Linux Kernel V2.5最新動向

Miracle Linux／CTO
OSDL／Board of Directors
よしおかひろたか

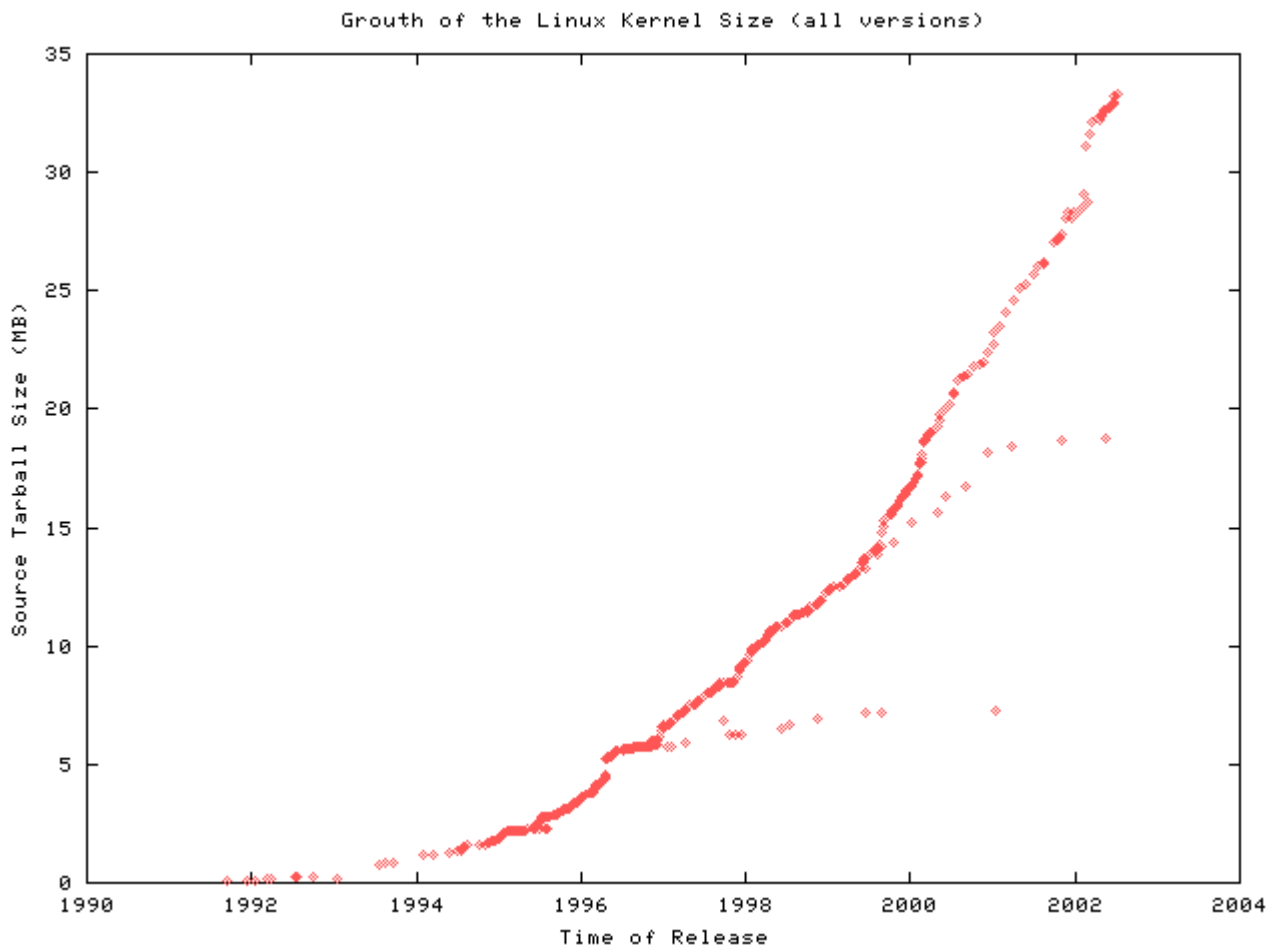
はじめに

- Linuxの動向を主にエンタープライズ系に焦点をあて、紹介する。
- Linuxの歴史
- 現状
- 今後の動向
- OSDLの紹介と役割

Linuxの歴史

- 1991年8月25日、comp.os.minixにLinus Torvalds (21歳)が投稿したのがきっかけで世界に知られることになる。V0.01 tar.gz 71K
- V1.0 1994/3 (1.2MB)
- V1.2 1995/3 (1.8MB)
- V2.0 1996/6 (5.6MB)
- V2.2 1999/1 (12.5MB)
- V2.4 2001/1 (23.2MB)
- V2.5 2001/11 (28.0MB)
- V2.5.28 2002/7/24 (33.3MB) **最新版**

Linuxの規模



- http://virgo.umeche.maine.edu/misc/kernel_size/

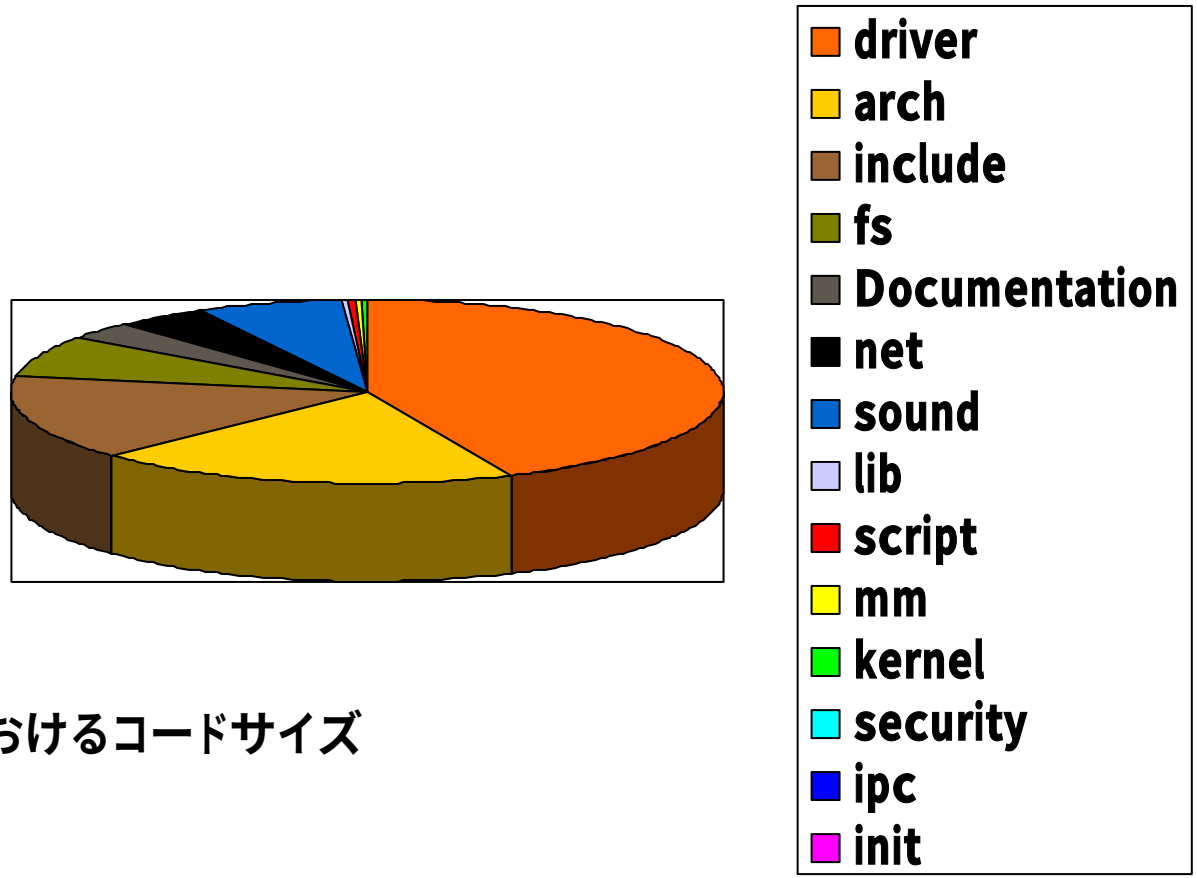
サポートするアーキテクチャ

バージョン	アーキテクチャ数	例
V1.0	1	i386
V2.0	6	alpha,m68k,mips,ppc,sparc
V2.2	7	sparc64
V2.4	13	arm,ia64,mips64,sh,s390,…
V2.5.x	17	crusoe,x86_64,…

サポートするファイルシステム

バージョン	ファイルシステム数	例
V1.0	9	ext/ext2/minix
V2.0	16	fat/ufs/vfat
V2.2	25	autofs/coda/hfs
V2.4	32	jffs/crashfs
V2.5.x	40	jfs/ext3

サブコンポーネントのサイズ



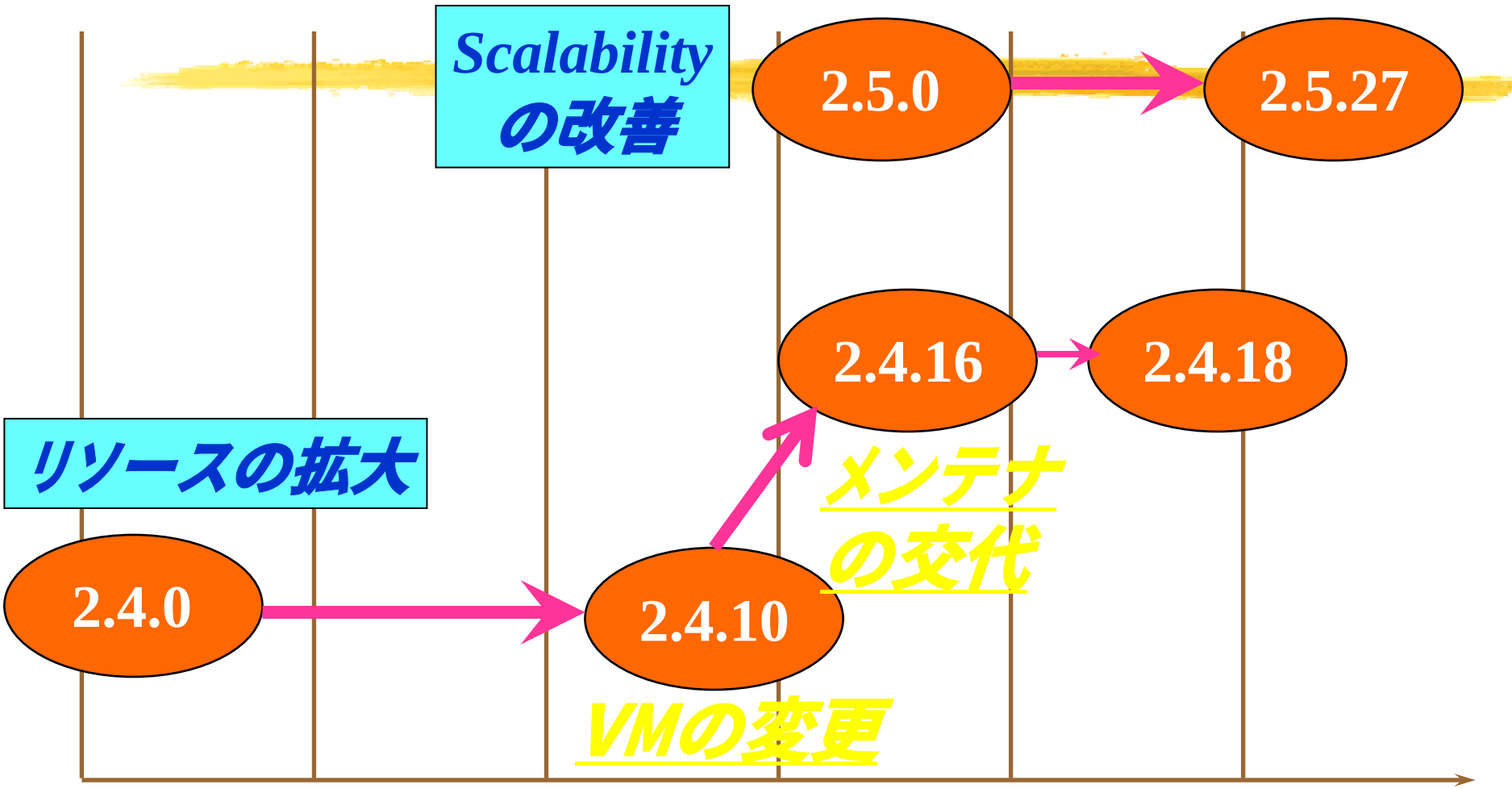
- 2.5.27におけるコードサイズ

Linuxの規模の拡大

- サポートするアーキテクチャ、ファイルシステム、デバイス等はバージョンごとに増加している。
- カーネルのサイズも増加している。

Linux Kernel の動向

kernel 2.4 から kernel 2.5 までの動向



Linux 2.5開発動向

- VMを2.4.10でRik van Riel氏のものから、Andrea Arcangel氏のものへ。
- メンテナーが2.4.16でLinusTorvalds氏からMarcelo Tosatti氏へ交代。
- 同時に2.5.0開発開始(2001年11月)
- Dave Jones氏が、2.4.xのバグフィックスなどを2.5.xへフォワードマージ
- バージョン管理システムBitKeeperの採用

Kernel 2.4の主な新機能 (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

Kernel 2.4の主な新機能 (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

Kernel 2.5の主な新機能 (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

Kernel 2.5の主な新機能 (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の主な新機能(3)

- Support for IDE TCQ(Tagged Command Queueing)
- New Quota System plugins
- Hotplug CPU
- Faster internal kernel clock frequency
- Direct pagecache <-> BIO disk I/O
- Linux BIOS
- Software suspend
- Remove the BIG IRQ lock (2.5.28)

Kernel 2.5の主な新機能(4)

- Linux Security Module
- EVMS (Enterprise Volume management System, beta)
- Dynamic Probes (beta)
- Page Table Sharing (beta)
- Async IO (beta)
- Better event logging for enterprise systems

開発途上のもの

- Scalable Statistics Counter
- Linux Kernel Crash Dumps (LKCD)
- Linux Kernel State Tracer (LKST)
- NFS V4
- Zerocopy NFS
- NUMA support
- Large Page Support

Kernel 2.5の新機能

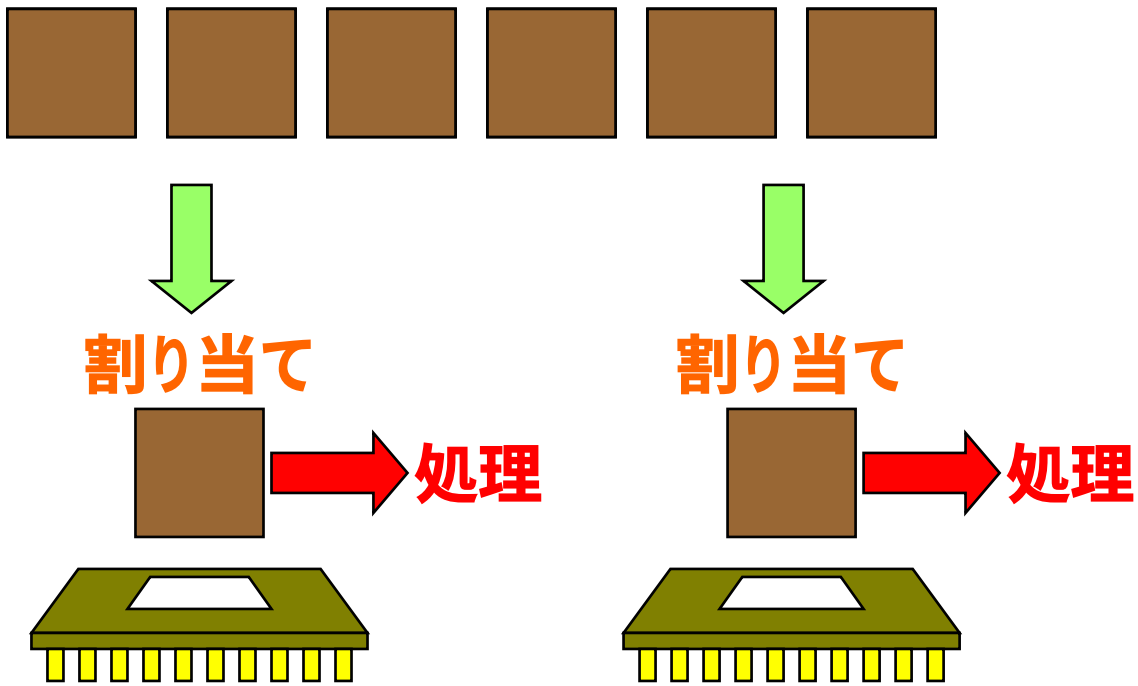
- Scalability を向上させる以下の機能を紹介
  - プロセス管理
    - O(1) スケジューラ
  - I/O
    - block I/O の改善
  - Network
    - NAPI

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。

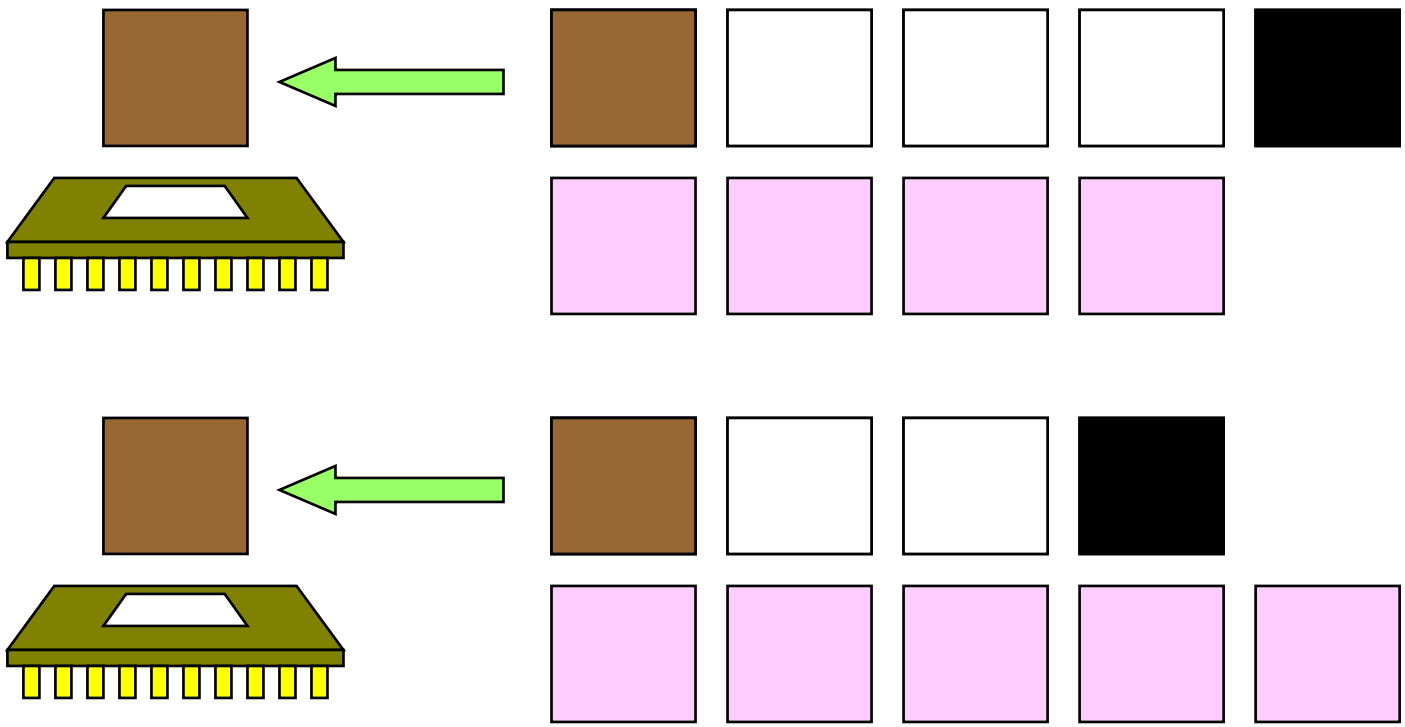


O(1)-スケジューラ

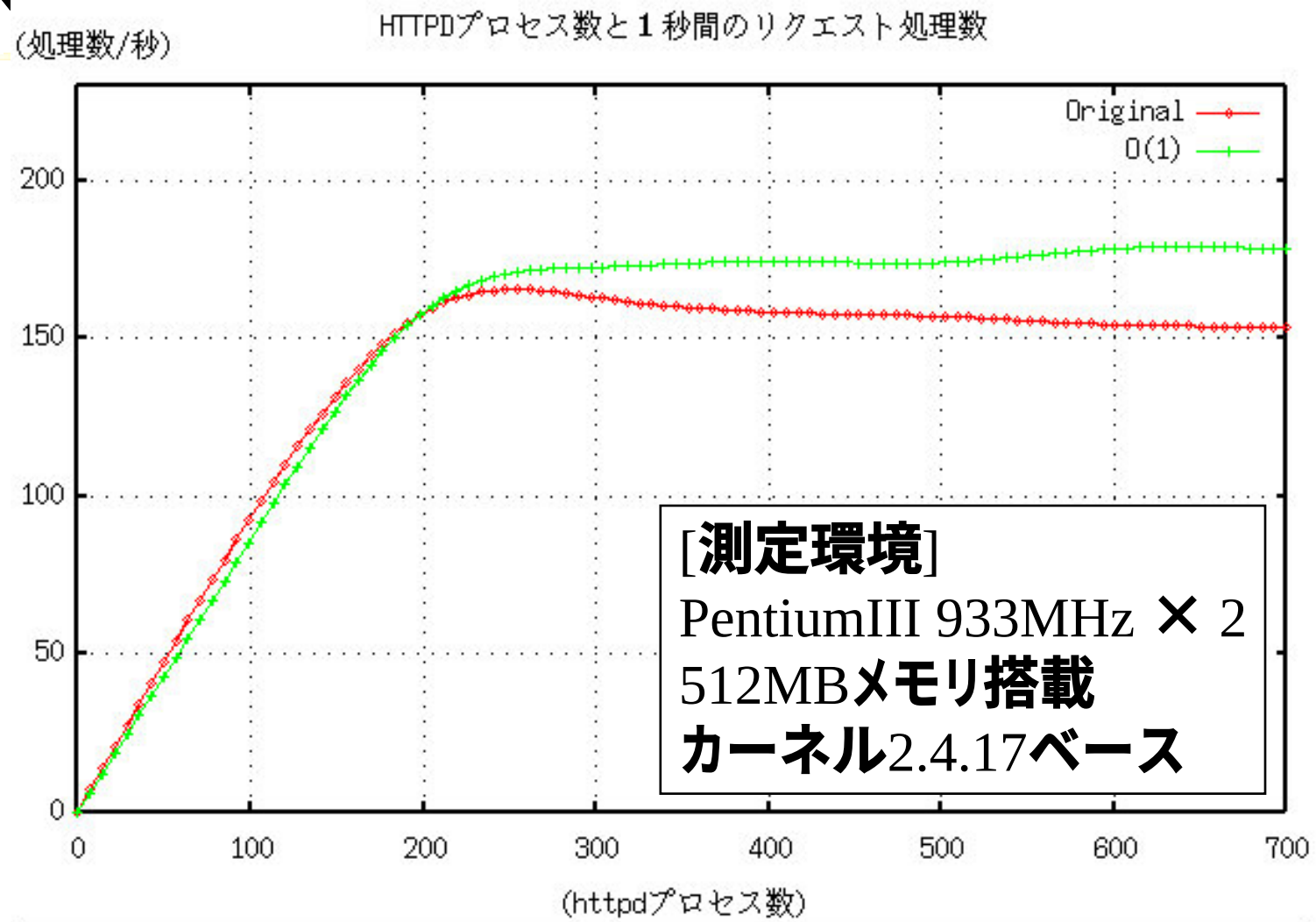
- CPUごとにactive/expiredキューを保有activeとexpiredキューが切り替わる
- 排他処理の削減
- プロセスのプライオリティに応じてキューイング
- x86 BSFLビットサーチ命令によるすばやいタスク選択
- プロセスのCPU移動を抑止。→タスクのCPU移動に伴うキャッシュバウンスを防ぐ効果

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



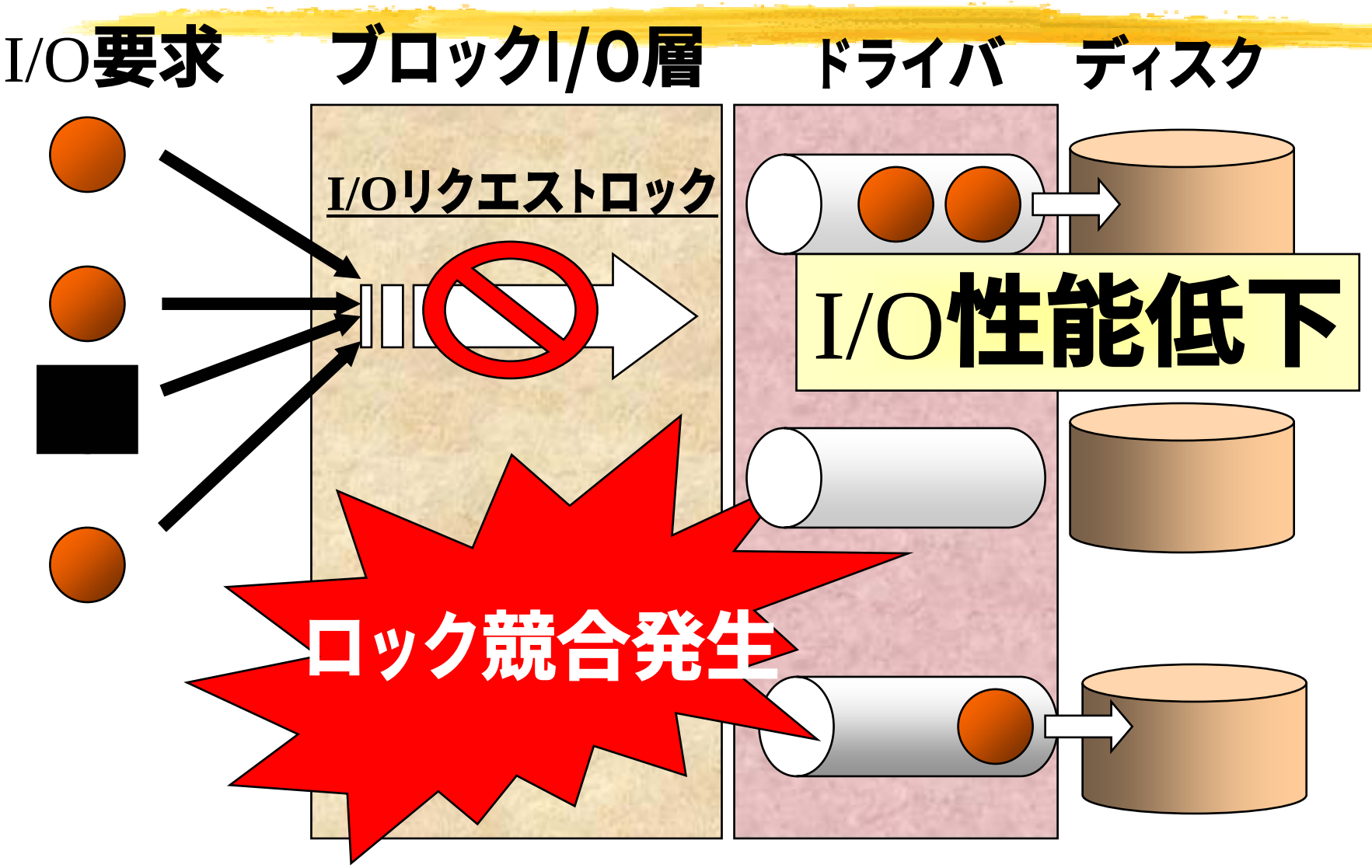
プロセス増加に伴うコスト

- 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。

IOサブシステム

- IOブロックレイヤーの改良

I/Oスケーラビリティの問題

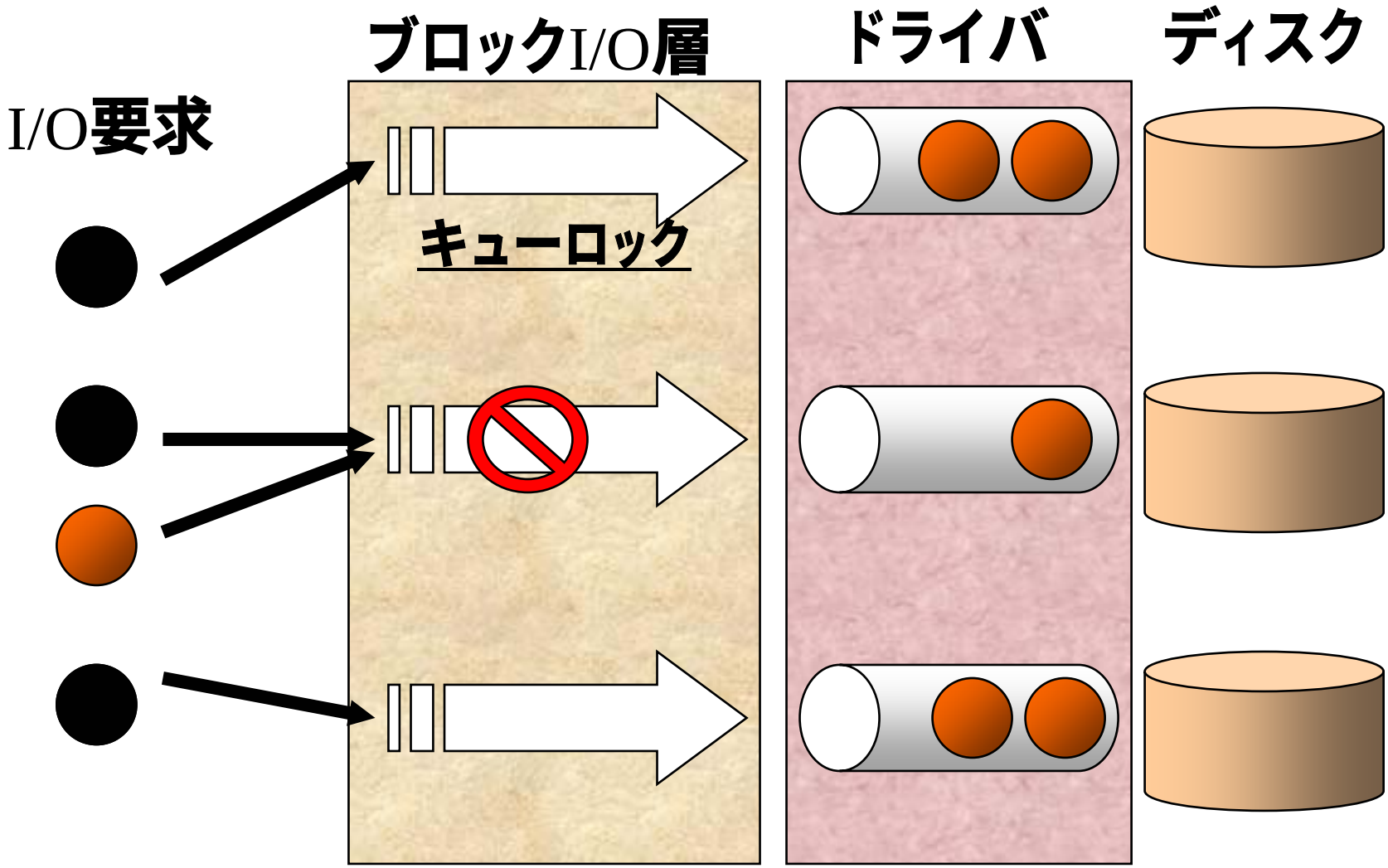


ブロックI/O層の改善

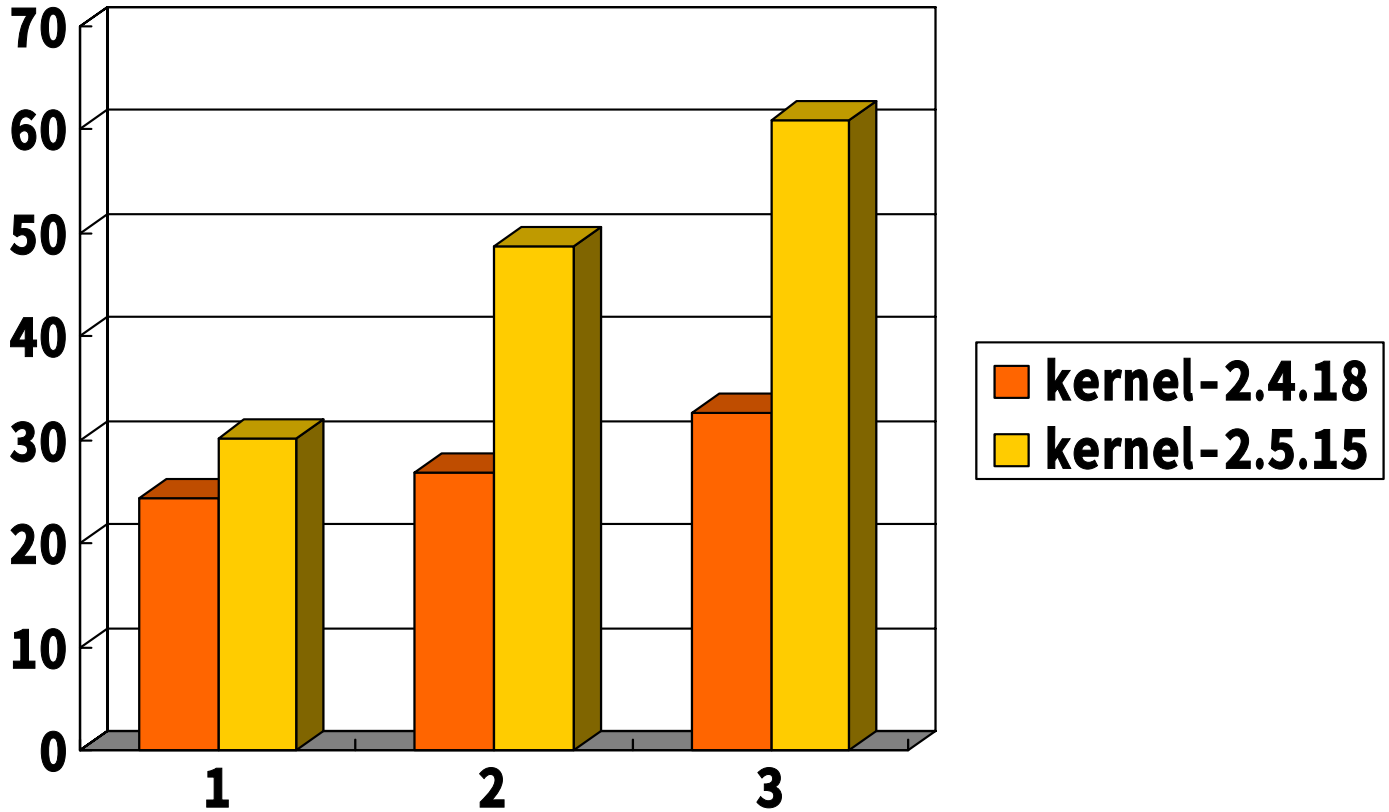
- I/Oリクエストロックの細分化
  - ✓複数デバイス使用時のスケーラビリティの向上
- ブロックI/O層の機能拡張

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

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

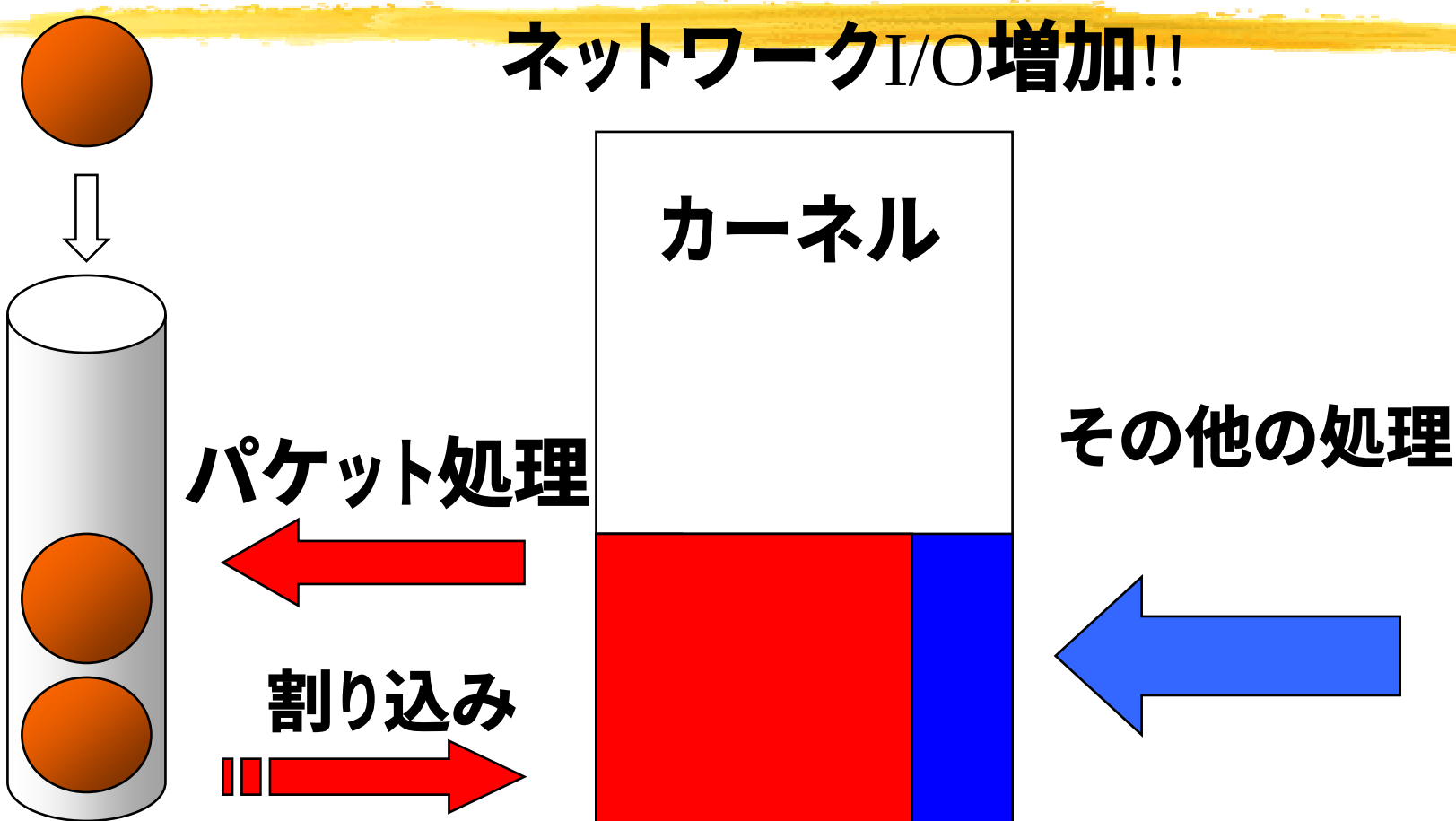
高負荷によるシステムへの悪影響

NAPIの導入

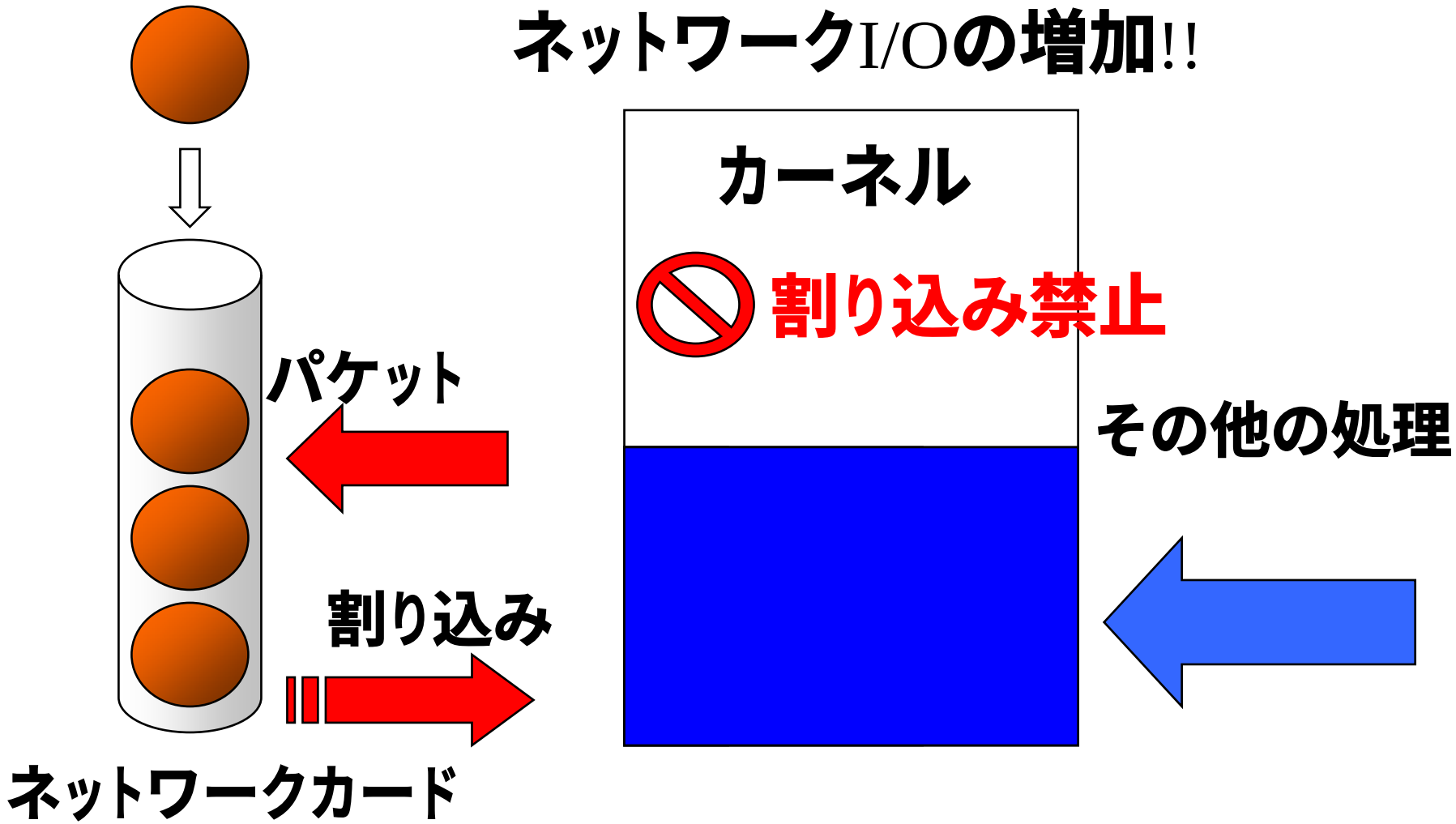
- Kernel 2.5.7から導入
- パケット処理方法の変更
- 3C59x,e1000,tulipドライバなど一部のNICのみ
- インターフェースが非互換のため、既存のドライバのままでは利用不可

現在のネットワークドライバ

ネットワークI/O増加!!



NAPIの仕組み



Linuxはどこへ向かうのか？

- Linuxは誰が開発しているのか？
  - V2.0くらいまで、Linusと仲間たち
  - V2.2くらいまで、カーネルハッカー
  - V2.3くらいから、企業のプロのプログラマが台頭
    例：IBM/SGI/HP/RedHat/SuSE/…

Linux Kernel Summit(LKS)

- 招待されたカーネルハッカーだけが参加できる会議。2001年3月に第一回が開催された。今年は6月Ottawa Linux Symposiumの前に開催された。
  - 今年はVA Linux Japanの高橋さんが参加
- Linuxカーネルの技術について議論する。

Linux Kernel Summit 2001

Linux Kernel Summit 2002

Database からの要求

- LKS2001でOracleのLance Larsh氏はデータベースの観点からみたカーネルへの要求を出した。
  - RAW I/O
  - Elevatorアルゴリズム
  - BLOCK I/Oレイヤ
  - Highmem
  - Async I/O
  - Large Page
  - Shared Page等々

Linux Kernel Summit

- LKS2001で議論された多くの項目は2.5に取り込まれた。
- LKS2002で議論されたいくつかの項目は2.6ないし3.0で取り込まれるだろう。
- 要求を明示するのは非常に重要である。

Linux Roadmap

- 古典的なバザールモデルは
  - 予算や権限、スケジュール責任を持ったプロジェクトリーダーはいない。コアメンバーはプロジェクトの調停役。
  - 明確なロードマップはない。
- ユーザーは、今後の動向が読めない。
  - 何らかのロードマップが必要

Linux Roadmap

- 必要とする機能の記述
- 開発プライオリティ
- 開発コミュニティ、実装 (あれば)

Linux Roadmap

- 開発者にとって
  - 開発の指針
  - リアルな顧客要求
  - 誰がどのような機能を開発しているか
- ユーザーにとって
  - 必要な機能が開発されるかの参考
  - Linux導入意思決定に利用

Example: Data Center Linux Requirements list

- Priority 1
  - Existing
    - Journaling File Systems
    - Volume Management
  - New Features
    - Clustered volume manager
    - Filesystem Access Control Lists
    - CPU hot plug
    - Kernel debug support: debugger, trace, crash dump
    - OS configuration and change logging
    - Diagnostics
    - Async I/O
    - Remote console

Data Center Linux (DCL)

- Priority 1 (Cont.)
  - Improvements
    - Hardened SCSI and Fiber Channel drivers
    - Very large file and filesystems
    - Multi-path I/O
    - Tape support
    - Posix N:M thread support
    - Event logging
    - Scaleability improvements
    - I/O performance improvements
    - Open Cluster API and tool set
    - NFS performance
    - Virtual Memory for large databases

OSDLの役割

- エンタープライズ系のLinuxの開発を推進することを目的として設立されたNPO
  - 大規模なコンピュータ資源を提供
    - テスト環境(Scalable Test Platform)
    - プロジェクト支援
  - Data Center Linux/Carria Grade Linuxの推進
  - Roadmapの策定
  - http://www.osdl.jp/
  - http://www.osdl.org/

OSDLを利用するプロジェクト

- プロジェクトリーダーはOSDLにプロジェクトの計画を提出し、施設の利用のスケジューリングをうける。
- OSDLはオープンソースプロジェクトを支援する。OSDLがプロジェクトを運営するわけではない。

OSDLのプロジェクト例

- **実施中**
  - Database Opensource Test Suite
  - Database Test One Development
  - I/O scalability
  - filesystem workload
  - kernel benchmarking and measurement
  - Linux Scalability Effort (LSE)
  - Linux Super Page Kernel
  - Score on IA64

OSDL Projects

- **提案中**
  - Distributed PostgreSQL
  - I/O subsystem Enhancement
  - Linux Kernel State Tracer (LKST)
  - Samba Large Resource Availability

OSDLインフラストラクチャ

- **大規模な計算機資源**
  - 1 way　20**台以上**
  - 2 way　60**台以上**
  - 4 way　10**台以上**
  - 8 way　10**台以上**
  - 16 way　2**台以上**
  - IA64　10**台以上**
  - storage ＞ TB

Scalable Test Platform (STP)

- Linux Kernelの自動テスト環境である。
- Kernelの各バージョンがあらかじめ準備されている。
- 利用者は独自のパッチをテストすることもできる。
- 標準的なベンチマークテストなどもあらかじめ準備されている。

STP(Scalable Test Platform)

- **利用可能なベンチマーク例**
  - aim9
  - bash-memory
  - bonnie++
  - dbench
  - iozone
  - LTP
  - lmbench
  - Unixbench
  - **等々**

STPの利用

- 利用者はSTPにパッチの登録をメールで依頼
- STPからパッチIDを得る
- STPを利用してテストを起動する
- STPはメールで結果を利用者に知らせる
- 結果はWebに登録されるので、利用者はレビューする。
- 上記のプロセスは自動化されている。

OSDLプロジェクト募集中

- エンタープライズ系に注力したLinuxないしオープンソースプロジェクトを募集している。
  - Scalability
  - Reliability
  - Availability
- 個人では用意できない大規模環境を提供
- 日本からの積極的な利用を〜
- まづはSTPを利用してみよう。

Linux開発にどう参加するか？

- ベンチマーク？
- パッチの開発？

おまけ：カーネル読書会

- YLUG(横浜Linuxユーザー会)の有志で不定期(1〜2ヶ月に一回)に開催。
  - Linuxカーネルを肴に楽しむ Just for fun
  - 通常OSDLで開催
  - http://www.ylug.org/

Links

- **カーネルのステータス**
  - http://www.kernelnewbies.org/status/latest.html
  - http://kerneljanitor.org/ (To do list)
- **カーネルソース**
  - http://www.kernel.org/

まとめ

- Linuxの動向を紹介した
- OSDLについて紹介した