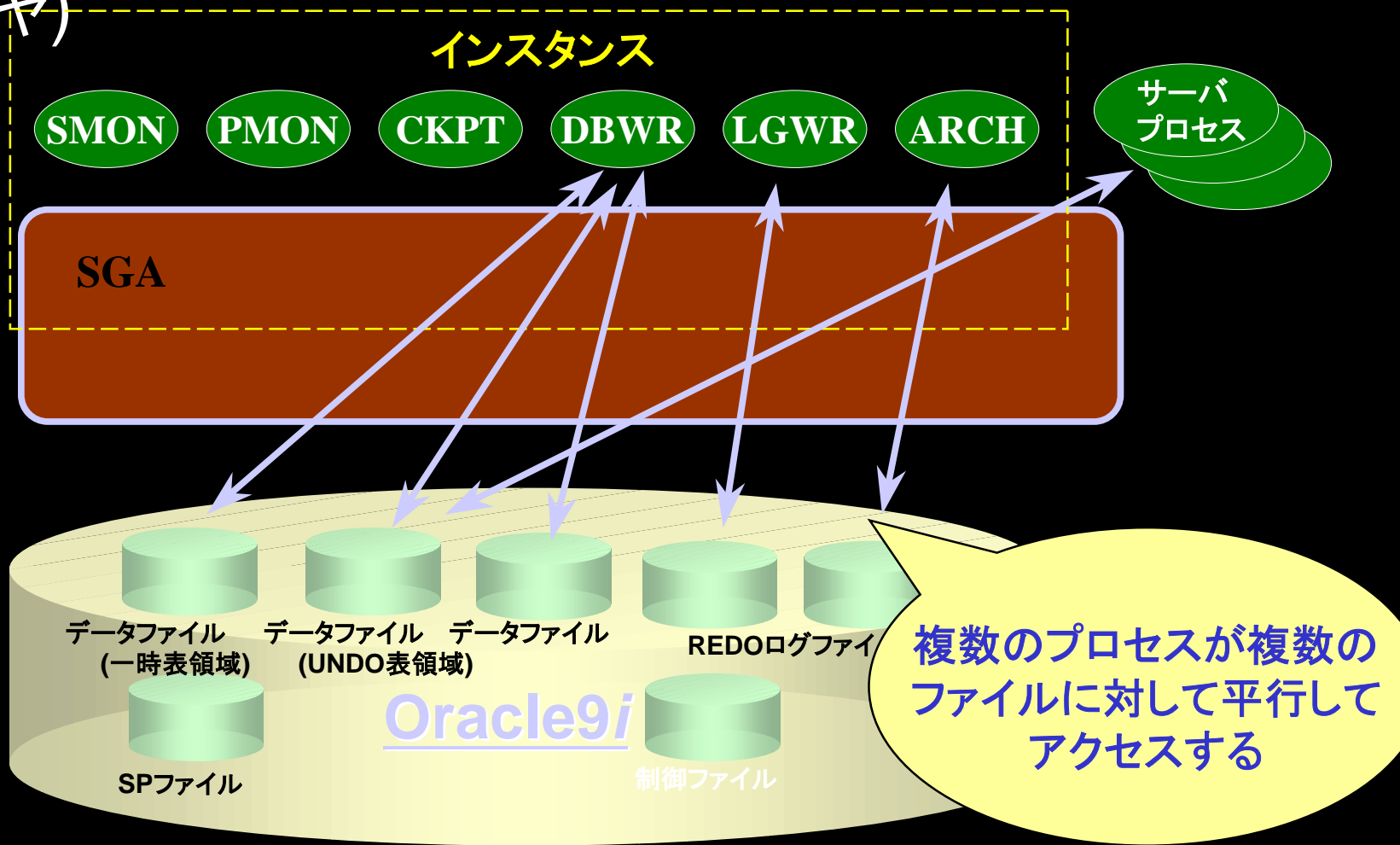


# O(1)スケジューラ測定結果

- 2CPUで測定。CPU数が増加すればより顕著に

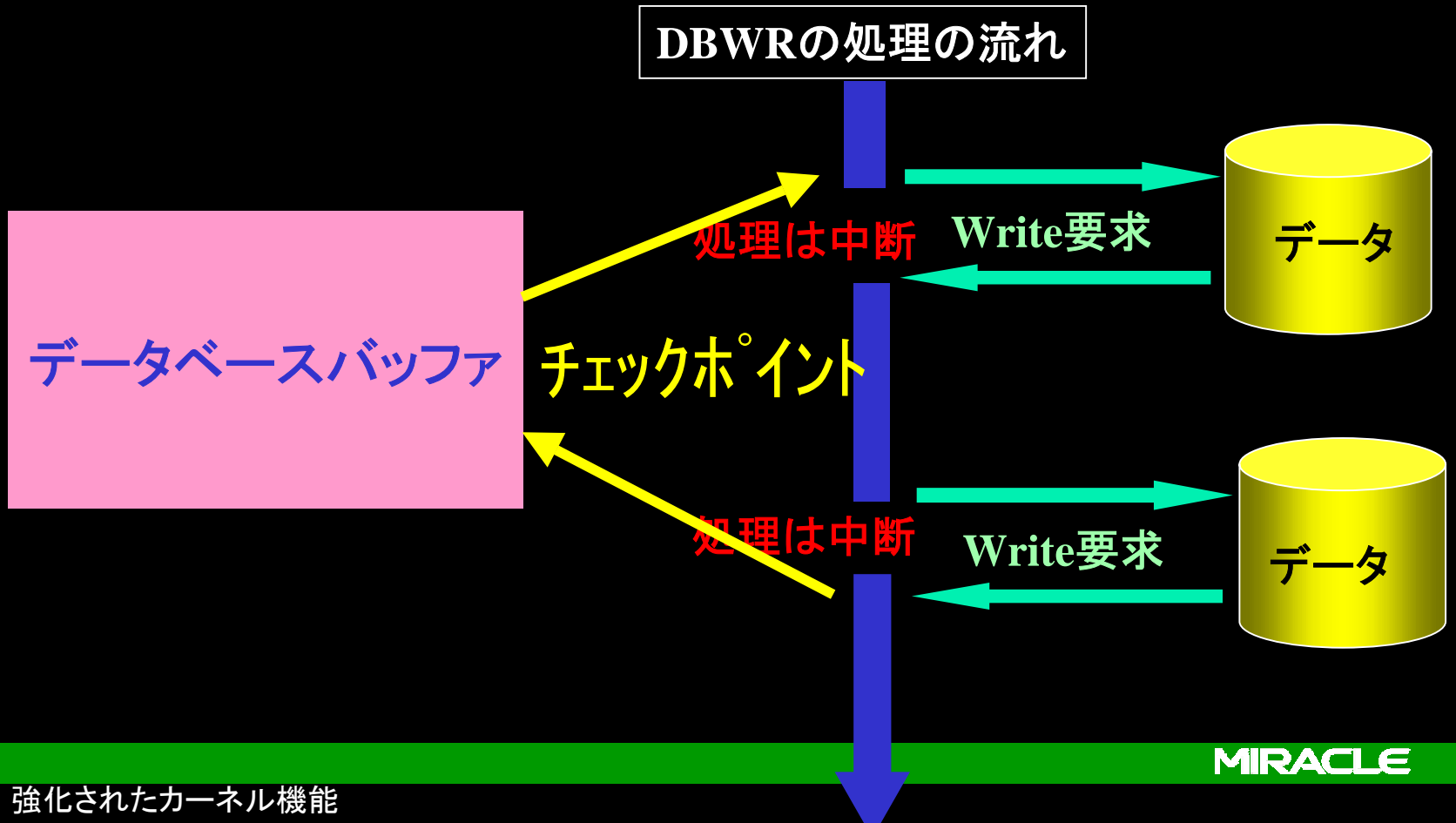


# 非同期 I/O (Oracleデータベースのアーキテクチャ)



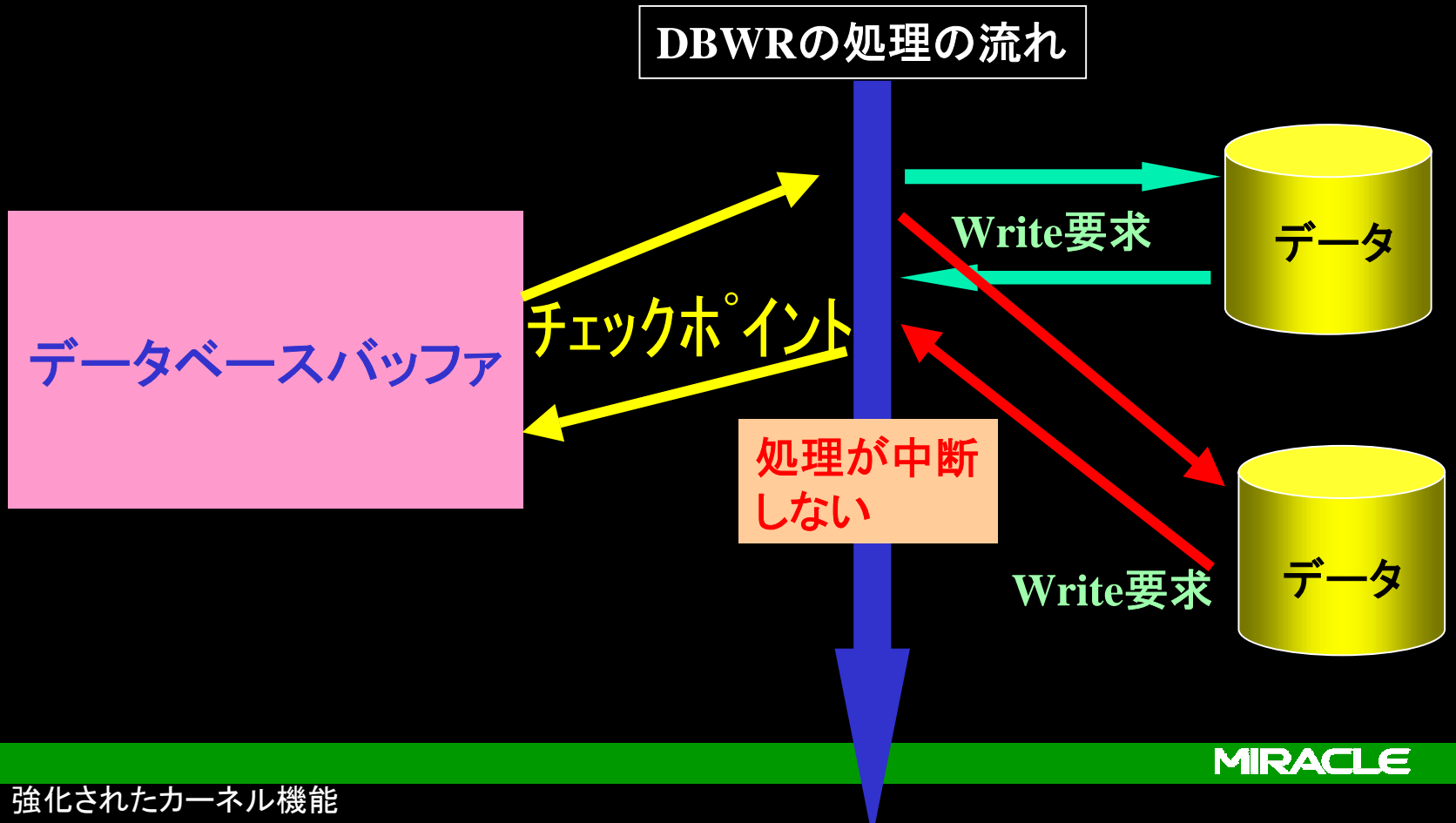
# 非同期 I/O (Oracle DBWRのスループット向上)

- 従来のデータベースI/O処理
  - 高負荷なシステムではチェックポイントに時間がかかる
  - チェックポイントのWriteはO\_SYNC



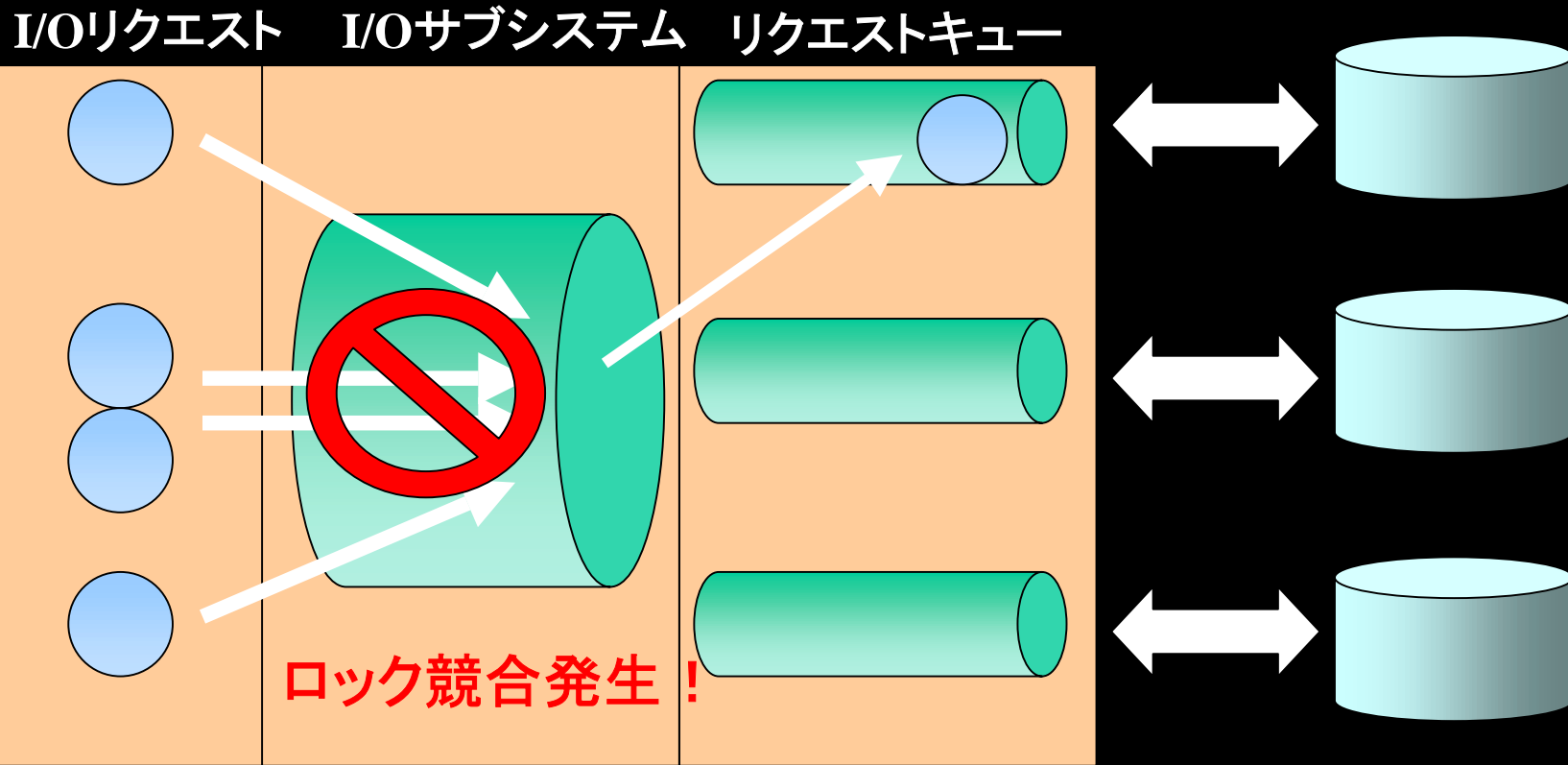
# 非同期 I/O (Oracle DBWRのスループット向上)

- Oracle 9i R2より非同期I/Oをサポート
  - 高負荷なシステムでのチェックポイント完了時間短縮



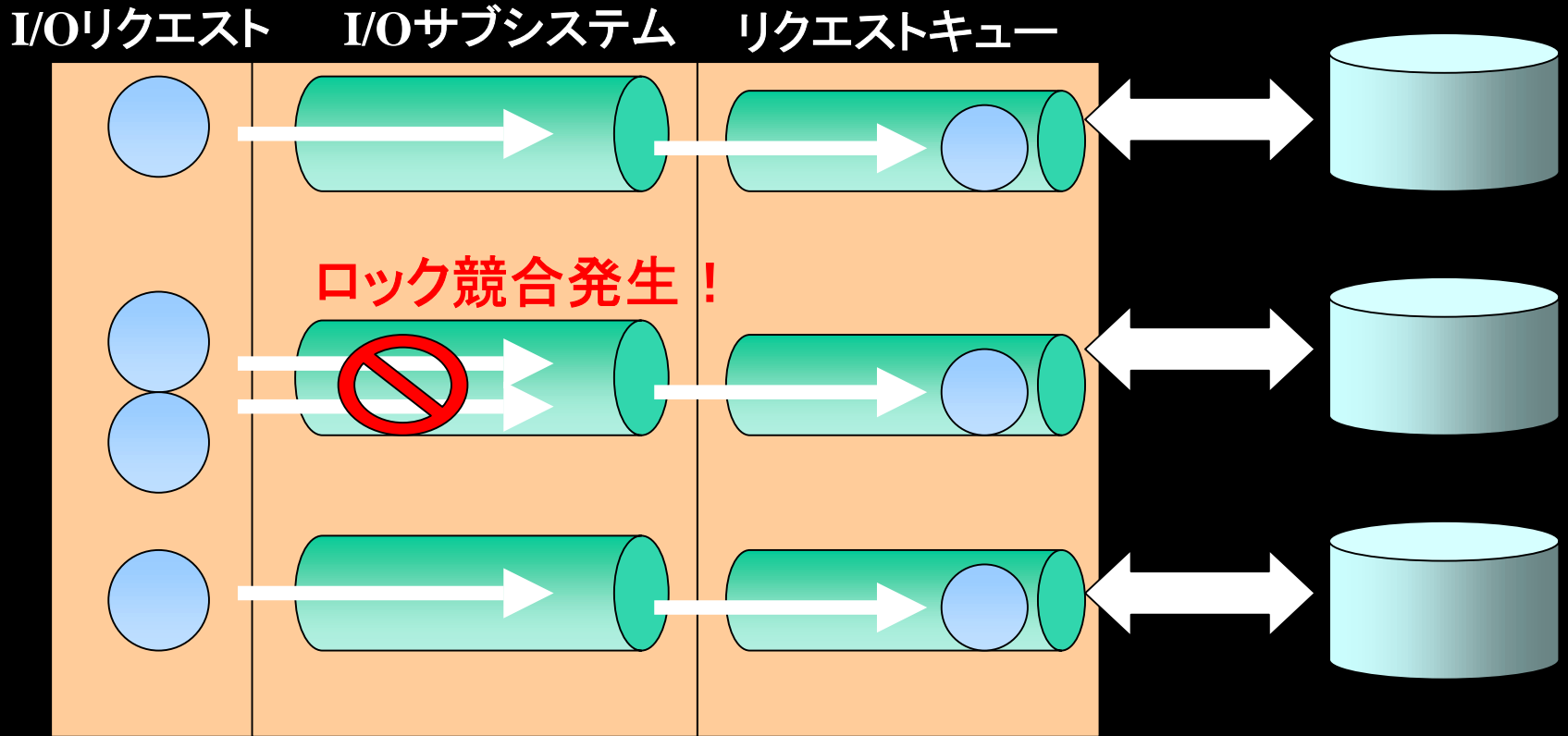
# I/Oリクエストロックの細分化

- I/Oリクエストロックとは？
  - カーネルがI/Oリクエストを制御するためのロック



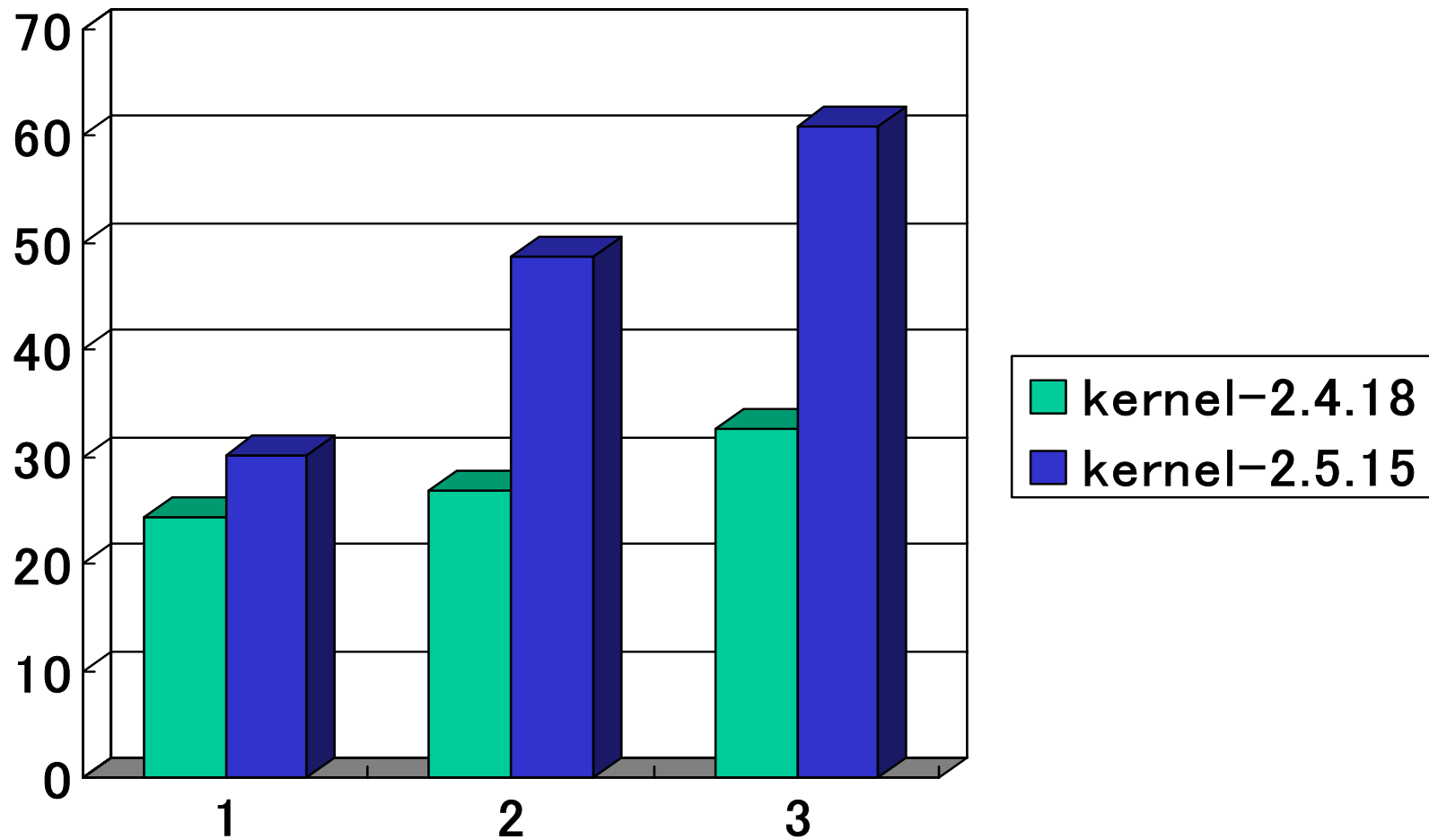
# I/Oリクエストロックの細分化

- ロック競合頻度の減少によってI/O性能が向上



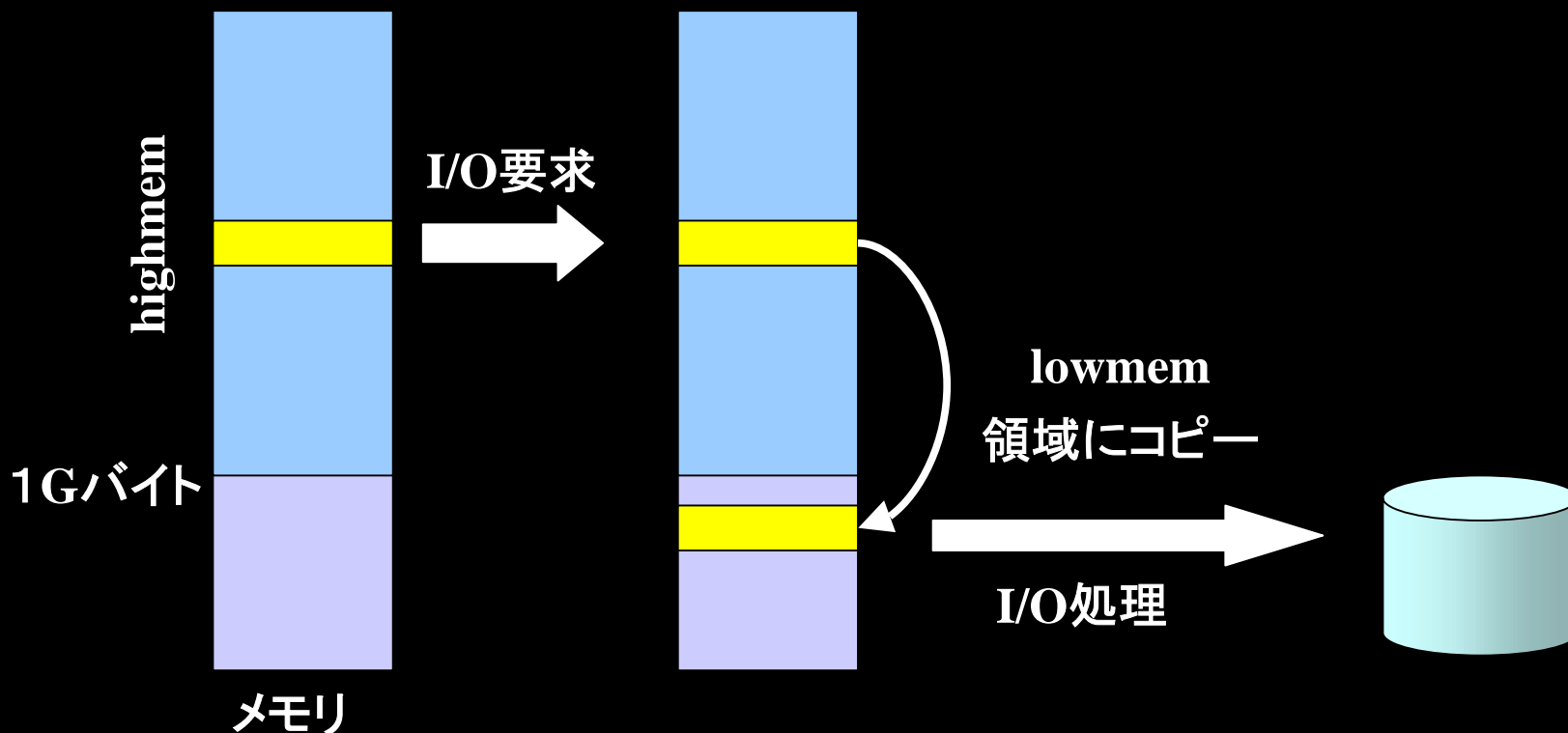


# 多重I/O性能の測定結果



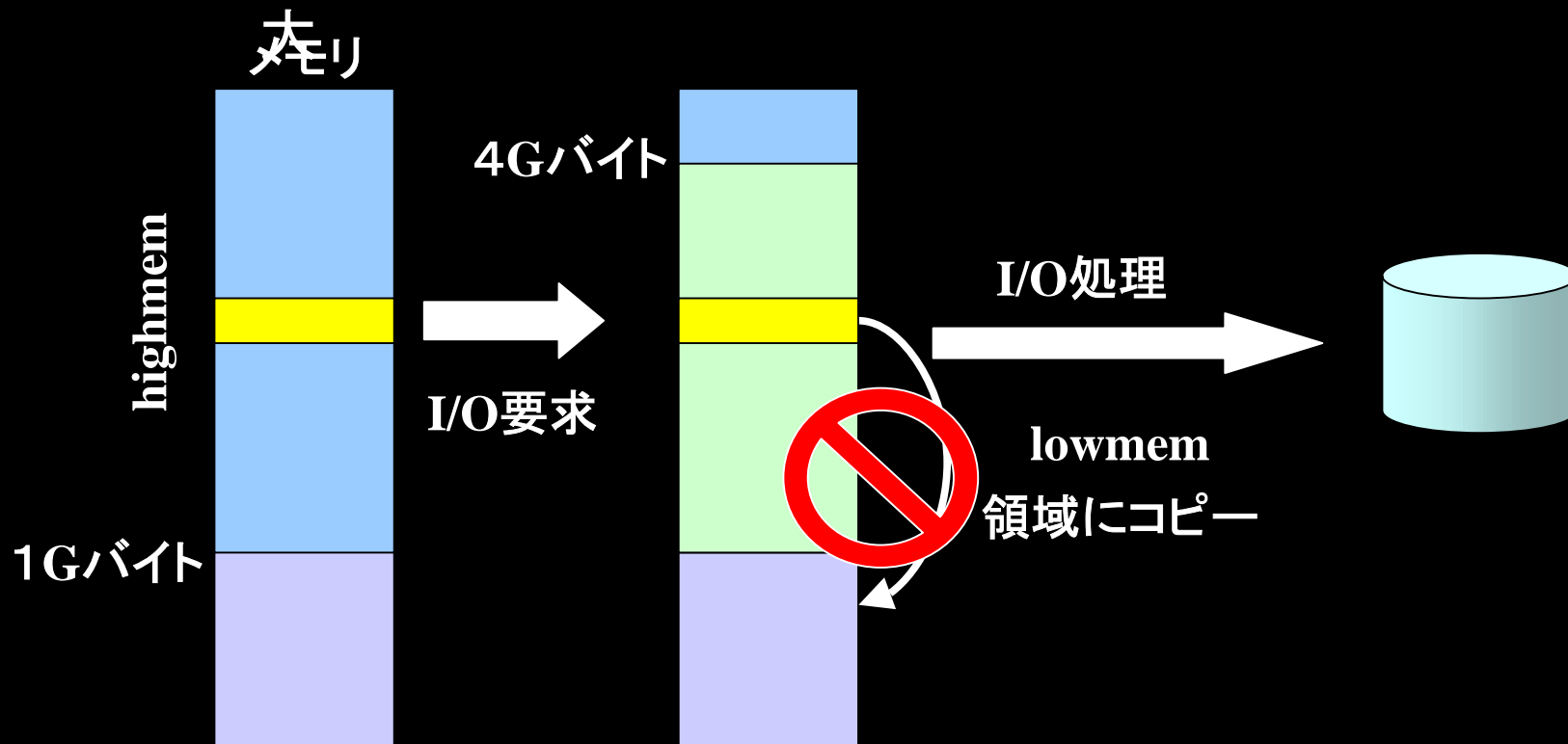
# バウンズバッファ処理の改善

- バウンズバッファ処理とは？
  - I/Oデバイスが直接扱えないメモリ領域に置かれたデータを、扱えるメモリ領域に一度コピーしてから、I/O処理を行う。

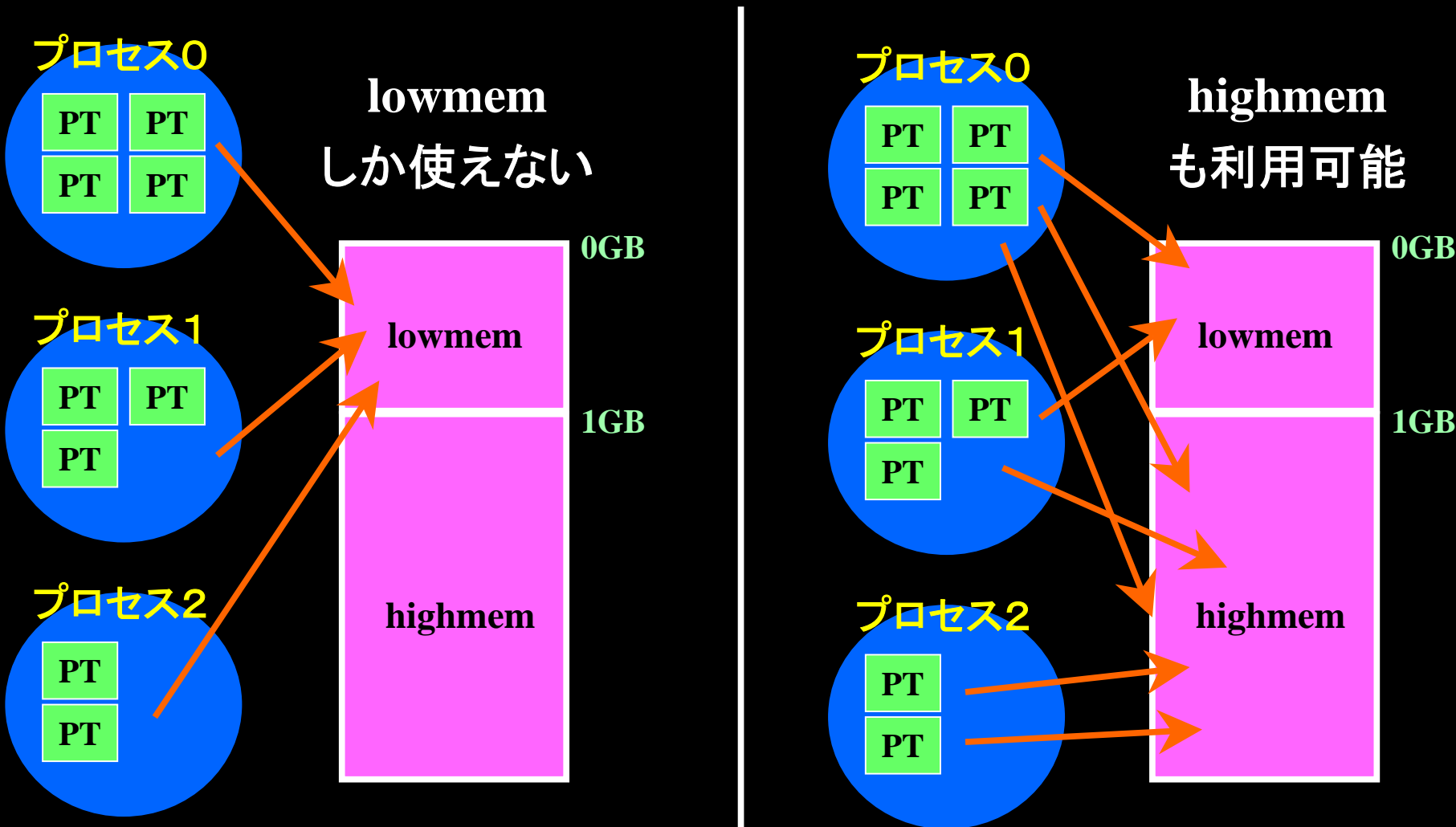


# バウンズバッファ処理の削減

- バッファコピー処理の削減によるI/O性能向上
  - バウンズバッファ処理を必要としないメモリ領域が拡大



# ページテーブルのhighmem領域の利用

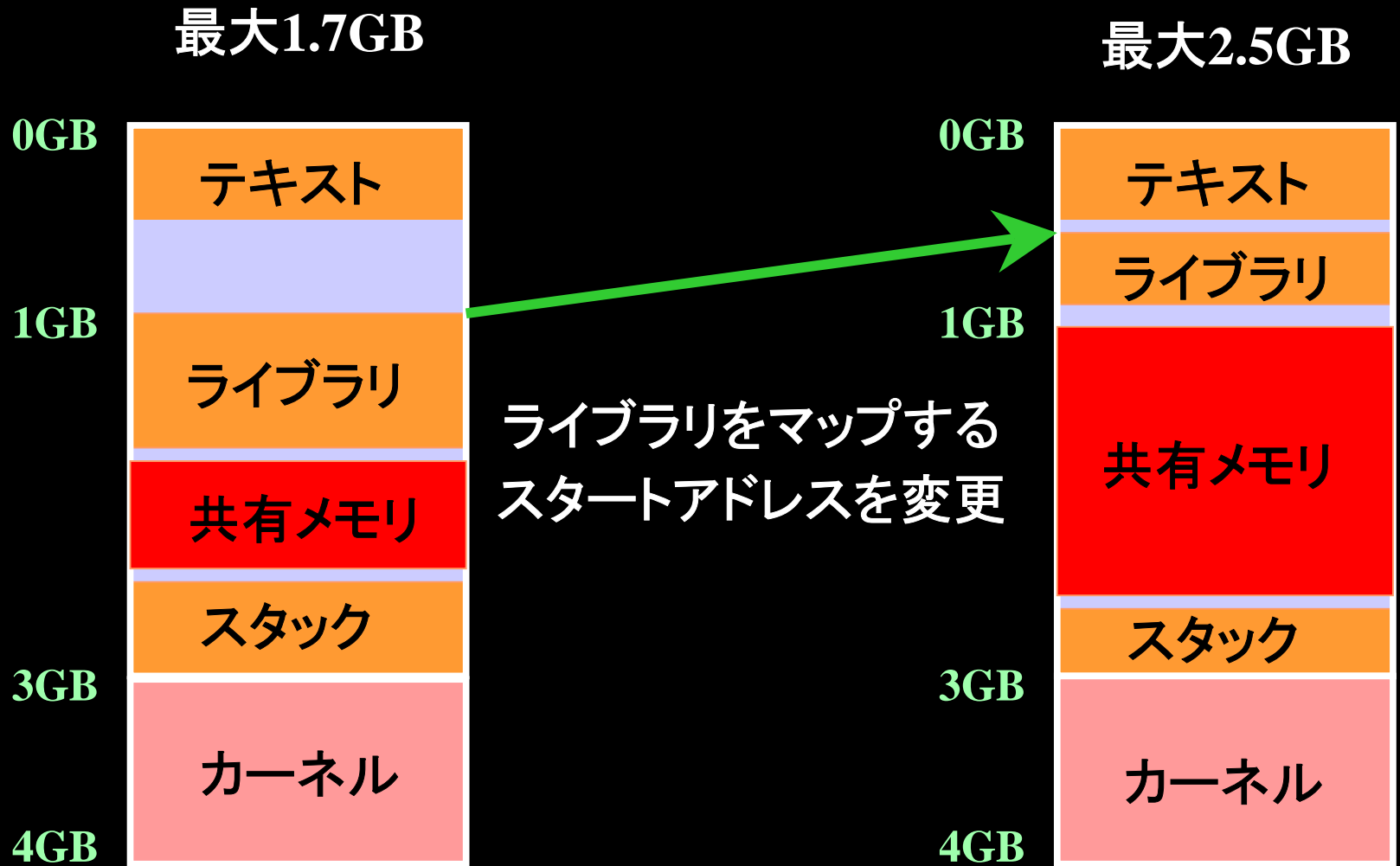


# 大規模 Oracle DB 向け機能

# Oracle DB 向けに強化された機能一覧

- 共有メモリとして利用可能な仮想メモリ空間の拡大
- ラージページ共有メモリ
- Oracle VLM 機能

# 共有メモリとして利用可能な仮想メモリ空間の拡大



# ラージページ共有メモリ

共有メモリ8MB

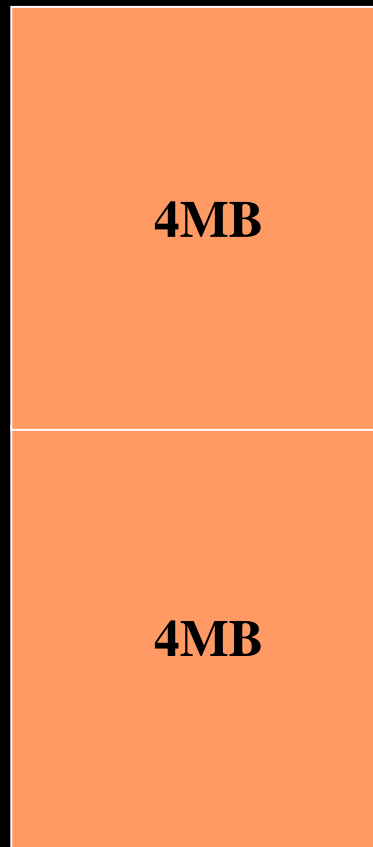
4k	4k	4k	4k
4k	4k	4k	4k
4k	4k	4k	4k
4k	4k	4k	4k

→ 2048ページ  
→ TLBミスが  
発生



ラージページ化

共有メモリ8MB

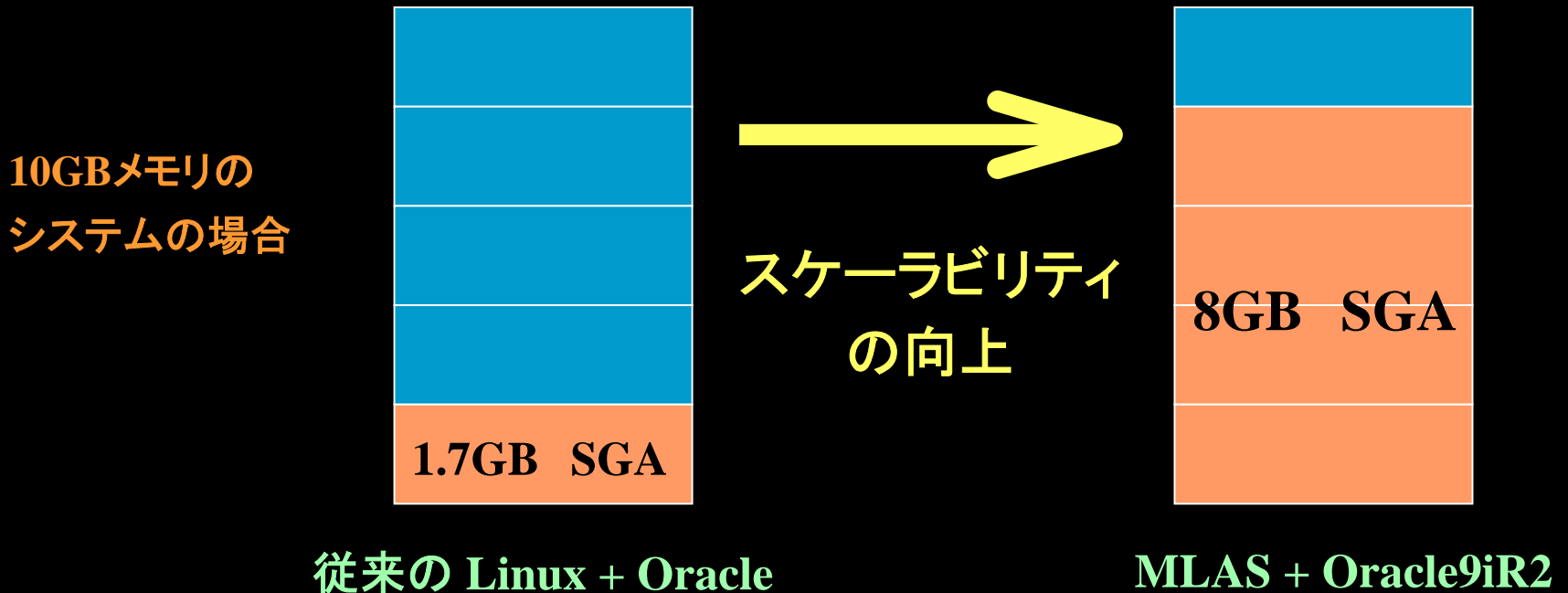


→ 2ページ  
→ TLBミスが  
発生しない



# Oracle VLM 機能

- 大規模メモリを Oracle のデータベースバッファキャッシュとしてフルに利用可能
  - 従来ではSGAのサイズは最大約1.7GBという制限
  - MLAS+Oracle9iR2 ではメモリファイルシステム使って物理メモリをフルに利用可能



# 保守管理性を向上させる機能

# LKCD, LKST

- LKCD (Linux Kernel Crash Dump)
  - システムクラッシュ発生時のメモリダンプを採取、解析するための機能
- LKST (Linux Kernel State Tracer)
  - カーネルの内部情報の変化を記録、確認するための機能

再現性の低い障害も確実に障害情報の採取が可能に

# MIRACLE

【お問い合わせ先】

info@miraclelinux.com

<http://www.miraclelinux.com>

## ミラクル・リナックス株式会社 【無断転載を禁ず】

この文書はあくまでも参考資料であり、掲載されている情報は予告なしに変更されることがあります。ミラクル・リナックス(株)は本書の内容に関していかなる保証もいたしません。また、本書の内容に関連したいかなる損害についても責任を負いかねます。又、本資料の著作権は特に指定されている箇所を除いて、ミラクル・リナックスが有します。ミラクル・リナックスが著作権を有するコンテンツにつきましては、ミラクル・リナックスに対して無断で複製、改変、頒布などを行うことはできません。

MIRACLE LINUX の製品名、ロゴ、サービス名などは、ミラクル・リナックスが所有するか、使用権許諾を受けている商標もしくは登録商標です。その他、本 Web サイトに掲載されている他社の製品名、ロゴなどは、それぞれ該当する各社が所有する商標もしくは登録商標です。

MIRACLE