Enterprise Linux Kernel

～ エンタープライズ向けカーネル機能の紹介～

ミラクル・リナックス株式会社
製品本部技術部
伊東達雄

2002年5月29日

MIRACLE

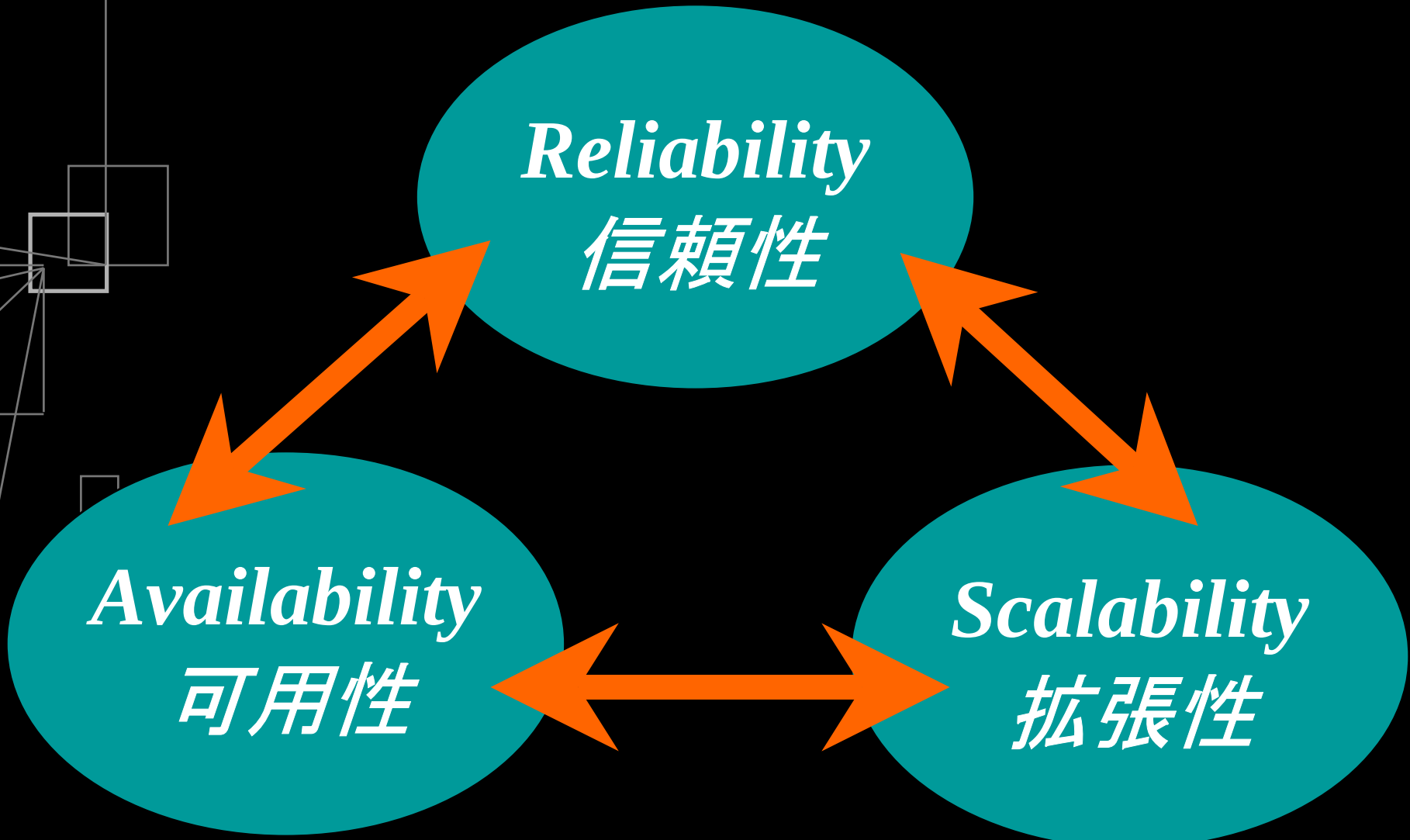
Agenda

- 導入
  - What is Enterprise ？
  - What is kernel ？
- kernel 2.4 から kernel 2.5 までの動向
- Enterprise 領域へ貢献する kernel 2.5 の新機能の紹介
- Enterprise 領域における Miracle Linux の取り組み

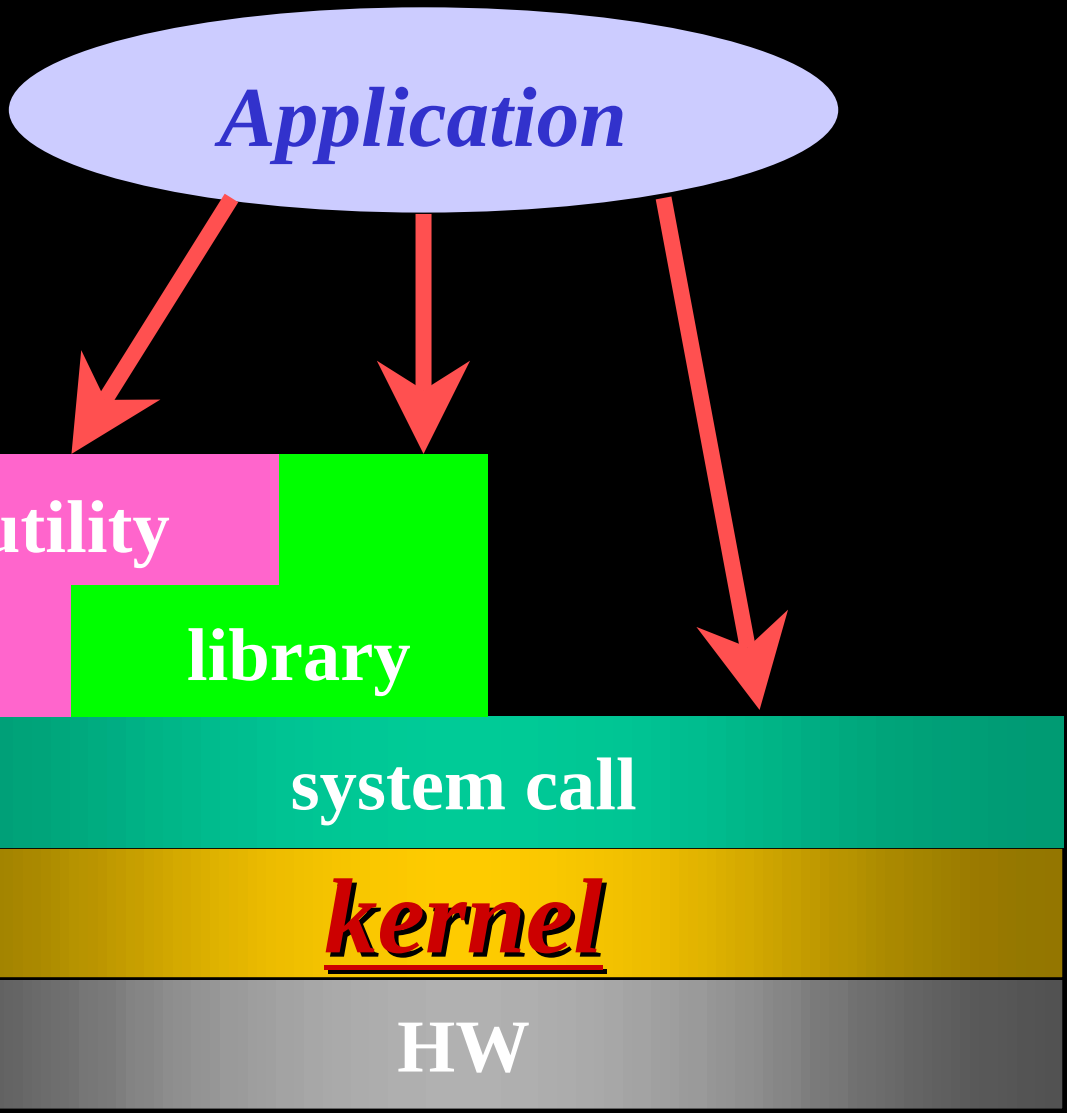
What is Enterprise ? (1)

- クリティカルな e ビジネスにおけるコンピュータシステム
  - 大量なデータを処理
  - 素早いレスポンスタイム
  - ノンストップでサービスを提供

What is Enterprise ? (2)

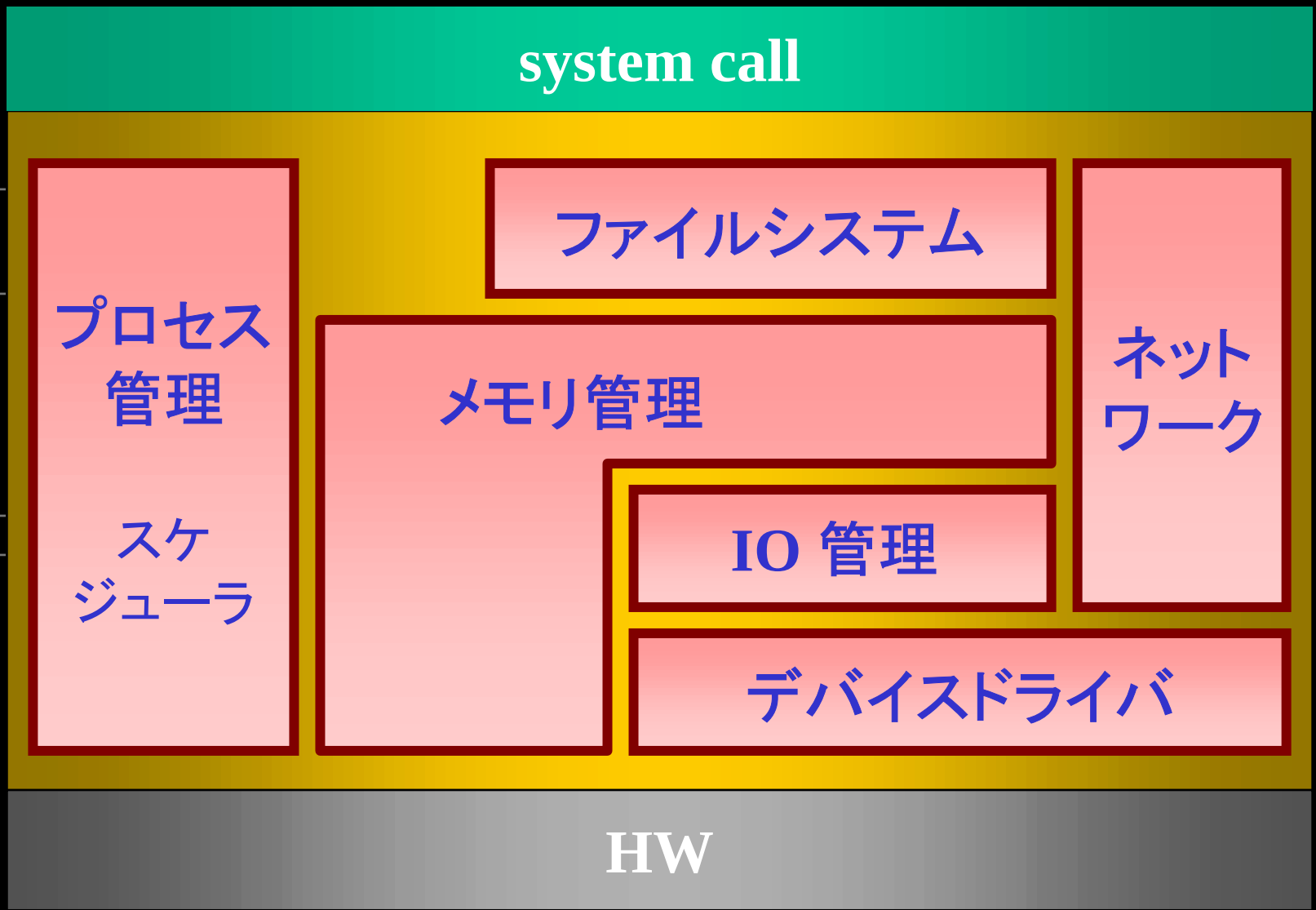


What is kernel ? (1)



OS

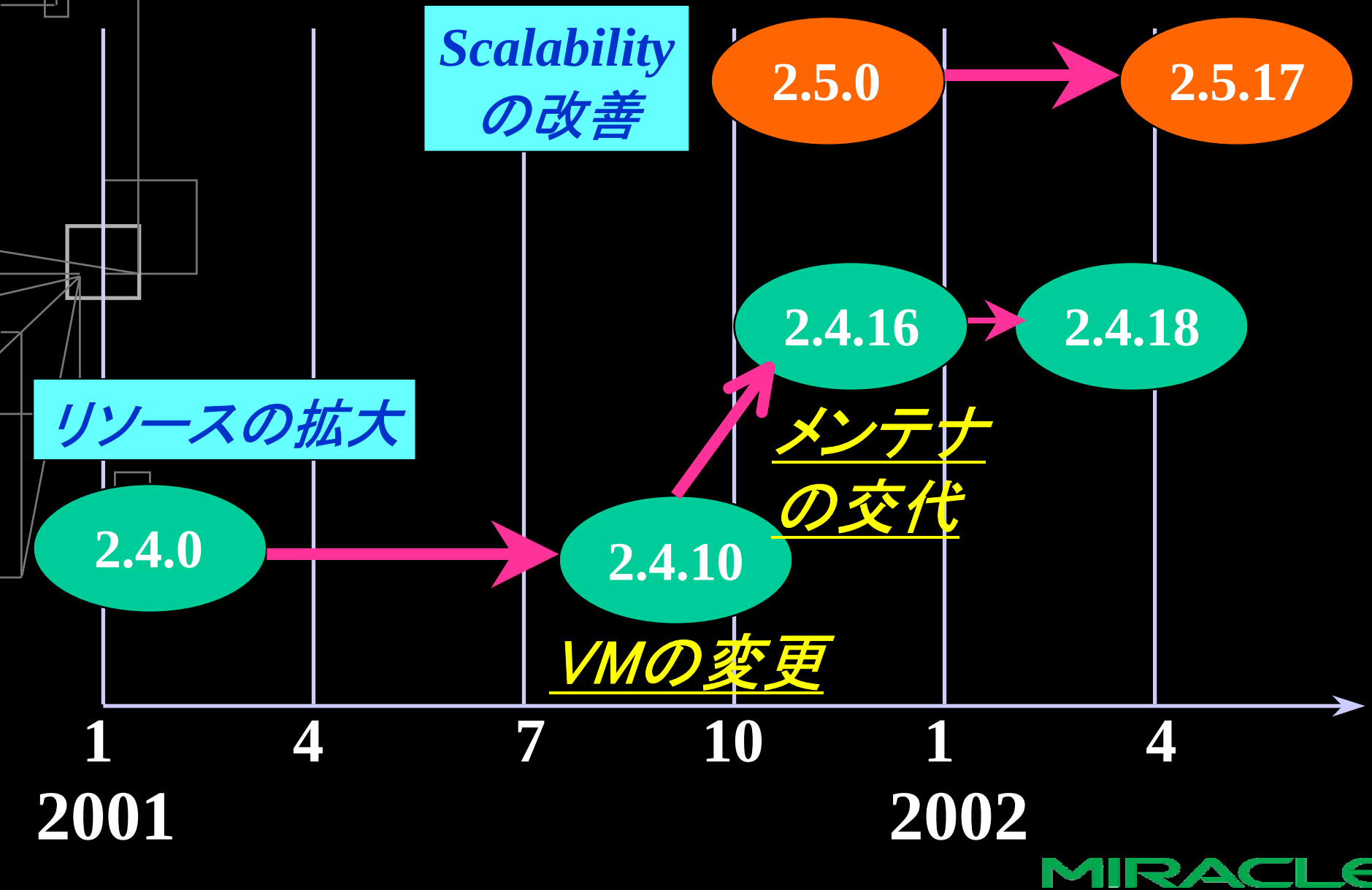
What is kernel ? (2)



kernel 2.4 から kernel 2.5 までの動向 (1)

- 安定版カーネル(偶数)
  - 2.2, 2.4
- 開発版カーネル(奇数)
  - 2.3, 2.5

kernel 2.4 から kernel 2.5 までの動向 (2)



Enterprise 領域へ貢献する kernel 2.4 リリース時の新機能 (1)

- 64GB の物理メモリをサポート
- LFS (最大 4TB のファイルサイズ)
- プロセス数無限
- ユーザ数/グループ数の拡大
- SMP における大幅な性能改善
- ファイルキャッシュの改善

Enterprise 領域へ貢献する kernel 2.4 リリース時の新機能 (2)

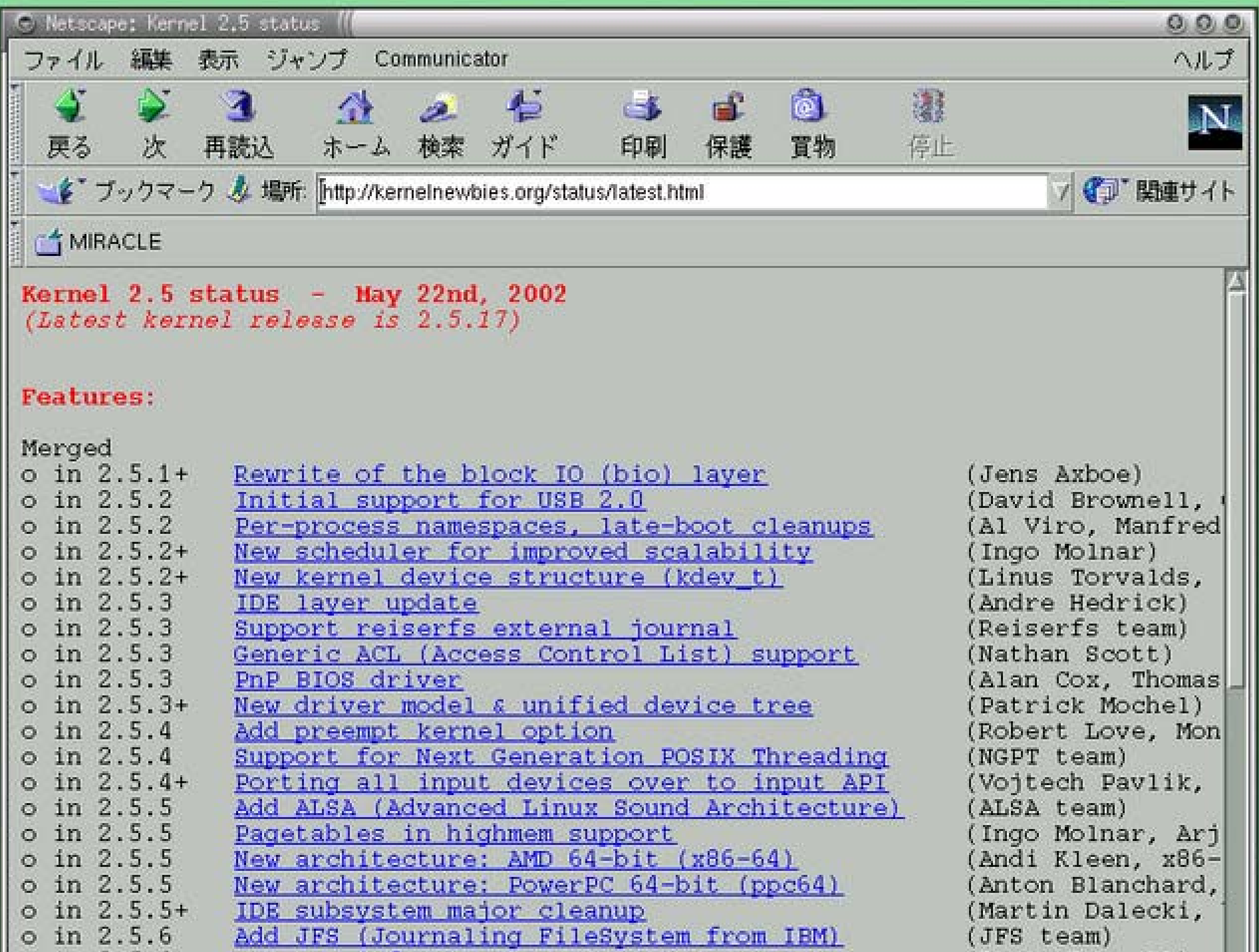
- NFS version 3 対応
- raw デバイス
- ジャーナリングファイルシステム

Enterprise 領域へ貢献する kernel 2.5の新機能 (1)

- block IO (bio) 層 の改善
- O(1) スケジューラ
- New kernel device structure (kdev_t)
- ACL サポート
- preemption の改善
- pagetables in highmem
- AMD 64 bit サポート
- PowerPC 64 bit サポート

Enterprise 領域へ貢献する kernel 2.5の新機能 (2)

- JFS
- NAPI
- system call interface for task affinity
- radix-tree pagecache
- smarter IRQ balancing
- Fast walk dcache
- rewrite buffer layer
- rmap (reverse map) VM



Kernel 2.5 status - May 22nd, 2002
(Latest kernel release is 2.5.17)

Features:

Merged

o in 2.5.1+	Rewrite of the block IO (bio) layer	(Jens Axboe)
o in 2.5.2	Initial support for USB 2.0	(David Brownell,
o in 2.5.2	Per-process namespaces, late-boot cleanups	(Al Viro, Manfred
o in 2.5.2+	New scheduler for improved scalability	(Ingo Molnar)
o in 2.5.2+	New kernel device structure (kdev_t)	(Linus Torvalds,
o in 2.5.3	IDE layer update	(Andre Hedrick)
o in 2.5.3	Support reiserfs external journal	(Reiserfs team)
o in 2.5.3	Generic ACL (Access Control List) support	(Nathan Scott)
o in 2.5.3	PnP BIOS driver	(Alan Cox, Thomas
o in 2.5.3+	New driver model & unified device tree	(Patrick Mochel)
o in 2.5.4	Add preempt kernel option	(Robert Love, Mon
o in 2.5.4	Support for Next Generation POSIX Threading	(NGPT team)
o in 2.5.4+	Porting all input devices over to input API	(Vojtech Pavlik,
o in 2.5.5	Add ALSA (Advanced Linux Sound Architecture)	(ALSA team)
o in 2.5.5	Pagetables in highmem support	(Ingo Molnar, Arj
o in 2.5.5	New architecture: AMD 64-bit (x86-64)	(Andi Kleen, x86-
o in 2.5.5	New architecture: PowerPC 64-bit (ppc64)	(Anton Blanchard,
o in 2.5.5+	IDE subsystem major cleanup	(Martin Dalecki,
o in 2.5.6	Add JFS (Journaling FileSystem from IBM)	(JFS team)

Enterprise 領域へ貢献する kernel 2.5の新機能の紹介

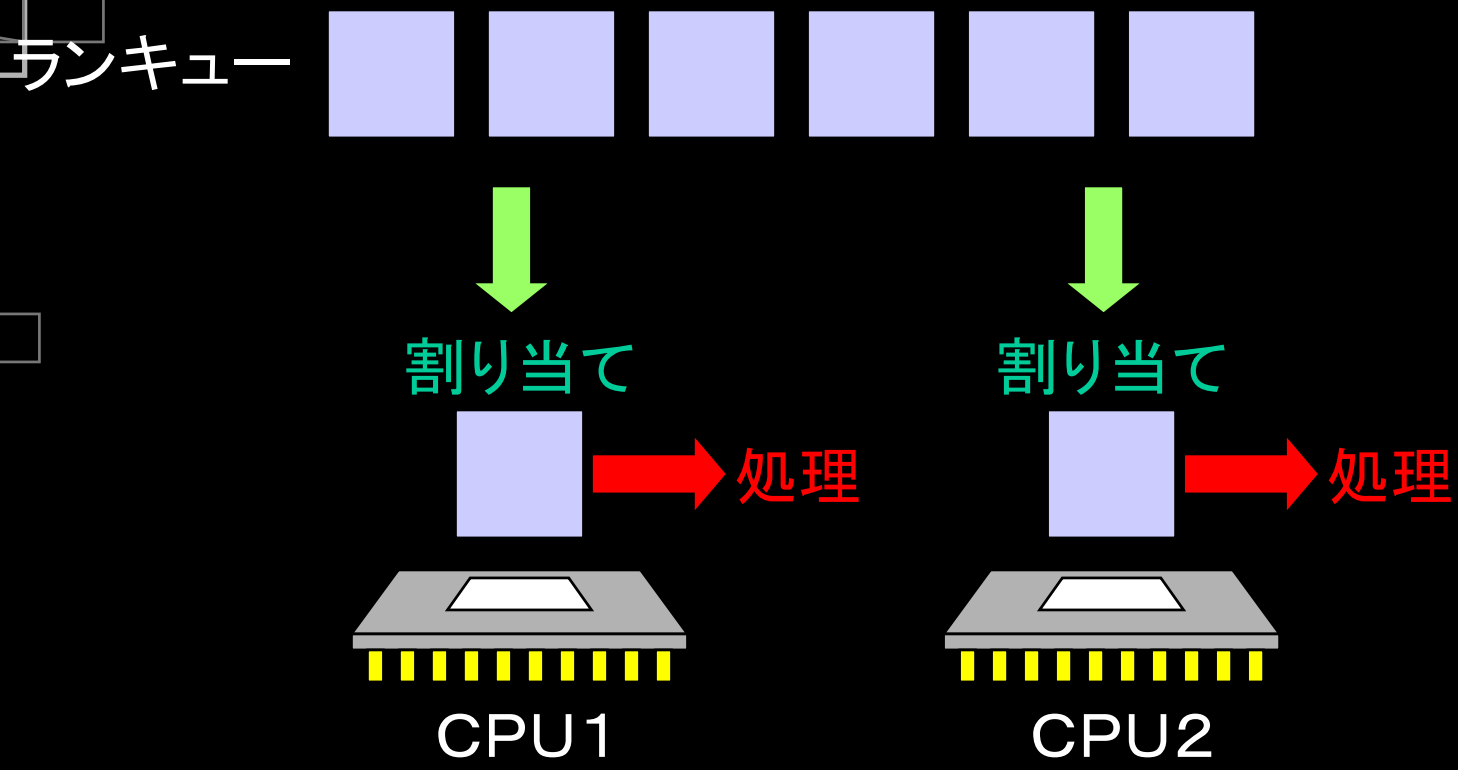
- Scalability を向上させる以下の 4 つの機能をピックアップ
  - プロセス管理
    - O(1) スケジューラ
  - I/O
    - block I/O の改善
  - Network
    - NAPI
  - メモリ管理
    - rmap (reverse map) VM

従来のスケジューラ

- システム全体でランキューを一つだけ保持。
- SMPシステム上ではランキュー走査のために処理をシリアライズ。
- プロセススイッチの際にランキューを線形走査。
- プロセスの多いシステム上ではスケジューラ自体のコストが高くなる。

従来のスケジューラ

- システム全体で1つのランキューを線形走査。

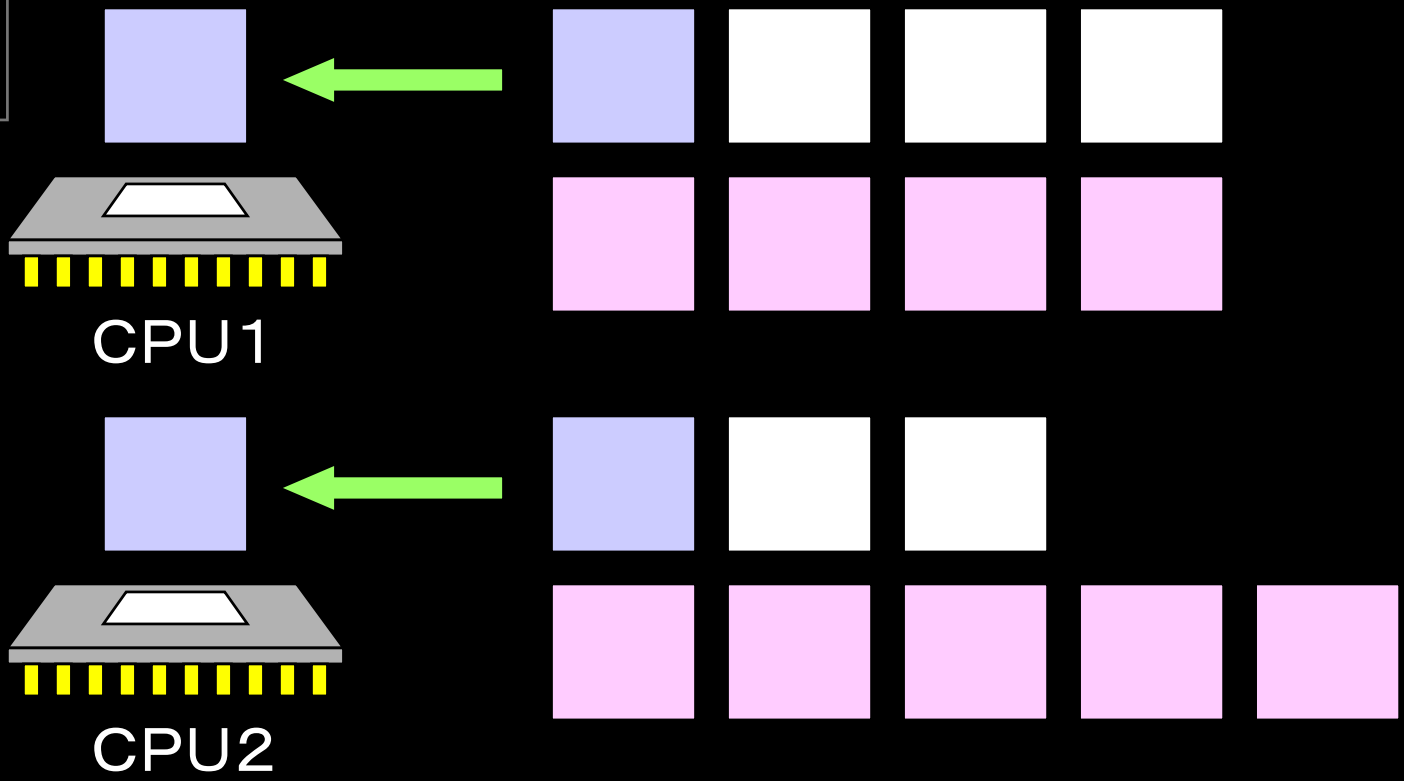


O(1)-スケジューラ

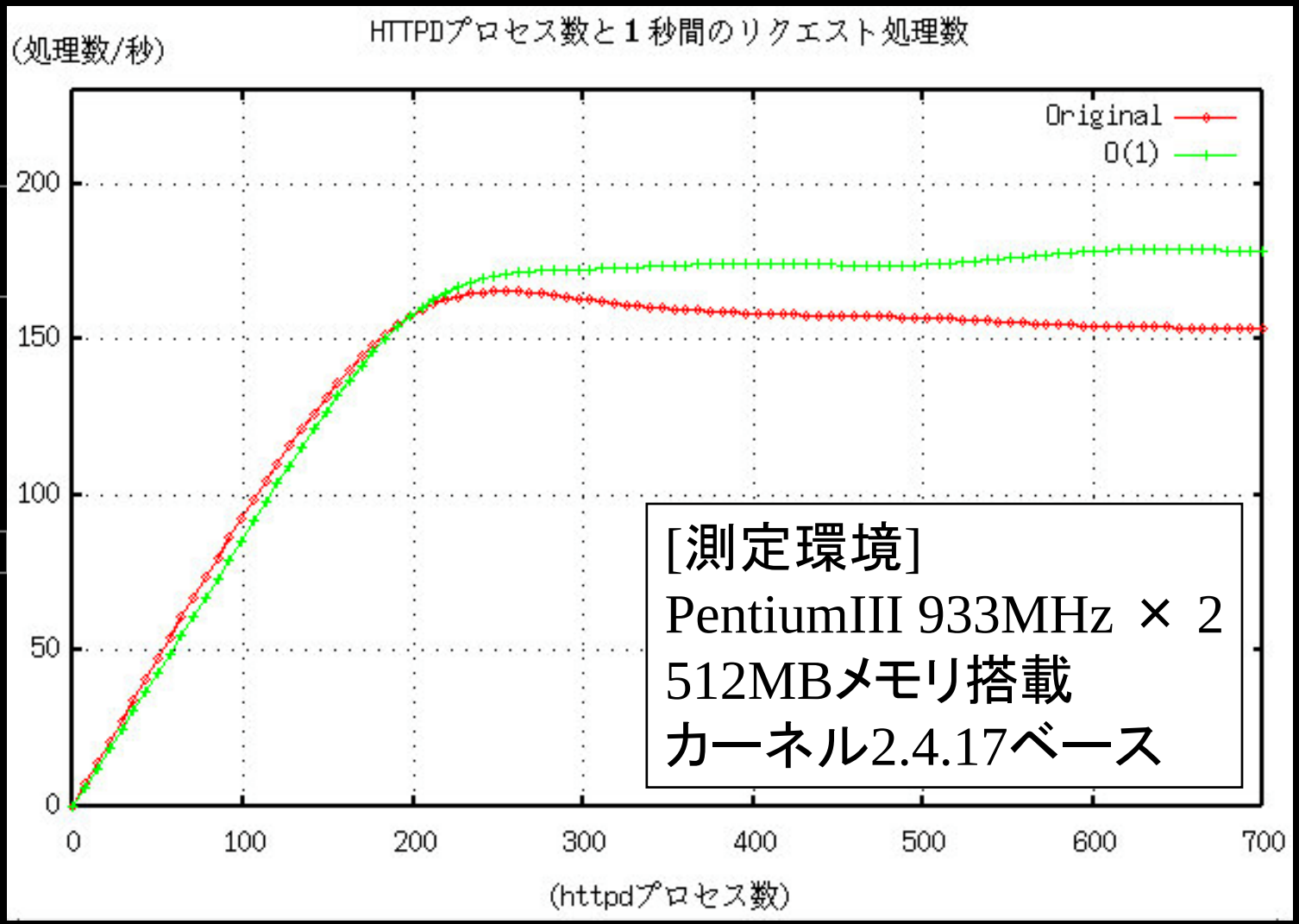
- CPUごとに active/expired キューを保有
  active と expired キューが切り替わる
- 排他処理の削減
- プロセスのプライオリティに応じてキューイング
- x86 BSFL ビットサーチ命令によるすばやいタスク選択
- プロセスのCPU移動を抑止。→タスクのCPU移動に伴うキャッシュバウンスを防ぐ効果

O(1)-スケジューラ

- CPUごとに2つのキュー。



プロセス増加に伴うコスト



MIRACLE

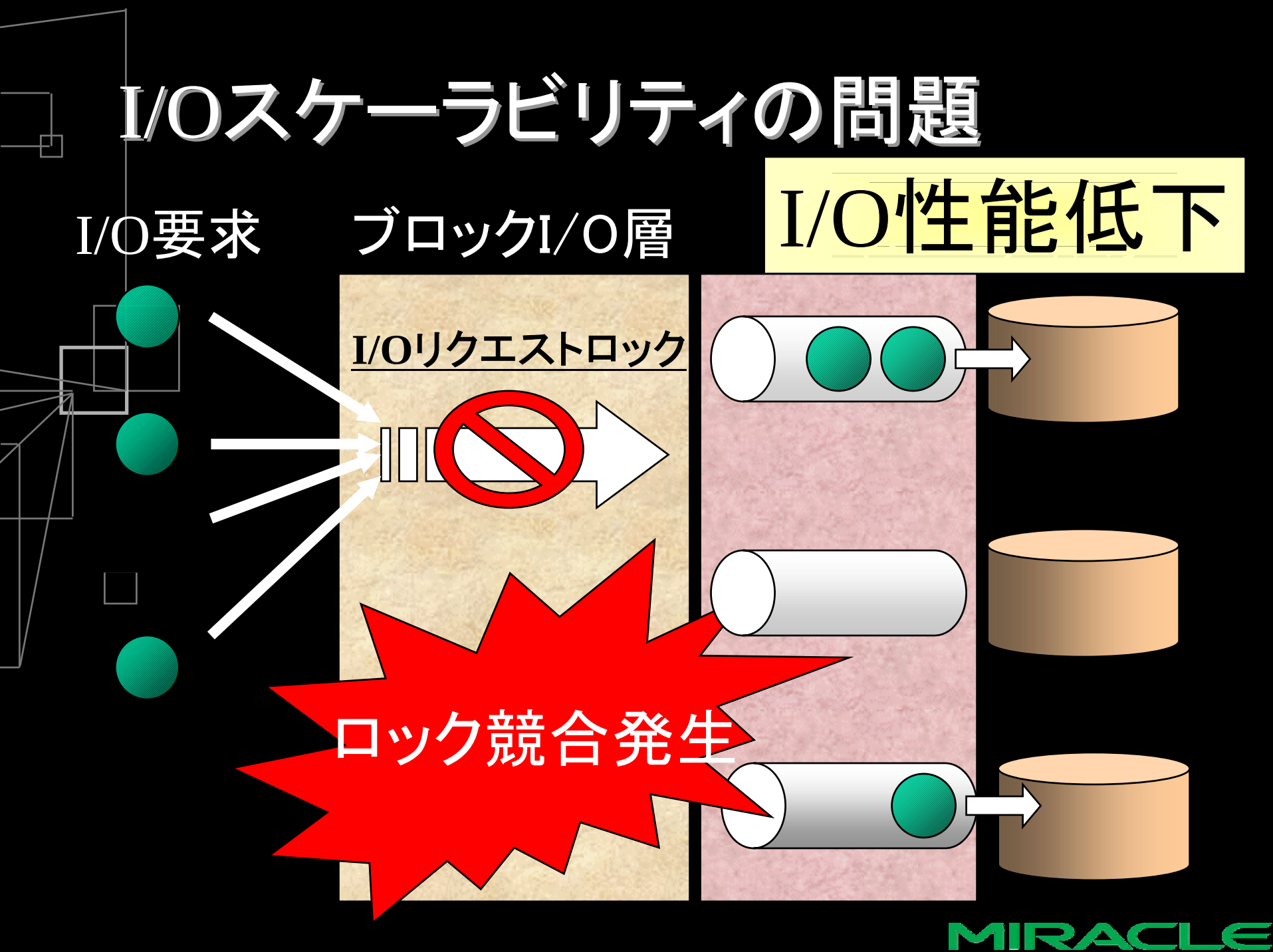
プロセス増加に伴うコスト

- Originalスケジューラはプロセス数増加に伴い、処理可能なリクエスト数減。
  - リストの線形走査のコスト。
  - ランキュー走査のシリアライズ。
- O(1)スケジューラはプロセス数が増加しても処理可能なリスクエスト数に影響しない。

Enterprise領域で求められるI/O処理

- 大容量
  - ✓Kernel 2.4から最大サイズ拡大
- 高信頼性
  - ✓ジャーナリングファイルシステム
- 高可用性
  - ✓LVMサポート
- スケーラビリティ

I/Oスケーラビリティの問題

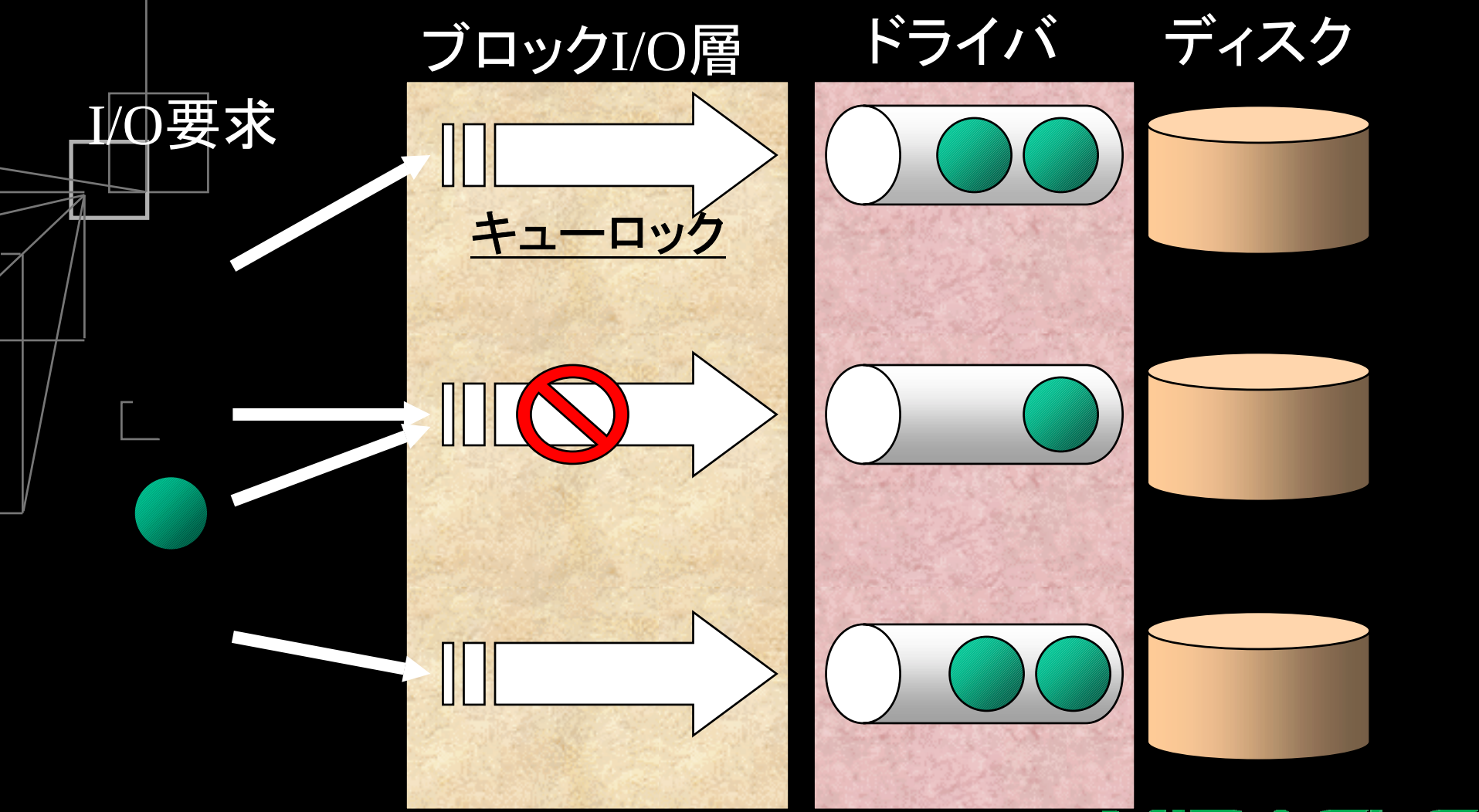


ブロックI/O層の改善

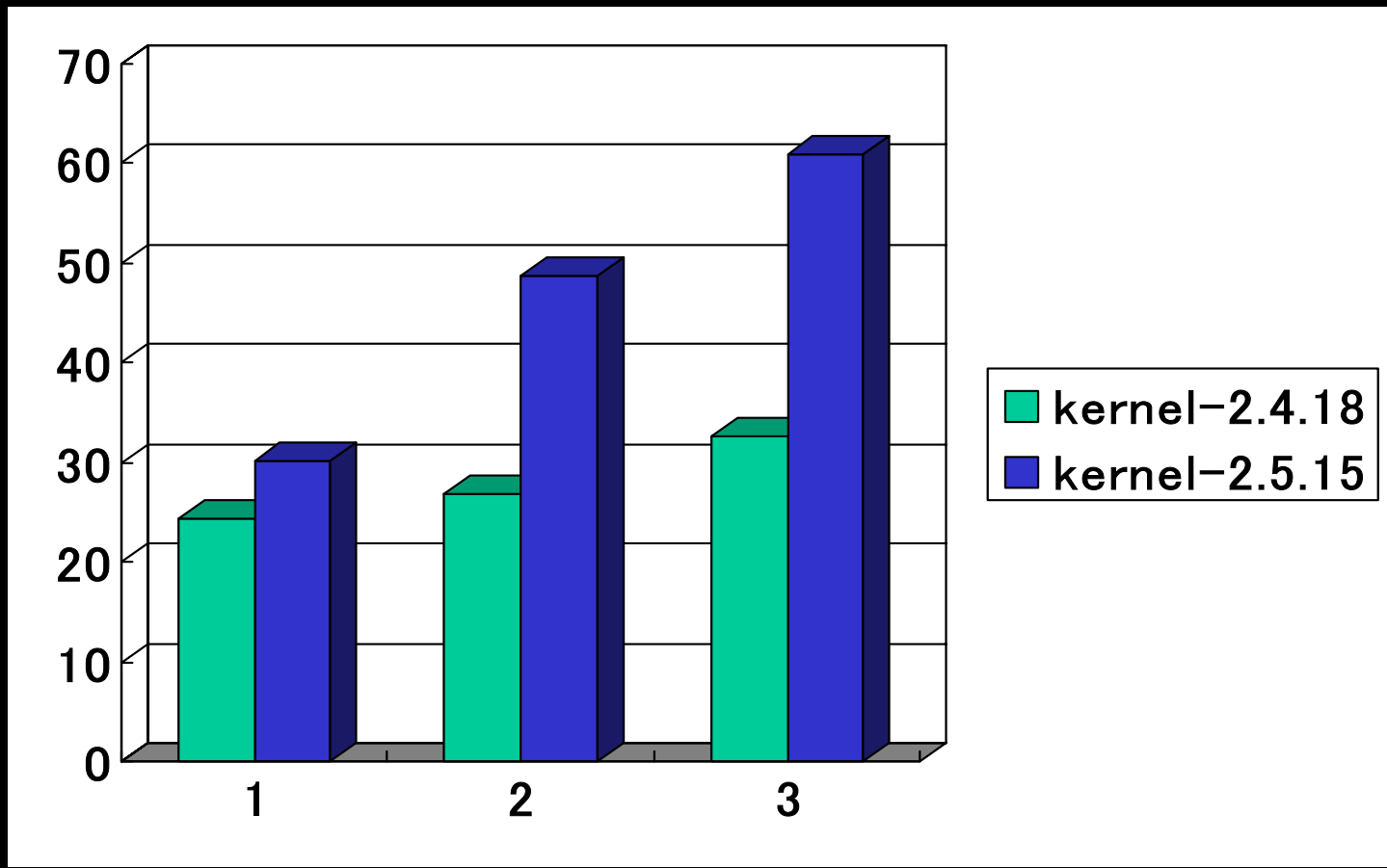
- I/Oリクエストロックの細分化
  - ✓複数デバイス使用時のスケーラビリティの向上
- ブロックI/O層の機能拡張

ロック競合の解消

I/O性能向上



多重I/Oの性能測定結果



複数デバイスを用意することで、
I/O性能の向上を図ることができる

ブロックI/O層の機能拡張

- より細かいI/Oリクエストの制御機能の提供
  - ✓ 例：バリアI/Oリクエスト
- デバイスの特性に合わせて、ブロックI/O層の処理を変更
  - ✓ 例：エレベータアルゴリズム

ただしデバイスドライバ側での対応も必要

ネットワークI/Oの問題

- ネットワーク帯域幅の必要性増加
  - ✓ クラスタシステム、NAS等
- ネットワークデバイスの高速化
  - ✓ ギガビットイーサネットの普及
  - ✓ 10Gbps Ethernetの導入へ

□ カーネルへの影響は？

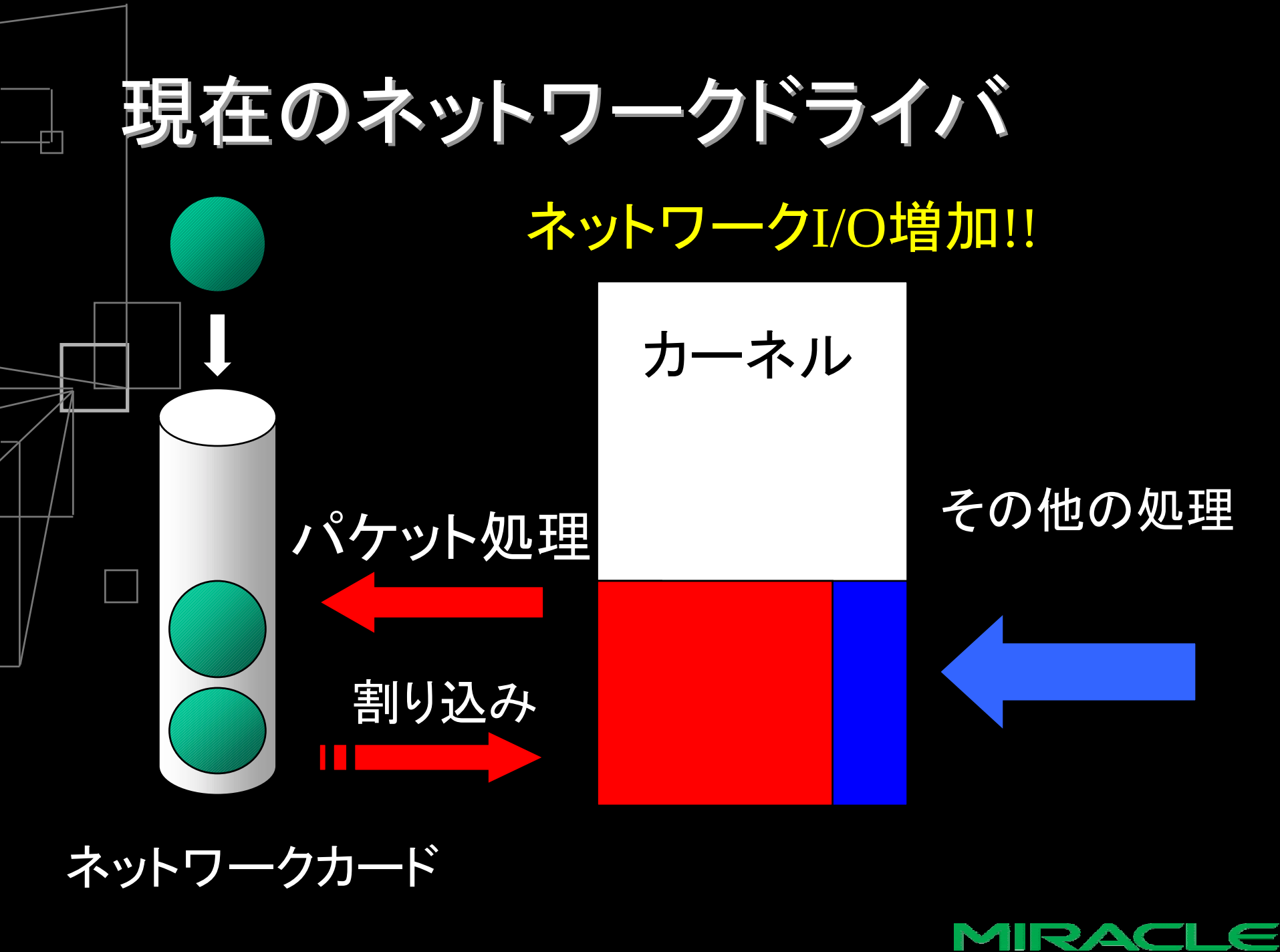
割り込み処理の増加

高負荷によるシステムへの悪影響

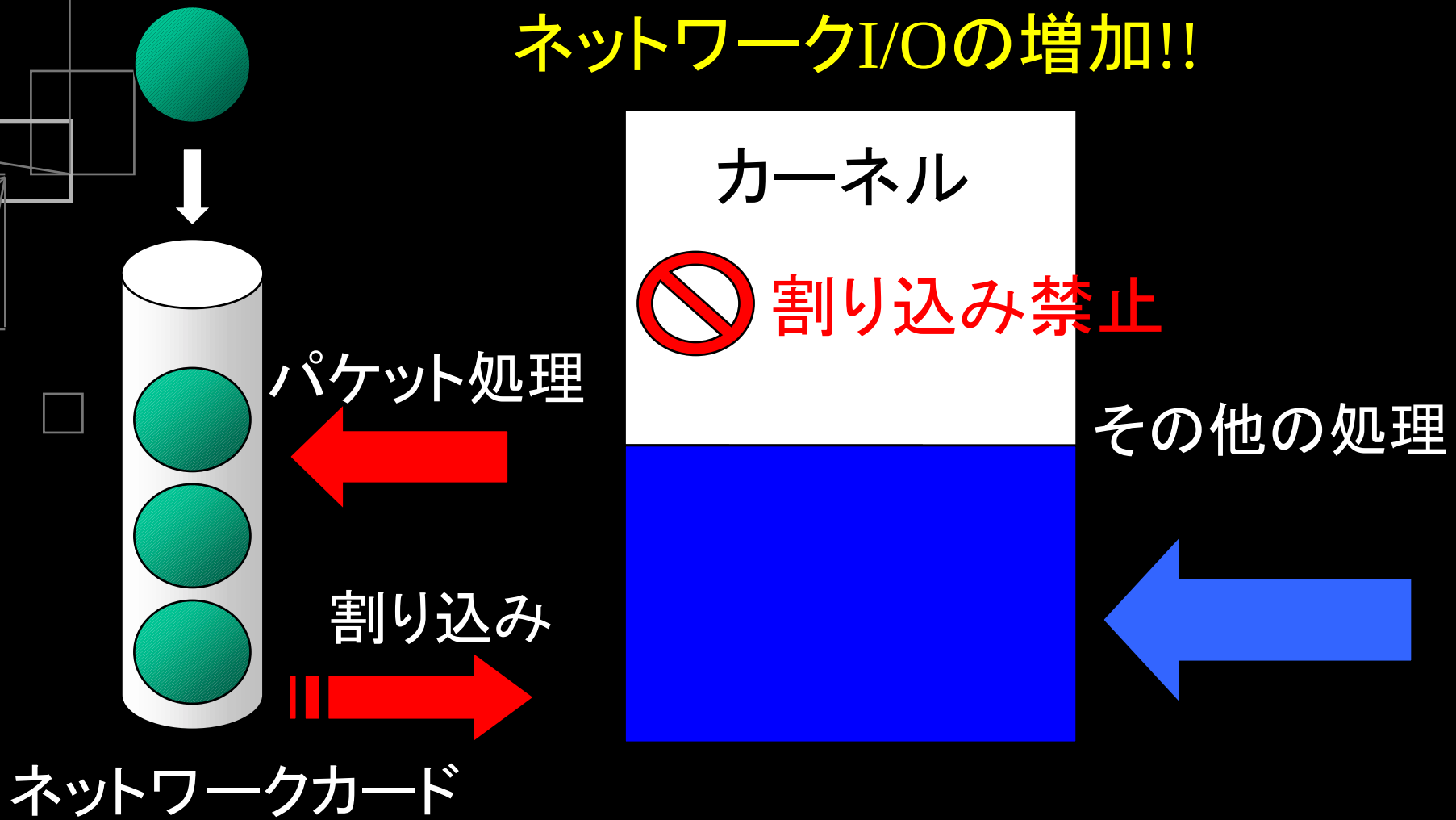
NAPIの導入

- Kernel 2.5.7から導入
- パケット処理方法の変更
- 3C59x,e1000,tulipドライバなど一部のNICのみ
- インターフェースが非互換のため、既存のドライバのままでは利用不可

現在のネットワークドライバ



NAPIの仕組み



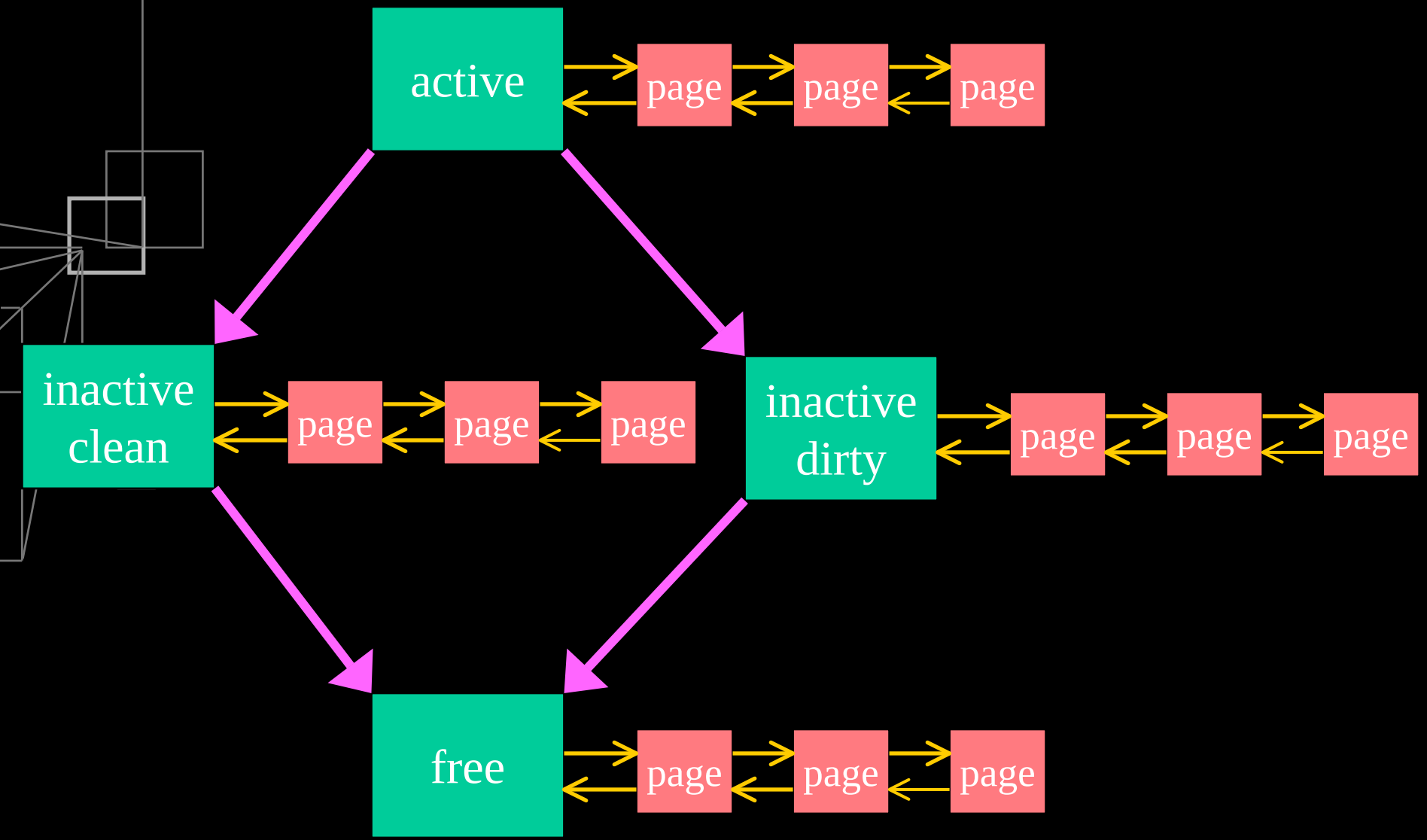
rmap VM: VM への要求

- VM はカーネルのクオリティを決定する中心的な機能
- Enterprise 領域においてはギガ単位のメモリサイズを効率的に管理することが求められる
  - メモリ負荷が高い状況においても、安定して動作する信頼性
  - パフォーマンス

rmap VM:目的と実装

- 効率的にページアウト処理を行うことによって、性能・信頼性を高める。
- 以下の機能を改良/追加することに実現
  - ページの状況に応じた こまやかなリスト管理
  - aging
  - reverse map

rmap VM: ページのリスト管理



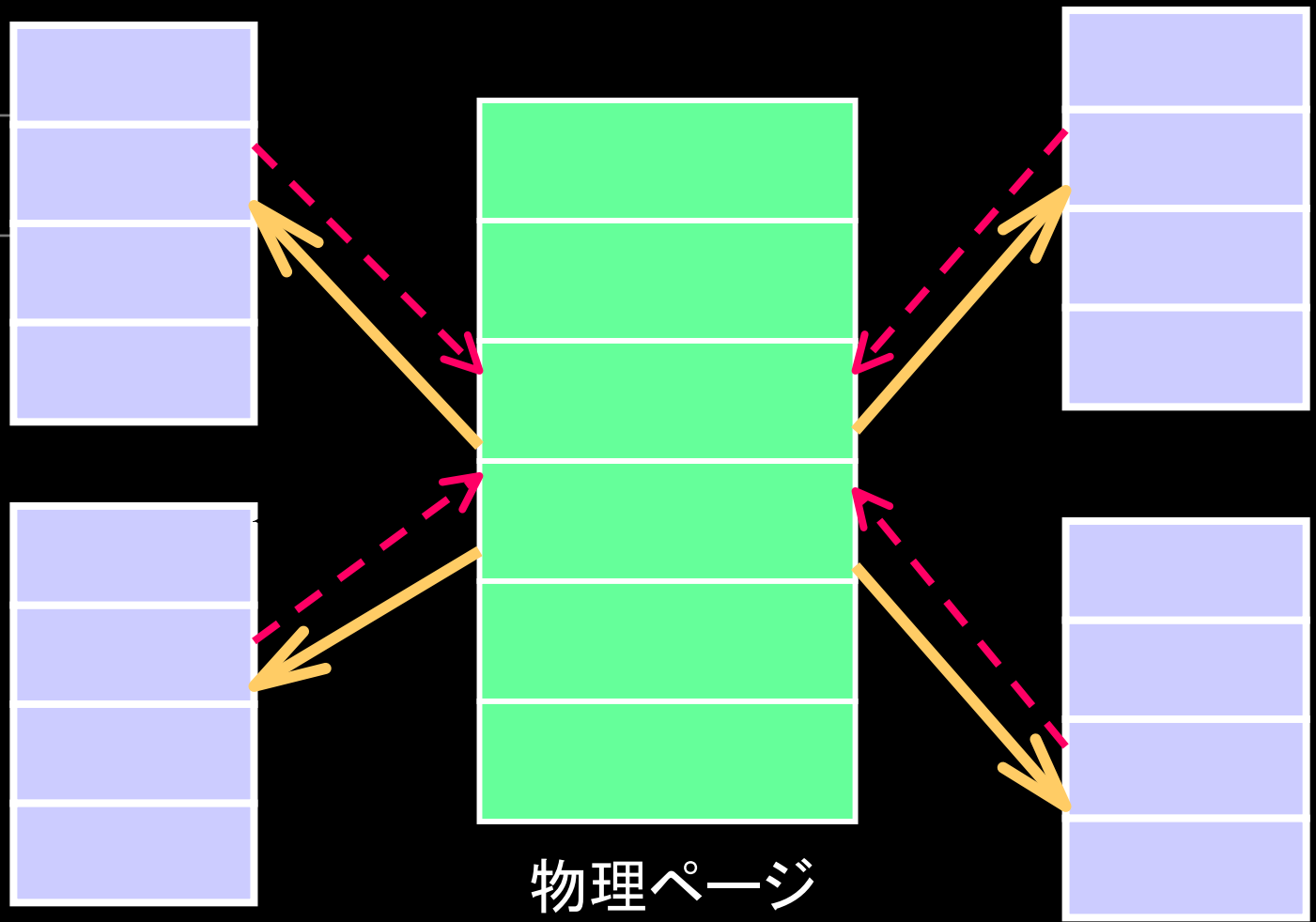
rmap VM: aging

- page out すべきプロセスページの選択
  - [従来] PTE の Access ビットで判断 (x86)
    - 0 or 1
  - [rmap] page [illegible] age カウンタで判断
    - 0 [illegible]

真に不必要なページだけ
page out !!

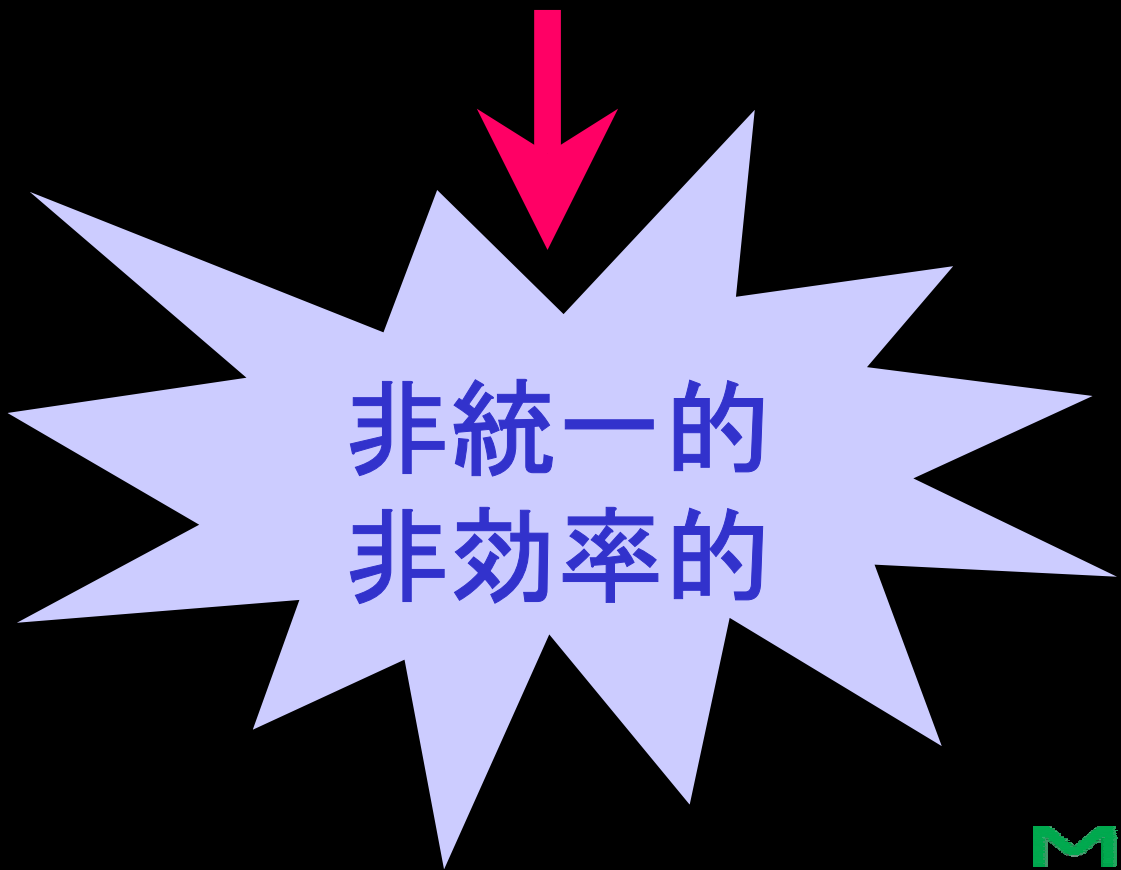
rmap VM: reverse mapping(1)

- プロセスの仮想ページと物理ページの関係



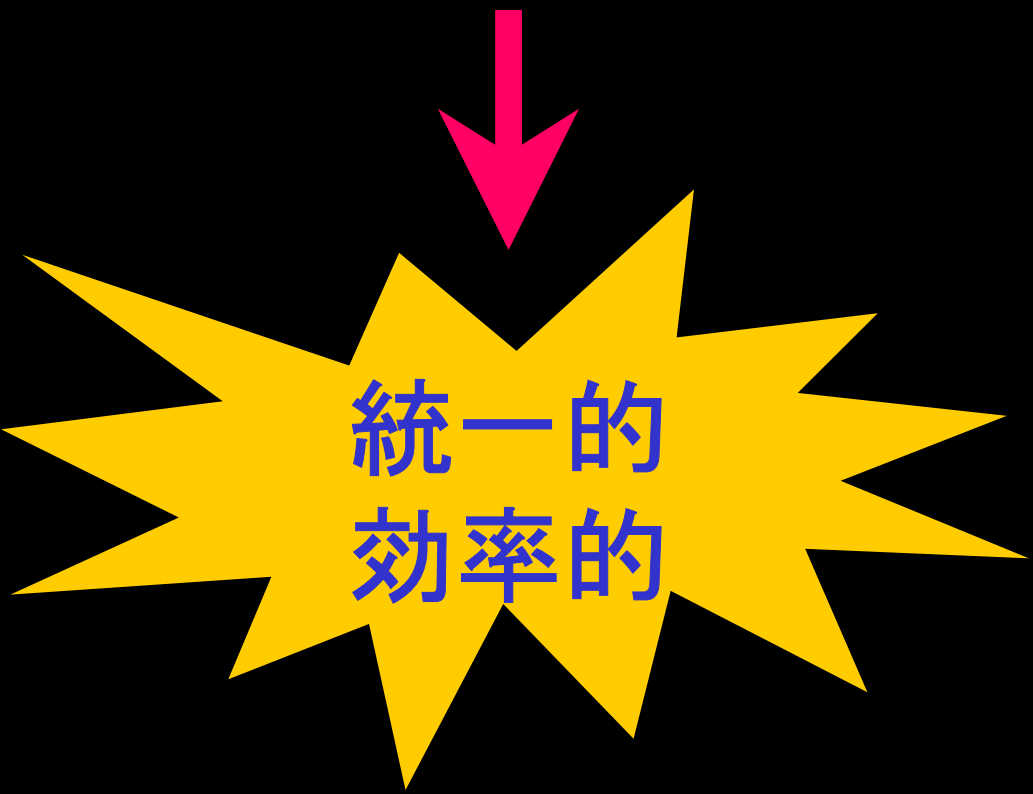
rmap VM: reverse mapping(2)

- 従来
  - ファイルキャッシュは　物理ページから　スキャン
  - プロセスページは仮想ページからスキャン



rmap VM: reverse mapping(3)

- rmap
  - ファイルキャッシュ/プロセスページともに物理ページからスキャン



rmap VM: まとめ

- Page Out に最適なページを抽出するのに手間をかけすぎると性能が犠牲になる
- 簡単に抽出しすぎると必要なページがPage Out してしまう
- rmap VMは相反する両者をバランス良く満たす手法

Reliability/Scalabilityを高めるメモリ管理機構を提供

Enterprise 領域における Miracle Linux の取り組み

- 信頼される Linux OS の提供
- OSDL への積極的な協力
- クラスタソフトウェアとの連携

信頼される Linux OS の提供 (1)

- e ビジネスに安心して使えるサーバ OS を目指す
  - 徹底的な検証
  - コアな問題をも解決するサポート力

信頼される Linux OS の提供 (2)

- 注力するコンポーネント
  - kernel
    - 大規模サーバ向けのチューニング
    - データベースサーバとして十分性能を発揮できるよう VM/IO を改善
    - 4社協業(IBM, NEC, 日立, 富士通)の成果の取り込み
      - LKCD (Linux Kernel Crash Dump)
      - LKST (Linux Kernel State Tracer)

信頼される Linux OS の提供 (3)

- 注力するコンポーネント
  - Web - DB
    - Oracle, postgreSQL
    - apache, php, Tomcat, Java
  - ファイルサーバ
    - NFS, Samba

信頼される Linux OS の提供 (4)

- *Goal !!*

MIRACLE

OSDL への積極的な協力

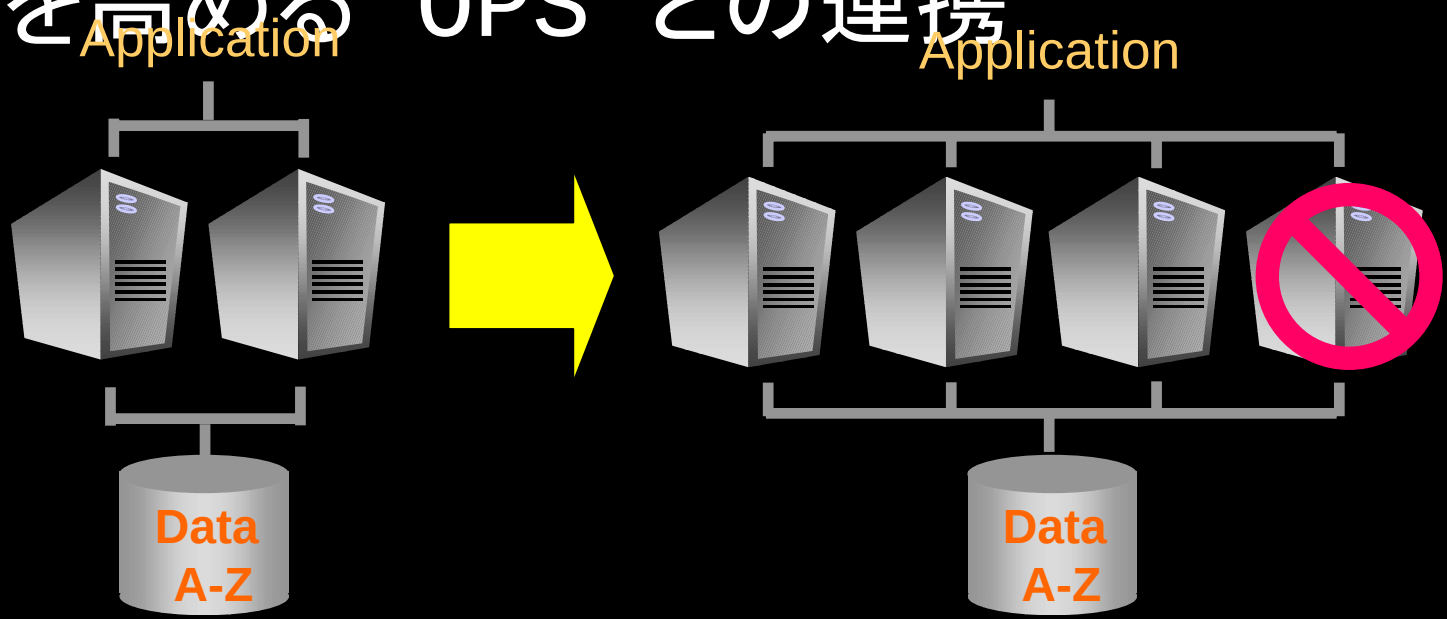
- OSDL
  - Linux の Enterprise 機能を高めるためにオープンソース開発者にリソース等を提供
  - http://www.osdl.org
  - http://www.osdl.jp (日本のサイト)
- Miracle Linux の貢献
  - スポンサー
  - Board Member として参画
  - Samba 関連のプロジェクトを登録し推進

クラスタソフトウェアとの連携 (1)

- Linux Server の Reliability/Availabilityを高めるクラスタソフトウェアとの連携
  - CLUSTERPRO
  - Life Keeper
  - Local Cluster
- MIRACLE LINUX にバンドルした製品も提供

クラスタソフトウェアとの連携 (2)

- Oracle DB の Reliability/ Availability/ Scalability を高める OPS との連携



Oracle Real Application Cluster

【お問い合わせ先】

info@miraclelinux.com

http://www.miraclelinux.com