解　説

ミッションクリティカルシステムのためのLinux

真鍋義文　日本電信電話（株）NTT サイバースペース研究所　manabe.yoshifumi@lab.ntt.co.jp

ミッションクリティカルシステムは，企業の基幹業務等に使用されるコンピュータシステムのことを指し，365日24時間無停止での稼働が求められることが多い．また，ミッションクリティカルシステム上のデータの紛失や漏洩はその企業にとって重大な損失・信用失墜を招きかねない．したがってミッションクリティカルシステムには非常に高い信頼性・可用性・セキュリティ等が要求される．Linuxカーネルの機能拡充および安定化に伴い，ミッションクリティカルシステムにもLinuxを導入する気運が近年高まってきている．本稿では，ミッションクリティカルシステム向けのLinuxについて概説する．

カーネルの機能拡充状況

オープンソースのオペレーティングシステムの代表となっているLinuxは，1991年に最初のバージョン0.01が当時ヘルシンキ大学の学生であったLinus Torvaldsによってリリースされた．それ以降，Linus Torvaldsを中心とした開発コミュニティによるいわゆる「バザール方式」と呼ばれる開発方式により，多くの改良・機能追加が行われてきている．

バージョン2.6における新規機能

2003年12月に公開されたLinuxカーネルバージョン2.6では，ミッションクリティカル向けの機能が多く取り込まれており，すでに多くの商用ディストリビューションで採用されている．主な機能としては，大規模なシステムを構築することを可能とする，64bitCPU対応，非対称マルチプロセッサ対応や大規模ファイルシステムのサポートなど，スケーラビリティ関連の機能があげられる．

大規模システム向けの機能変更としては，スケジューラの変更があげられる．2.6ではO(1)スケジューラと呼ばれる新たなスケジューラが導入された．2.4以前のスケジューラでは，マシン上に1つのランキューのみが存在していた．したがって，マルチプロセッサ環境においては，複数のCPUが1つのランキューを取り合い，ランキューを獲得できなかったCPUがアイドル状態に陥ったり，CPU間でプロセス移動が頻繁に発生したりする．このため，マルチプロセッサシステムでは性能のスケーラビリティが良くないという問題点が指摘されていた．2.6ではこの問題を解決するために，CPUごとにランキューを用意している．これにより，原則的に同じプロセスは同じCPU上で実行されることとなり，キャッシュ等を有効に利用できる．CPUごとの負荷に偏りが発生した時にはキュー間でのプロセスの移動を行い，バランスを取る．また，1つのランキューは優先度ごとのスロットを用意しているため，プロセスの存在している優先度最大のスロットから次に実行するプロセスを選択すればよい．このようにプロセス選択の手間が実行可能プロセス数に依存しないという意味で，このスケジューラはO(1)スケジューラと呼ばれている．

また，セキュリティ対応としてSE Linux（Security-Enhanced Linux）がマージされた．SE Linuxは，米国防省が定めたセキュリティ評価基準TCSEC（Trusted Computer System Evaluation Criteria）のB1クラス相当の強制アクセス制御機能を提供する．この機能を用いると，

ポリシーファイルにおける設定に基づいたアクセス制御を，すべてのユーザとプロセスに強制することが可能となる．これにより，あるユーザに対し与えてはならないアクセス権をポリシーに反して設定することはrootであってもできなくなる．これにより，システムのセキュリティを大幅に向上させることが可能となる．

このほかにも，マルチパス入出力，ダイレクト入出力，IPv6，ゼロコピー NFS，IPSecなどの機能が導入された．

最近の追加機能

2.6リリース後，最近になってカーネルに導入されたミッションクリティカル向け機能としてInfiniBand対応，kexec, kdump, CPUホットプラグ, relayFSなどがあげられる．

InfiniBandは，次世代の高速インターコネクト規格である．1チャネルで一方向あたり2.5Gbpsの帯域幅が規定され，全二重通信と1本のケーブルに最大12チャネルを収容可能であることにより最大60Gbpsの帯域幅の通信を可能とする．プロセッサに対する負荷が小さく，低レイテンシという特長を持っているため，クラスタシステムにおいてInfiniBandは有用と考えられている．

kexecは実行中のカーネルから他のカーネルを直接ブートして実行する機能である．いわゆるBIOSやカーネル初期化の部分を実行することなくkexecは新しいカーネルの起動を行う．これにより，システムのリブートを高速に実行することが可能となる．この機能を用いてkdumpが実現されている．

kdumpは異常時にカーネルのダンプを取るメカニズムである．カーネル異常時には，カーネルが正しい動作を行うとは限らないので，異常が発生したカーネルによる正しいダンプ出力は期待できない．したがって，ダンプ取得機能を通常のカーネルの外に置いておく．これがkdumpの基本的な考え方である．kdump実行のため，カーネル初期化時にkdump用のメモリ空間を確保しておく．この空間は通常のカーネル実行時には使用されない．panic発生など動作中のカーネルがクラッシュしたときには，kexecを用いて，確保されていた領域にある，ダンプを取るためだけの機能を持ったカーネルを起動する．この時BIOSによる初期化は行われないので異常が発生した時点のメモリイメージは保存されている．これにより異常時にもメモリイメージをダンプとして取得することが可能となる．これは異常発生時の障害解析を容易にする機能である．

ホットプラグは動作中にデバイスを抜き差しする機能である．CPUやメモリのホットプラグによりシステムを停止させることなく故障したユニットを交換することが可能になり，可用性は大きく向上すると期待される．

relayFSは仮想ファイルシステムであり，カーネル空間からユーザ空間への大量のデータ転送を高速に行うことを目的として作成された．従来，たとえばデバッグのためprintkを挿入して実行し，大量の情報を書き出した場合にデータを紛失することがあった．これは通常，printkの出力をファイルに書き込む場合にユーザプロセスであるsyslogd経由で行うので，カーネル実行中にsyslogdが動作しないため書き出しが間に合わなくなるためである．relayFSは独立な通信チャネルを確保することにより，カーネル空間からユーザ空間への大量のデータ転送を高速に行っている．この機能はデバッグ以外の用途に利用できる可能性も高い．

カーネルリリース方針の変更

2005年にはカーネルのリリース方針が変更された．従来は2.3, 2.5といったマイナーバージョンの番号が奇数のものは新規機能を導入して試験を行う開発版で，2.4, 2.6といったマイナーバージョンの番号が偶数のものでは新規機能の導入は行わずに機能の安定化を目標としており，安定版と呼ばれていた．しかし，カーネルの機能が複雑化してきているために，このリリース方式ではマイナーバージョン間のリリース間隔が大きくなってきて新規機能が安定版になかなか導入されないという問題を生んでいる．

したがって，今後のリリースについては以下のような方針が採られることとなった（**図-1**）．

- 開発者からの提案（patch）はAndrew Mortonの管理するmm-treeにマージされる．このmm-treeが開発版に相当する．
- mm-treeにマージされたpatchのテストが行われる．
- Linus Torvaldsがmm-treeからmainlineに採用できるpatchをピックアップし，2.6.x-rc（release candidate）としてリリースする．
- 2.6.x-rcのパッチはさらにテストされ，問題がなければバージョン2.6.xとして正式にリリースされる．
- リリース後のバグフィックスは2.6.x.yの形でリリースされる．

この方針変更により，以前から言われていた2.7の開発開始は当面行われないと思われる．

上記のリリース方針において，すべてのpatchがmm-treeを経由しているわけではないことや，rc版に対するテストが不十分であるためにリリース後のバグフィックスが多く，カーネルがなかなか安定しないことなどの問

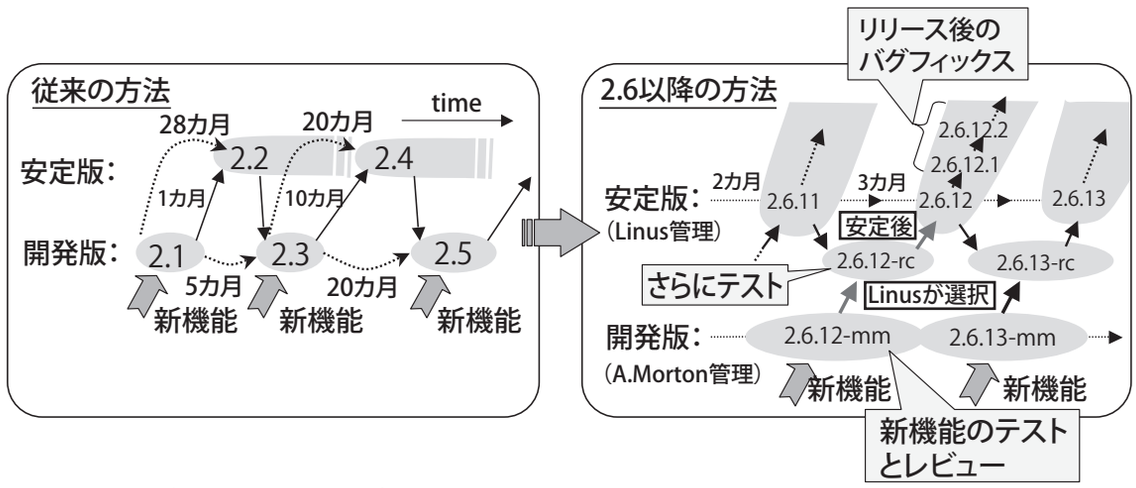

図 -1　Linux カーネルのリリース手順の変更

題点がすでに指摘されている．rc に対するリリース前テストのさらなる強化など，改善すべき点はまだ多いと思われる．

今後導入が期待される機能

今後カーネルに導入される可能性の高いミッションクリティカル向け新機能として仮想化（virtualization），クラスタファイルシステム（cluster filesystem）などがあげられている．

仮想化技術は，1 台の計算機を複数の計算機であるかのように論理的に分割し，複数の OS（ゲスト OS）を同時に動作させる技術である．これにより，Linux, FreeBSD, NetBSD, Windows などの異なる OS を同一の計算機上で同時に実行できるのでサーバ台数を減らすことが可能となる．仮想的に別々の計算機上で実行されているために，あるゲスト OS を停止させる必要が生じた時にも他のゲスト OS はその影響を受けることなく実行可能である．よって，複数のアプリケーションを同一の OS 上で同時に実行させるより安全に実行可能となる．また，多数のプロセッサを搭載する計算機において，昼間はオンラインジョブ処理サーバに多くのプロセッサやメモリを割り当て，夜間はバッチジョブ実行サーバに多くのプロセッサやメモリを割り当てるなど，リソースの柔軟な割り当てにも利用可能である．

クラスタファイルシステムは，ローカルファイルシステムと NFS の組合せよりも処理速度が速く，Linux クラスタでファイルを共有する場合に有用と期待される機能である．クラスタファイルシステムでは，複数のノード間で仮想的に単一のファイルシステムを共有できる．複数ノードへのファイルの複製配置により，故障時のフェイルオーバも可能である．

また，ミッションクリティカルシステムにおけるローカルファイルシステムには従来のジャーナリングファイルシステムより高い可用性が必要であるため，NTT を中心として Linux 上のログ構造化ファイルシステム NILFS の開発を行っている[1]．2005 年 10 月に最初のバージョン 1.0.0 が公開された．ext2 などの伝統的なファイルシステムでは，ファイルシステムに対する書き込み操作（ディレクトリやファイルの作成，削除，移動など）を行う場合にディスクブロックの上書きを行う．この書き込み時に電源障害が発生した場合に，そのデータブロックの内容は保証されない．

このような問題に対処するために ext3 などのジャーナリングファイルシステムが開発されてきた．ジャーナリングファイルシステムでは，書き込み操作を実行する前に，ジャーナルと呼ばれるログ領域にまず操作内容を記録する．この記録が終了したのち，本来のディスクへの書き込み操作を実行する．これにより，正常にシステムがシャットダウンされなかった場合にも，システム再起動時にジャーナルに基づいてファイルシステムを正しい状態に復旧させることができる．しかし，ディスク書き込みに I/O のスケジューリングを行うエレベータアルゴリズムが適用されているなどの理由でファイルシステムの要求と異なる順序でディスクへの書き出しが行われていた場合には，電源障害時にジャーナルファイルが破損することがある．その場合には破損したジャーナルログファイルを元にしたジャーナリングが行われる可能性があり，ジャーナリングファイルシステムにおいても正しい復旧は保証されない．

ログ構造化ファイルシステムではファイルの管理情報

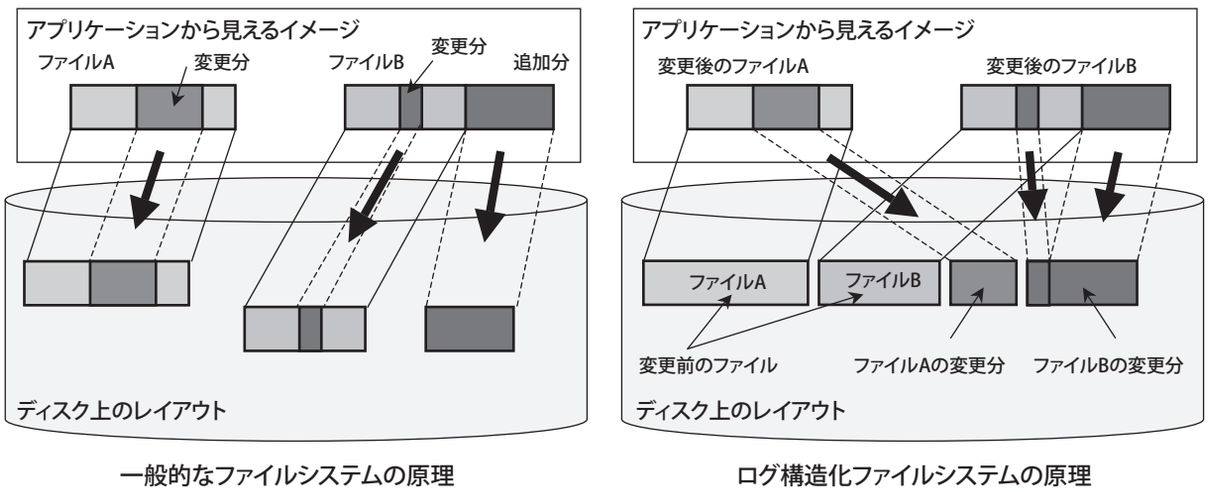

図 -2　ログ構造化ファイルシステム

やデータを上書きするのではなく，それらの変更部分のみをディスクの空き領域に書き込む（**図 -2**）．したがってすべての書き込みはそれまでに書かれた領域を変更しないので，電源障害などの場合にも存在するほかのブロックを破壊することがなく，高い可用性を達成できる．また，すべての書き込みは連続した領域への書き込みであるので，高速な書き込みが可能となる．さらに，一連の書き込みが完了した時点で，時刻情報付きのスナップショットを自動生成し，スナップショットへの瞬時リカバリが可能という特徴を持つ．スナップショットはオペレーションミス等も救済できるのでシステムの保守性の向上にも役立つと考えられている．

これら新規機能の導入により，Linux はミッションクリティカル向けとしてもトップレベルの機能を持った OS になっていくと予想される．

ミッションクリティカル分野適用に向けた各種団体の活動

Linux のミッションクリティカル分野への適用に向けて，各種団体の活動も活発に行われている．本章ではミッションクリティカル向けの活動をしている唯一の国際団体 OSDL，国内で活動を行っている団体である日本 OSS 推進フォーラムの活動内容について述べる．

OSDL の活動

オープン・ソース・デベロップメント・ラボ（Open Source Development Labs Inc.）は Linux の成長とエンタープライズでの Linux 採用を促進することを目的とした非営利団体（NPO）である[2]．2000 年に設立され，現在 Linus Torvalds およびメンテナの Andrew Morton が在籍している．OSDL には現在 CGL（Carrier Grade Linux），DCL（Data Center Linux），DTL（DeskTop Linux），MLI（Mobile Linux Initiative）という 4 つのワーキンググループにおいて技術面での活動を行っている．そのうちミッションクリティカル分野をターゲットにしている DCL ワーキンググループにおける活動について以下では述べる．

DCL ワーキンググループにおいては市場および現状の技術レベルを調査し，ミッションクリティカルシステムにおける Linux 適用に求められる要件の文書 DCL Goals and Capabilities を作成している．最新版のバージョン 1.1 は 2005 年 2 月にリリースされた．その技術的要件 Technical Capabilities は 8 つの分野に分類されている．

- スケーラビリティ：大規模システムを Linux で構築できるための要件．
- 効率（performance）：Java の実行効率など，ビジネス向けの効率要件．
- 信頼性・可用性・保守性（reliability/availability/serviceability）：365 日 24 時間の利用を可能とするための要件．
- 管理性（manageability）：システムモニタリングやログ収集解析など，管理のためのツール類に関する要件．
- クラスタリング：負荷分散や可用性のためのクラスタリングに関する要件．
- 標準化準拠：ディストリビューション間移行を容易と

するLinux Standard Baseなどの他の標準化への準拠.
- セキュリティ：顧客データ流出や不正侵入などを防止するためのセキュリティ要件.
- 使いやすさ（usability）：管理者のシステム構築・管理を容易にするツール類に関する要件.

これらの技術的要件と現在のLinuxの機能とのギャップを調査し，ギャップを埋める技術的活動を行っている．具体的な各技術項目に関しては誰でも参加可能なSIG（Special Interest Group）を構成して活動を行っている．現在，DCL関係で活動中のSIGはClustering, Security, Storage, HotPlugの4つであり，メモリ／CPUのホットプラグ機能やNFSv4のテストなどが行われている．

このほか，OSDLではLinuxの信頼性向上を目指す活動も行われている．プロジェクトDOUBTでは，Linuxで提供されている機能が正しく動作するか否かのテストを行う各種プログラムを提案し，調査結果を公開している[3]．たとえば，ファイルシステムが同期書き込みを正しくサポートしているか否かをテストするためのツールとしてdiskioがオープンソースソフトウェアとして公開されている．diskioを使用すると同期書き込みのシステムコールを出してから終了するまでの時間を測定することが可能である．もしその時間がディスクの書き込みにかかるはずの時間よりも短ければ，正しい処理を行っていないと結論づけることができる．diskioを用いることにより，調査時点で最新の，バージョン2.6.8などのext3で正しく同期書き込みが行われていなかったことが初めて明らかになった☆1．

ほかにもストレージの取り外し後にファイルにアクセスした時にエラーメッセージを正しく返すか否かというテスト結果も公開されている．Linuxの多くのファイルシステムで正しくエラーメッセージが出力されないことが明らかになっている．これはLinuxカーネルの信頼性向上を目指す非常に数少ない活動の1つである．この分野のさらなる活性化が望まれる．

日本OSS推進フォーラムの活動

日本OSS推進フォーラムは，日本の情報システムのユーザ，ベンダ，学識経験者の有識者が参集して，オープンソースソフトウェアの活用上の課題解決に向けての取り組みを行うために2004年に組織された[4]．また，日本・中国・韓国が連携した北東アジアOSS推進フォーラムの日本側の構成員でもある．事務局を独立行政法人情報処理推進機構が務めている．ワーキンググループとしてデスクトップWG，開発基盤WG，ビジネス推進WG，サポートインフラWG，人材育成WG，標準化・認証WGがあるが，本稿ではミッションクリティカル分野の技術部門に関連の深い開発基盤WGの活動について紹介する．

開発基盤WGにおいては，オープンソースソフトウェアの性能・信頼性評価によるシステム設計・構築ノウハウの共有，および障害解析ツールの開発とノウハウの共有を行うことを目的としている．今までの開発成果としてはダンプデータ解析ツールAlicia，ディスク割り当てツールDAVなどがあげられる．どちらもオープンソースソフトウェアとして公開されている．

Linuxのダンプデータ解析ツールとしてはすでにcrashおよびlcrashが存在するが，解析のスクリプトを記述する際にlcrashはsialという言語で記述する必要がある．またcrashにおいては拡張コマンドとしてC言語で記述する動的ライブラリを用いる方法しかない．そこで，crashおよびlcrash（対応予定）をPerlのインタフェースによりラッピングして解析スクリプト記述を容易にしたツールがAliciaである．これにより，コマンドの実行結果を変数に格納してPerlで処理することなどが可能になり，ダンプデータ解析の効率向上に役立つと期待される．

DAVはディスクの割り当て状況を可視化するツールである．ディスクのフラグメンテーションによる性能劣化を調査するために使用できる．ディスク全体の表示のほか，指定したファイルのみの状況表示が可能であるため，特定のファイルのフラグメンテーションに起因する性能劣化の検出も容易である．また，デフラグツールの性能評価にも利用することができる．現在，対応しているファイルシステムはext2/ext3のみであり，他のファイルシステムへの対応が望まれている．

Linux導入に向けたユーザ等の動向

Linuxカーネルの成長に伴い，ベンダのLinux対応やユーザの導入も加速しつつある．経済産業省が総務省などと共同で策定している政府IT調達指針もほぼ完成しており，間もなく公開される予定である．

ガートナージャパンなどが運営する新電子自治体共同研究会が2004年9～10月に行った調査によると，日本の自治体の53%がLinuxサーバを利用しており，サーバ台数比率でも11.4%に達している[5]．また，矢野経済研究所とインプレスが2005年6～7月に行った調査では，企業や公共団体においてLinuxサーバを利用している率は38%と，UNIXサーバの26.4%を上回っている[6]．このように，企業・自治体等でのLinux導入は確実に増加し

☆1 その後のバージョンでこのバグはフィックスされている．

つつある．金融機関等，特にミッションクリティカル性が要求されるユーザにおいてもLinuxでのシステム構築の実例が増えてきている．

しかし企業ユーザのLinux利用にあたっての問題点がいくつか存在する．以下では主な問題点とそれに対する取り組みについて述べる．

Linuxディストリビューションの間に互換性がないと，アプリケーション・ベンダはディストリビューションごとの動作試験や異なるバージョンの開発を必要とするためにアプリケーション開発上の障害となる．またユーザも，異なるディストリビューション間でのシステム移行が困難になる．ディストリビューション間互換性の確保のため，米国の非営利団体Free Standards GroupではLinuxの標準規格Linux Standard Base（LSB）を制定している[7]．LSBはABI（バイナリレベルの標準インタフェース），標準コマンド，ファイル配置，設定ファイル，共有ライブラリなどの規定からなり，2005年10月にバージョン3.1が公開されている．

Linux導入についてのユーザの不安の大きな要因はサポートである．Linuxと他のオープンソースソフトウェアを組み合わせて利用して初めて不具合が発生する可能性もある．それぞれのソフトウェアごとに別のベンダにサポートを依頼すると，このような場合に障害の原因がどこにあるのかの切り分けをユーザができないとサポートを受けられない．このようなソフトウェアの組合せに対する動作検証サービスを提供する企業や，顧客のLinuxシステムの障害について，カーネルのソースコードやメモリダンプに基づいた詳細調査を行い，障害原因を追及して特定，回避方法を提案し，修正パッチを作成するサービスを提供する企業も存在する．このように，サポート面からも企業での利用が容易になりつつある．

企業においてLinuxの導入をためらわせるもう1つの理由に知的財産にまつわる問題があげられる．UNIXのコード流用問題や特許侵害問題である．

SCO社は，自社が著作権を保有するUNIXのコードをLinuxに流用したとして，IBMを相手取った訴訟を行っている．米連邦裁判所はこの裁判の陪審による審理を2007年2月としているのでそれまでこの訴訟が続く可能性がある．OSDLはこの問題に関し，Linux法的防御基金を設立してLinuxユーザがSCOからの訴訟に巻き込まれた場合に法的費用を負担することを宣言している．また，SCOが今までに一般に公開した情報の中には，流用であると認められるコードは存在していない．多くの識者はSCOが勝訴する可能性はないと述べているが訴訟の早い終結が待たれている．

もう1つの問題は，Linuxカーネルが他社の持つ特許を侵害している可能性である．知的財産権侵害訴訟から守るための保険を販売するOpen Source Risk Management（OSRM）の2004年の調査によると，Linuxカーネルには283件の特許侵害の可能性があると言われている[8]．しかし，このうちの60件の特許を有するIBMはLinuxカーネルに対して特許権を主張しないことを宣言している．このような動きを背景に，OSDLはオープンソースソフトウェアに対して無償利用が許諾された特許のデータベースPatent Commons Projectを公開している．また，2005年にはLinuxやその上のアプリケーションに関連する特許を買い取り，それらを無償で利用許諾することを目的としたOpen Invention Networkという会社が設立された[9]．さらに，複数の大手ディストリビュータは顧客に対し，権利侵害訴訟を起こされた場合の免責保証サービスを提供し始めた．したがって，Linuxのユーザ企業が訴えられるリスクは現在では非常に小さいと言える．

あとがき

本稿ではミッションクリティカルシステム向けのLinuxに関する動向についてまとめた．日本のユーザは欧米のユーザと比べてミッションクリティカルシステムに対する要求条件が厳しいと言われている．したがって，可用性・信頼性等の分野では日本のベンダ等が中心になって活動を行う必要があり，活動実績も増えてきている．この分野における日本発の技術がさらに増えることを期待したい．このような貢献を通じて，Linuxはミッションクリティカル向けOSとしていっそう利用されるようになるであろう．

参考文献
1）NILFSプロジェクト：http://www.nilfs.org/
2）OSDL：http://www.osdl.org/
3）DOUBTプロジェクト：http://developer.osdl.jp/projects/doubt/
4）日本OSS推進フォーラム：http://www.ipa.go.jp/software/open/forum/
5）ガートナージャパンプレスリリース：http://www.gartner.co.jp/press/pr20050106-01.pdf
6）矢野経済研究所プレスリリース：http://www.yano.co.jp/press/2005/050928.html
7）Free Standards Group：http://www.freestandards.org/
8）OSRMプレスリリース：http://www.osriskmanagement.com/press_releases/press_release_080204.pdf
9）Open Invention Network：http://www.openinventionnetwork.com/
（平成17年12月1日受付）