

SGI ProPack™ for Linux® Start Here (日本語版)

007-4558-004JP

制作協力

著作 : Julie Boney

編集 : Susan Wilkening

イラスト : Chrystie Danzer

制作 : Glen Traefald

著作権

© 2003 Silicon Graphics, Inc. All rights reserved. このマニュアルには下記の商標・著作に関する注意にある通り、サード・パーティが著作権を保有している部分が含まれます。本書の内容の一部あるいは全部について、Silicon Graphics, Inc. から事前に文書による明確な許諾を得ず、いかなる形態においても複写、複製、翻案することは禁じられております。

LIMITED RIGHTS LEGEND

The electronic (software) version of this document was developed at private expense; if acquired under an agreement with the USA government or any contractor thereto, it is acquired as "commercial computer software" subject to the provisions of its applicable license agreement, as specified in (a) 48 CFR 12.212 of the FAR; or, if acquired for Department of Defense units, (b) 48 CFR 227-7202 of the DoD FAR Supplement; or sections succeeding thereto. Contractor/manufacturer is Silicon Graphics, Inc., 1600 Amphitheatre Pkwy 2E, Mountain View, CA 94043-1351 USA.

商標・著作

Silicon Graphics, SGI, SGI ロゴ, IRIX, Origin, Onyx, Onyx2, XFS は Silicon Graphics, Inc. の米国および各国における登録商標です。NUMAflex, NUMAlink, OpenMP, Performance Co-Pilot, SGI Advanced Linux, SGI ProPack, SGIconsole は Silicon Graphics, Inc. の米国および各国における商標です。

SGI Advanced Linux Environment 2.1 は Red Hat Linux Advanced Server 2.1 for the Itanium Processor をベースにしていますが、Red Hat, Inc. のサポート対象には含まれません。

Cray は Cray, Inc. の登録商標です。FLEXlm は Macrovision Corporation の登録商標です。Java は Sun Microsystems, Inc. の米国および各国における登録商標です。KAP/Pro Toolset, VTune は Intel Corporation またはその米国および諸国における子会社の商標です。Intel, Itanium, Pentium は Intel Corporation またはその米国および諸国における子会社の登録商標です。Linux は Linus Torvalds の登録商標であり、Silicon Graphics, Inc. は許可を得て使用しています。MIPS は MIPS Technology, Inc. の登録商標です。PostScript は Adobe Systems, Inc. の登録商標です。QLogic は QLogic Corporation の登録商標です。Red Hat は Red Hat, Inc. の登録商標です。Red Hat Linux Advanced Server 2.1 および RPM は Red Hat, Inc. の商標です。Windows は Microsoft Corporation の米国および各国における登録商標です。

表紙デザイン : Sarah Bolles (Sarah Bolles Design), Dany Galgani (SGI Technical Publications)

改訂履歴

バージョン	説明
001	2003 年 2 月 初版発行
002	2003 年 2 月 SGI ProPack v2.1.1 for Linux リリース改版 (含邦訳)
003	2003 年 4 月 SGI ProPack v2.1.2 for Linux リリース改版
004	2003 年 6 月 SGI ProPack v2.2 for Linux リリース改版 (含邦訳)

本リリースでの変更点

本書『SGI ProPack for Linux Start Here』の今回の改訂は SGI ProPack for Linux 2.2 リリースに対応します。

本書の主な変更

本書の今回の改訂では次の部分に変更・追加されています。

- 基本になる Linux ディストリビューションが Linux Red Hat Advanced Server 2.1 に変更されています。これを基本にした新しい SGI ベース・ディストリビューションが SGI Advanced Linux Environment 2.1 です。
- カーネルは `/boot/efi/EFI/sgi/` にインストールされます。この変更は `elilo v3.3` に対応するものです。詳細は 第3章の「拡張ファームウェア・インターフェイス (EFI: Extensible Firmware Interface)」を参照してください。
- アップデート用 CD およびブート用 CD が供給 CD の一覧から削除されました。これらの機能はインストール CD に含まれます。
- 第2章の「SGI ProPack 旧リリースから SGI ProPack 2.2 へのアップグレード」が新たに追加されました。

目次

本書の主な変更	v
図一覧	xi
表一覧	xiii
関連ドキュメント	xvi
Intel コンパイラ関連文書	xvi
その他の Intel 関連文書	xvi
SGI 文書	xvi
SGI ドキュメントの入手先	xviii
本書の表記法	xviii
読者からのフィードバック	xix
1. リリースの内容	1
ソフトウェアの概要	1
CD の内容	2
インストールの概要	4
2. ソフトウェアの計画とインストール	5
ソフトウェアの計画	5
ディスク・パーティション表	6
Linux カーネル・クラッシュ・ダンプ	6
SGI Advanced Linux Environment 2.1 ベース OS のインストール	7
SGI ProPack 旧リリースから SGI ProPack 2.2 へのアップグレード	10
SGI ProPack for Linux のインストール	11
System Controller Software 1.1 CD のインストール	13
ソフトウェアのアップグレード	13

損傷した Root ファイルシステムの修復	14
Root ファイルシステムのマニュアル・マウント	16
3. サポートされる製品	17
OS 拡張	21
CpuMemSets のサポート	21
Cpuset のサポート	22
完全システム・アカウンティング (CSA:Comprehensive System Accounting)	22
システム・パーティション	23
I/O サブシステム	23
PCI-X バス固定ナンバリング	24
Ethernet デバイスの固定命名	25
XSCSI サブシステム	27
XSCSI-SCSI サブシステム	27
XFS ファイルシステム	28
XVM ボリューム・マネージャ	28
HPC アプリケーションのツールとサポート	28
Message Passing Toolkit	28
Performance Co-Pilot	29
システム管理	29
PROM チップ	29
拡張ファームウェア・インターフェイス (EFI: Extensible Firmware Interface)	30
FLEXlm	32
SGIconsole	33
システム・コントローラ・ファームウェア	33
NUMA Tools	33
dlook コマンド	34
dplace コマンド	34
4. パフォーマンス・チューニング	35
システム構成の確認	35

シングルスプロセッサ・コードのチューニング	36
出力の正確性の確認	37
dddの使用	38
ヒープ破壊問題対策	39
チューニング済みコードの使用	41
チューニング必要部分の決定	41
コンパイラ・オプションの選択と適用	42
キャッシュ・パフォーマンスのチューニング	43
メモリ管理	43
chatrによるスタックとデータ・セグメントの変更	44
マルチプロセッサ・コードのチューニング	45
コードの並列化	45
コード中のボトルネックの検出	47
無効な共有の修復	47
dplaceとrunonの利用	48
MPIコードに対するdplaceの使用	48
OpenMPコードに対するdplaceの使用	48
環境変数によるパフォーマンス・チューニング	49
プロファイラとパフォーマンス・ツール	50
pfmonを使ったプロファイリング	50
profile.plスクリプト	51
MPIプログラムでのprofile.plの使用	51
VTuneを利用したリモート・サンプリング	52
GuideViewの利用	52
その他のパフォーマンス・ツール	53
索引	55

図一覽

図 1-1	SGI ProPack for Linux Release CD の内容	3
-------	--	---

表一覧

表 2-1	devfs ディスク・パーティション	6
表 3-1	SGI ProPack v2.2 for Linux 製品	18
表 3-2	EFI コマンド	30

このマニュアルについて

このマニュアルは、SGI ProPack for Linux に関する情報を解説します。本書は次の各章から構成されます。

- 第1章「リリースの内容」では、このリリースの主な機能、CD の内容、および関連ドキュメントについて説明します。
- 第2章「ソフトウェアの計画とインストール」では、ディスクのパーティション設定、システムの利用を始める際に必要となる設定ファイルの作成と配置、SGI ProPack for Linux のインストールの手順についてフローチャート形式で説明します。
- 第3章「サポートされる製品」では、SGI Altix 3000 システムでサポートされる製品コンポーネントを示します。
- 第4章「パフォーマンス・チューニング」では、シングルプロセッサ環境およびマルチプロセッサ環境でのプログラムのチューニングについて解説します。

本書の内容とその他の SGI ProPack for Linux ドキュメント、ディストリビューション CD に含まれる RPM に関するその他のドキュメントは CD "SGI ProPack v2.2 for Linux - Documentation CD" に入っています。このドキュメント CD の内容を見るには、Web ブラウザ・プログラムを使って `index.html` ファイルを開いてください。このオンライン・ファイルには本書作成後に記述された追加情報が含まれていることがありますので、目を通すようにしてください。

メモ：ソフトウェアおよびドキュメントに関するリリース時の最新情報は SGI ProPack for Linux Documentation CD のルート・ディレクトリにある `README.TXT` ファイルに格納されています。

関連ドキュメント

有用な情報が記載されたドキュメントと参考文献の一覧を次に示します。

Intel コンパイラ関連文書

Intel コンパイラに関する資料はシステム・ディスクのコンパイラがインストールされているディレクトリの `/docs` サブディレクトリにあります。Intel コンパイラのインストール後、下記の資料が参照可能です。

- 『Intel C++ Compiler User's Guide』 (*c_ug_lnx.pdf*)
- 『Intel Fortran Compiler User's Guide』 (*for_ug_lnx.pdf*)
- 『Intel Fortran Programmer's Reference』 (*for_prg.pdf*)
- 『Intel Fortran Libraries Reference』 (*for_lib.pdf*)

その他の Intel 関連文書

次の参考文献は Itanium (旧名称 IA-64) アーキテクチャやその関連事項について解説しています。

- 『Intel Itanium 2 Processor Reference Manual for Software Development and Optimization』
(<http://developer.intel.com/design/itanium/manuals> からオンライン入手可)
- 『Intel Itanium Architecture Software Developer's Manual』
(<http://developer.intel.com/design/itanium/manuals> からオンライン入手可)
- 『Introduction to Itanium Architecture』
(<http://shale.intel.com/softwarecollege/CourseDetails.asp?courseID=13> からオンライン入手可、要セキュア接続)

SGI 文書

SGI からは次の文書が入手可能です。

- 『Linux Device Driver Programmer's Guide』

デバイス・ドライバのプログラミング、構成、制御についての情報が記載されています。

- 『Message Passing Toolkit: MPI Programmer's Manual』
SGI システム用に最適化された業界標準メッセージング・プロトコルについての解説です。
- 『Origin 2000 and Onyx2 Performance Tuning and Optimization Guide』
MIPS/IRIX システムでのパフォーマンス・チューニングと最適化についての記述ですが、ハードウェアや OS に依存しない一般的なガイドラインについても述べられています。
- 『Performance Co-Pilot for Linux User's and Administrator's Guide』
Performance Co-Pilot (PCP) ソフトウェア・パッケージの解説です。PCP は SGI システム用に開発されたパフォーマンス解析ツールで、Linux システム上でも動作します。
- 『Resource Administration Guide for Linux』
Linux OS 搭載 SGI システム管理者向けのリファレンス・ガイドです。
- 『SGI Altix 3000 User's Guide』
SGI Altix 3000 システムのアーキテクチャ概要と主なコンポーネントについての解説です。システムの電源投入とシャットダウンの基本操作、トラブルシューティング、設置と運用に関する安全基準などの情報も含まれます。
- 『SGI ProPack v2.2 for Linux Release Notes』
ソフトウェアおよびドキュメントに関するリリース時の最新情報が納められています。この文書は SGI ProPack for Linux Documentation CD のルート・ディレクトリにある README.TXT ファイルに格納されています。
- 『SGIConsole 1.1 Start Here』
SGIConsole の概要が解説されています。
- 『SGI L1 and L2 Controller Software User's Guide』
システム・コンソールから実行する L1/L2 コントローラ・コマンドの解説書です。これらのコマンドは SGI Altix 3000 Series システムのモニタリングと管理に使用します。
- 『XFS for Linux Administration』
XFS についての解説です。XFS はオープン・ソースのジャーナル・ファイルシステムで、高速回復性、ダイレクト I/O、スペース・プリアラケーション、アクセス・コントロール・リスト、quotas など、業務用ファイルシステムに必要な機能を備えています。

SGI ドキュメントの入手先

SGI が公開するドキュメントは下記の方法で入手できます。

- SGI Technical Publications Library (<http://docs.sgi.com>) にアクセスする方法。ドキュメントは各種のフォーマットで提供されます。オンライン・ブック、リリース・ノート、マン・ページ、その他の文献の最新版が入手できます。
- マン・ページは、コマンドラインから「`man <title>`」と入力すると表示されます。

本書の表記法

本書では、次の表記が使用されます。

書式	内容
<code>command</code>	コマンド名、ファイル名、ルーチン名、パス名、シグナル、メッセージ、プログラミング・コードは等幅フォントで表示されます。
<i>variable</i>	変数、新出の用語は斜体で表示されます。
user input	用例中、ユーザが入力する文字は太字の等幅フォントで表示されます。
<code>manpage(x)</code>	マン・ページのタイトルは、かっこ内にセクション番号が表記されます。
[GUI 要素]	ウィンドウ、画面、ダイアログ・ボックス、メニュー、ツールバー、アイコン名、ボタン、ボックス、フィールド、リストなどの GUI (Graphics User Interface) 要素名は、角かっこ <code>[]</code> で囲んで表示されます。

読者からのフィードバック

この文書の技術的な正確性、内容、または編成についてのご意見がありましたら、SGI までご連絡ください。ご意見とともに、マニュアルのタイトルと文書番号をお知らせください。(オンライン文書の文書番号は、文書の 1 ページ目に記載されています。印刷文書の文書番号は、本文ページ下端に記載されています。)

弊社には次のいずれかの方法でご連絡いただけます。

- 電子メールで連絡される場合は、下記アドレスをご利用ください。
`techpubs@sgi.com`
- Technical Publications Library Web ページ(下記 URL)にアクセスされる場合は、Feedback 欄からコメントを送信してください。
`http://docs.sgi.com`
- ご購入元のカスタマー・サービス担当にご連絡される場合は、問題点が SGI のバグ・トラッキング・システムにファイルされるようご相談ください。
- 書面でのお問い合わせは、下記住所まで郵送ください。

Technical Publications
SGI
1600 Amphitheatre Pkwy, M/S 535
Mountain View, California 94043-1351 USA

- FAX で送信される場合は、+1 650 932 0801 “Technical Publications” 宛までお送りください。

お送りいただいたご意見には迅速に対応いたします。

リリースの内容

この章では SGI ProPack for Linux の概要と CD 内容の構成、ドキュメントに関する情報、およびインストール手順の概要について説明します。

ソフトウェアの概要

Linux はオープン・ソースと Linux コミュニティを特徴とする共同体制により開発されており、大規模サーバやスーパーコンピュータを提供するコンピュータ・メーカーに対して新しい開発モデルの方向性を示しています。大規模コンピュータ・システムに Linux OS を採用する利点は、より高度なソフトウェア保護、そしてメーカー製のオペレーティング・システムやハイ・パフォーマンス・コンピューティング (HPC)・アプリケーション / コードと、サード・パーティやユーザによるソフトウェア開発との間で、より優れた統合化を実現できることです。また、Linux コミュニティによる OS の進化と企業メーカーによる開発が連動することで、企業単独によるオペレーティング・システム開発とは異なる方向から、革新と進歩の新たな可能性がもたらされます。

SGI ProPack for Linux 製品は、技術や芸術分野で要求される大規模なデータ処理と計算量のために Linux OS と Itanium プロセッサが適用される局面において理想的な処理機能とパフォーマンスを備えています。SGI ProPack for Linux は Red Hat Linux Advanced Server 2.1 をベース OS として、各種の機能が追加、拡張されています。SGI ProPack for Linux は SGI Altix 3000 Series の任意の構成で動作するよう設計されています。このリリースでサポートされる SGI ハードウェア・プラットフォームおよび OS 構成の範囲は、下記 URL から入手できるドキュメントに詳しく説明されています。

<http://support.sgi.com/linux>

CD の内容

SGI ProPack for Linuxには、次のCDが含まれています(各CDの内容は図1-1に示されています)。

- SGI Advanced Linux Environment 2.1
 - SGI Advanced Linux Environment 2.1 CD1 - Installation CD 1 of 2
 - SGI Advanced Linux Environment 2.1 CD2 - Installation CD 2 of 2
 - SGI Advanced Linux Environment 2.1 CD3 - Source Code CD 1 of 2
 - SGI Advanced Linux Environment 2.1 CD4 - Source Code CD 2 of 2
- SGI ProPack v2.2 for Linux
 - SGI ProPack v2.2 for Linux CD1 - Open/Free Source Software
 - SGI ProPack v2.2 for Linux CD2 - Proprietary Software

メモ：この CD には SGI システム専用ソフトウェアが含まれています。再配布は認められません。パッケージに印刷されているライセンス条項を参照してください。

- SGI ProPack v2.2 for Linux CD3 - Documentation CD
- System Controller Software 1.1 CD (for IRIX and Linux)

メモ：このソフトウェアは Origin / Onyx 3000 Series システム (MIPS プロセッサ・ベースの SGI システム)、および SGI Altix 3000 Series システムに対してのみライセンスされます。この CD は SGI ライセンス条項と使用制限に従って使用する必要のあるソフトウェアを含みます。この CD の SGI ライセンス条項では、CD 内容が SGI システムにのみインストール可能であるということが規定されています。

SGI Advanced Linux Environment 2.1 および SGI ProPack v2.2 for Linux CD の内容の一覧は 3 ページの図 1-1 を参照してください。ソフトウェアとドキュメントに関するリリース時の最新情報は SGI ProPack for Linux Documentation CD のルート・ディレクトリにある README.TXT ファイルに格納されています。

* SGI Advanced Linux Environment 2.1		
	SGI Advanced Linux Environment 2.1 Installation CD 1 of 2 CD1	インストーラ本体
	SGI Advanced Linux Environment 2.1 Installation CD 2 of 2 CD2	
	SGI Advanced Linux Environment 2.1 Source Code CD 1 of 2 CD3	ベース OS とその他のオープンソース・アプリケーションのソースコード
	SGI Advanced Linux Environment 2.1 Source Code CD 2 of 2 CD4	
* SGI ProPack v2.2 for Linux		
	SGI ProPack v2.2 for Linux Open/Free Source Software CD1	GPL/LGPL ライセンスおよびその他のオープンソース・ライセンス・ソフトウェア、 最新 SGI プラットフォームと NUMA サポート、カーネルのパフォーマンス改善、 CpuMemSets とパフォーマンス評価ツール、XSCSI インフラストラクチャ、 QL-SCSI、XFS、CSA、LKCD、kdb
	SGI ProPack v2.2 for Linux Proprietary Software CD2	システム PROM、XVM、cpuset、NUMAtools (dlock、dplace など)、 Array Services、FLEXIm、システム・パーティション・ソフトウェア、MPT
	SGI ProPack v2.2 for Linux Documentation CD CD3	ドキュメント・セット、FAQ、HOWTO、マン・ページ
* System Controller Software 1.1 CD		
	System Controller Software 1.1 CD	システム・コントローラ・ソフトウェアおよびファームウェア (IRIX/Linux 用)

図 1-1 SGI ProPack for Linux Release CD の内容

SGI のオープン・ソースに関連する開発計画の詳細は、次の Web サイトで公開されています。

<http://oss.sgi.com>

その他のオープン・ソース（LKCD など）情報は、次の Web サイトで公開されています。

<http://sourceforge.net>

インストールの概要

通常、SGI ProPack for Linux は SGI プラットフォームにプリインストールされて出荷されます。再インストールが必要な場合は、まず SGI Advanced Linux Environment 2.1 CD セットをインストールして、次に SGI ProPack v2.2 Open/Free Source Software (CD1)、そして最後に SGI ProPack v2.2 Proprietary Software (CD2) の順でインストールしてください。インストールの詳細は第 2 章「ソフトウェアの計画とインストール」で説明されています。

メモ： ProPack v2.2 リリースにはアップデート CD はありません。アップデートのためのデータはすべて SGI Advanced Linux Environment 2.1 CD セットに含まれています。

ソフトウェアの計画とインストール

この章では SGI ProPack for Linux のインストールに必要な計画と手順について説明します。

SGI Altix 3000 システムには、ベース Linux ディストリビューション (SGI Advanced Linux Environment 2.1) と SGI ProPack ソフトウェアがプリインストールされています。この章では、これらのソフトウェアを再インストールする必要があるときに CD からインストールする手順について説明します。

セキュリティ強化のため、Linux で root ログインするにはパスワードが必要です。プリインストールされているソフトウェアの場合、デフォルトのパスワードは `sgisgi` です。ログイン後、このパスワードは別のパスワードに置換えてください。

SGI ProPack ソフトウェアは SGI Advanced Linux Environment 2.1 ディストリビューションと共にインストールした場合のみ動作します。これより古いディストリビューションや他社のディストリビューションは、SGI ProPack ソフトウェアと共に動作しません。

システムのインストールと設定を始める前に、第1章「リリースの内容」を読み、SGI ProPack for Linux ソフトウェアの機能と構成の概要を理解してください。xvi ページの「関連ドキュメント」に示される書籍も参考になります。

ソフトウェアの計画

ここでは、同梱されるソフトウェアについて知っておく必要のあるディスクとファイルの情報を説明します。

ディスク・パーティション表

SGI から出荷されるシステムは、表 2-1 に示されるレイアウトでパーティションが設定されています。

表 2-1 devfs ディスク・パーティション

デバイス	マウント・ポイント	サイズ
<code>/dev/xscsi/pci01.03.0-1/target1/lun0/part1</code>	<code>/boot/efi</code>	500 MB
<code>/dev/xscsi/pci01.03.0-1/target1/lun0/part2</code>	swap	9 GB
<code>/dev/xscsi/pci01.03.0-1/target1/lun0/part3</code>	/	25 GB

デバイスの名前については 24 ページの「PCI-X バス固定ナンバリング」と 27 ページの「XSCSI サブシステム」も参照してください。

Linux カーネル・クラッシュ・ダンプ

LKCD (Linux Kernel Crash Dump) は `/var/dump` にファイルを作成します。`/var` ディレクトリのサイズは Performance Co-Pilot (PCP) と CSA (Comprehensive System Accounting) がそれぞれのサブディレクトリにログを蓄積するにつれて増えます。通常、これらのファイルは 100 MB を越えません。LKCD のデフォルト設定の詳細は `lkcd_config` (1) マン・ページを参照してください。

メモ： `lkcd` ダンプが生成されると、LKCD は `/var` ディレクトリを容易に数 GB 以上消費してしまいます。デフォルトの `root` ファイルシステムが比較的大きく設定されているのはこのためです。必要に応じて `/var/dump` を独立したファイルシステムにしたり、`lkcd` の設定を変えることができます。

SGI Advanced Linux Environment 2.1 ベース OS のインストール

ここでは、SGI Advanced Linux Environment 2.1 CD セット (CD1 と CD2) のインストール手順について説明します。コンソール設定のオプションとその手順については、『SGI Altix 3000 User Guide』の「Monitoring Your Server」を参照してください。

メモ： インストール画面には、前のステップに戻るためのボタンとインストールを中断するためのボタンがあります。これらのボタンを使用するには、Tab キーを押して選択したいボタンをハイライト表示させ、Enter キーを押します。

1. SGI Advanced Linux Environment 2.1 CD1 をシステムの CD-ROM ドライブにセットし、システムを再起動します。
2. システム起動メッセージの表示中に、デバイス・マッピングを確認します。次のようなメッセージ出力をチェックしてください。

Device mapping table

```
fs0  : Pci(2|1)/Ata(Primary, Master)/CDROM(Entry1) (この行が CD-ROM)
fs1  : Pci(1|1)/Scsi(Pun0/Lun1)/HD(Part1,Sig00000000)
fs2  : Pci(1|1)/Scsi(Pun0/Lun2)/HD(Part1,Sigg1)
blk0 : Pci(2|1)/Ata(Primary, Master)
blk1 : Pci(2|1)/Ata(Primary, Master)/CDROM(Entry1)
blk2 : Pci(1|1)/Scsi(Pun0/Lun1)
blk3 : Pci(1|1)/Scsi(Pun0/Lun1)/HD(Part1,Sig00000000)
blk4 : Pci(1|1)/Scsi(Pun0/Lun1)/HD(Part2,Sig00000000)
blk5 : Pci(1|1)/Scsi(Pun0/Lun1)/HD(Part3,Sig00000000)
blk6 : Pci(1|1)/Scsi(Pun0/Lun1)/HD(Part4,Sig00000000)
blk7 : Pci(1|1)/Scsi(Pun0/Lun2)
blk8 : Pci(1|1)/Scsi(Pun0/Lun2)/HD(Part1,Sigg10)
```

3. **Shell>** プロンプトが表示されたら、次のように CD-ROM デバイス名を指定します。

```
Shell> fs0: (CD-ROM デバイス名を指定して移動)
fs0:\>
```

4. 「**elilo**」と入力して Enter キーを押し、CD からブートします。ELILO boot: プロンプトが表示されたら、再度 Enter キーを押します。

```
fs0:\> elilo (ブートローダを指定)
LoadPe: using PE image entry point
LoadPe: using PE image entry point
ELILO boot: (Enter キーを押して Linux カーネルを起動)
```

5. CD がブートされ、インストールが始まります。

6. インストールで使用する言語と、システムのデフォルトとして設定する言語を選びます。**OK** を選択して次に進みます。

メモ：現時点では、コンソールでのアジア言語はサポートされません。SGI では、アジア系言語で使用するシステムのデフォルト・システム言語として英語 (English) を選択することを推奨します。

7. 次の画面では、インストール手法を選択できます。現時点で有効な選択肢は [Local CDROM] のみです。**OK** を選択して次に進みます。
8. SGI ProPack 情報画面が表示されます。**OK** を選択して次に進みます。
9. 次の画面ではマウスの有無を選択します。SGI Altix 3000 システムでは [None] を選択します。**OK** を選択して次に進みます。
10. インストール・タイプ選択画面が表示されます。アップグレードを行う場合は、[Upgrade Existing Sysem] を選択して 10 ページの「SGI ProPack 旧リリースから SGI ProPack 2.2 へのアップグレード」に進んでください。ディスク上の既存のデータを消去してシステムを新規にインストールする場合は、[Custom] を選択して次のステップに進んでください。
11. パーティション画面が表示されたら、[Autopartition] を選択して、**OK** を選択します。警告メッセージが表示されたら [IGNORE] を選択します。
12. パーティション方法を尋ねられるので、[Remove all partitions on this system] を選択します。SGI Altix 3000 システムに複数の SCSI ディスクが含まれる場合は、どのドライブにインストールするかを尋ねられます。SGI Altix 3000 出荷時の構成では、1 つのシステム・ディスクに OS 全体がインストールされています。出荷時同様に 1 つのシステム・ディスクに OS 全体をインストールしたい場合は、ここでインストール対象のディスクとして 1 つだけ選択します。デフォルトでは、複数のデバイスが選択された状態になっています。**OK** を選択して次に進みます。
13. デバイスのデータ消去に関する警告メッセージが表示されたら、[Yes] を選択してパーティションのレイアウトを実行することを承認します。
14. 次の画面では、必要に応じてパーティションのレイアウトを変更できます。パーティションを変更した場合は、/boot/efi ファイルシステム・パーティションのタイプが vfat に、スワップ・パーティションのタイプが swap に、その他のファイルシステムが xfs に、それぞれ設定されていることを確認します。**OK** を選択してパーティションのレイアウトを確定します。

15. ネットワーク・インターフェイスの設定に *Bootp* または *DHCP* を使用して次に進むにはそのまま **OK** を選択します。スタティック IP を使用する場合は、必要なパラメータを入力してから **OK** を選択して次に進みます。

メモ：IX ブリックのバス 1: スロット 2 (IO9 のすぐ右) に Ethernet ボードを追加する場合は、25 ページの「Ethernet デバイスの固定命名」を参照してください（このようなボード追加はほとんどの場合必要ありません）。

16. 前の画面でスタティック IP を選択した場合は、システムのホスト名を尋ねられます。ホスト名を入力し、**OK** を選択して次に進みます。
17. 次の画面では、ファイアウォールの設定が必要かどうか尋ねられます。適切なオプションを選択し、**OK** を選択して次に進みます。
18. 次の画面では、システムがサポートする追加言語を選択できます。

メモ：追加言語を選択する場合、デフォルト言語として使用する言語も指定する必要があります。デフォルト言語はインストール完了後、システムで使用されるようになります。デフォルト言語はインストール後も変更できます。現時点では、コンソールでのアジア言語はサポートされません。SGI では、アジア系言語で使用するシステムのデフォルト・システム言語として英語 (English) を選択することを推奨します。

19. 次の画面ではシステムのタイム・ゾーンを設定します。適切な情報を指定して、**OK** を選択して次に進みます。
20. 次の画面では、システムの root パスワードが尋ねられます。root ログインに使用するパスワードを入力して、**OK** を選択します。
21. 次の画面ではシステムのコマンドを作成します。システムにログインする root 以外のコマンドを設定します。コマンドを作成する場合はここで各欄に必要な情報を入力して、**OK** を選択します。すべて空欄のまま **OK** を選択すると、コマンドを作成せずに次に進みます。
22. 前の画面でコマンドを作成した場合は、次のコマンドを作成する画面が表示されます。必要に応じてコマンドを追加していきます。**OK** を選択して次に進みます。
23. 次の画面では、認証の設定を行います。**OK** を選択して次に進みます。

24. [Package Group Selection] 画面が表示されます。ここでは、表示されるすべてのパッケージをインストール対象として選択する必要があります。[Everything] (リストの下端) を選択してください。Tab キーで **OK** をハイライト表示し、Enter キーを押して次に進みます。
25. [Installation to begin] 画面が確認のために表示されます。**OK** を選択して次に進みます。
26. インストールが開始されます。インストール中に SGI Advanced Linux Environment 2.1 CD2 をセットするように指示されます。指示に従って CD をセットし、**OK** を選択します。
27. インストールが完了したら **OK** を選択します。システムが再起動されます。この時点で CD をドライブから取り出せます。

以上でベース環境が更新されました。この時点ではまだインストールは完了していません。システム再起動後、11 ページの「SGI ProPack for Linux のインストール」を参照して SGI ProPack をアップグレードしてください。

SGI ProPack 旧リリースから SGI ProPack 2.2 へのアップグレード

ここでは、SGI ProPack 旧リリース (SGI ProPack 2.1, 2.1.1, 2.1.2) から SGI ProPack 2.2 へアップグレードする方法について説明します。コンソール設定のオプションとその手順については、『SGI Altix 3000 User Guide』の「Monitoring Your Server」を参照してください。

SGI ProPack 2.2 へのアップグレードはベース環境が変更されるため、従来のアップグレード手順と異なります。SGI Advanced Linux Environment 2.1 CD1 からのブートが必要になります。

7 ページの「SGI Advanced Linux Environment 2.1 ベース OS のインストール」のステップ 1 からステップ 9 までを実行すると、インストール・タイプの選択画面が表示されます。アップグレード・インストールを選択した場合は、続けて次の手順を実行してください。

1. [Upgrade Existing System] を選択してから **OK** を選択し、Enter キーを押します。
2. 警告メッセージを無視するために [Ignore] を選択して Enter キーを押します。SCSI ディスクが複数存在する場合は、ディスクの数だけ警告が繰り返されます。
3. システム上に複数の root ファイルシステムがある場合は、アップグレードしたい root パーティションを選択してから **OK** を選択し、Enter キーを押します。root ファイルシステムが 1 個だけの場合は、そのファイルシステムがアップグレードされる旨のメッセージ画面が表示されます。**OK** を選択し、Enter キーを押します。

4. [Customize Packages to Upgrade] 画面が表示されます。[No] を選択し、Enter キーを押します。
5. アップグレードすべきパッケージを検索中であることを示すメッセージ画面が表示されます。すべてのパッケージの検出に数分かかります。
6. [Upgrade to begin] 画面が表示されます。OK を選択し、Enter キーを押します。
7. アップグレード処理の途中で、SGI Advanced Linux Environment 2.1 CD2 をセットするよう求められます。メッセージが表示されたら CD をドライブにセットして Enter キーを押します。
8. インストールが完了したら OK を選択します。システムが再起動されます。この時点で CD をドライブから取り出せます。

以上でベース環境が更新されました。この時点ではまだインストールは完了していません。システム再起動後、11 ページの「SGI ProPack for Linux のインストール」を参照して SGI ProPack をアップグレードしてください。

SGI ProPack for Linux のインストール

ここでは、SGI ProPack v2.2 for Linux CD1 (Open/Free Source Software) と SGI ProPack v2.2 for Linux CD2 (Proprietary Software) のインストール手順を説明します。コンソール設定のオプションとその手順については、『SGI Altix 3000 User Guide』の「Monitoring Your Server」を参照してください。

1. システムに root でログインします。
2. SGI ProPack v2.2 for Linux CD1 をシステムにセットして、次のコマンドを入力して CD をマウントします。

```
mount /dev/cdrom /mnt/cdrom
```
3. 「/mnt/cdrom/INSTALL」と入力します。
4. SGI ProPack の [Welcome] 画面が表示されます。OK を選択して次に進みます。
5. 次の画面でインストールのタイプを選択します。[Yes] を選択してから、OK を選択します。

6. [Package Group Selection] 画面が表示されます。この画面では、インストールするパッケージ・グループが選択できます。上下矢印キーを押して選択行を移動し、スペースバーで選択します。パッケージ・グループを選択後 Tab キーを押して **OK** をハイライト表示し、Enter キーを押すと、選択したパッケージ・グループに対応する RPM がインストールされます。デフォルトではすべてのパッケージが選択されています。選択を変更した場合は、[SGI Proprietary] オプションが選択されていることを確認してください。XVM 等、SGI のライセンス・ソフトウェアを使用する場合は、[SGI Lincensed] が選択されていることを確認してください。ライセンス・ソフトウェアを使用しない場合はこのアイテムは選択しません。**OK** を選択して次に進みます。
7. [Installation to Begin] 画面が表示され、インストールのログが `/tmp/sgi-install.log` に保存されることが示されます。Tab キーを押して **OK** をハイライト表示し、Enter キーを押してください。
8. インストールが開始されます。[Package Installation] 画面にインストール中のパッケージ名とインストール経過時間が表示されます。

メモ：SGI ProPack インストールをシステム・コンソールから開始した場合は、パッケージのインストール中に "INIT: *version number reloading*" というメッセージが表示されます。このメッセージはエラーではないので無視できます。

9. インストール中に SGI ProPack v2.2 for Linux CD2 をセットするように指示されます。指示に従って CD をセットし、**OK** を選択して Enter キーを押します。
10. インストールが完了すると [Complete] 画面が表示されます。Enter キーを押してください。root のプロンプトに戻ります。
11. インストール完了後、インストールされた SGI ProPack for Linux software 環境を有効にするためにシステムを再起動してください。「reboot」と入力して Enter キーを押すと、システムを再起動できます。CD はドライブから取り出せます。

System Controller Software 1.1 CD のインストール

ここでは、System Controller Software 1.1 (for IRIX and SGI Linux) のインストール手順を説明します。このソフトウェアは MIPS プロセッサ・ベースの SGI システム Origin / Onyx 3000 Series および SGI Altix 3000 システムに対してのみライセンスされます。

1. システムに root でログインします。初期インストールで指定したパスワードを入力します。
2. System Controller Software 1.1 CD をシステムにセットして、次のコマンドを入力して CD をマウントします。

```
mount /dev/cdrom /mnt/cdrom
```

3. /mnt/cdrom/RPMS/ia64 ディレクトリに移動して、次のコマンドを入力します。

```
./install
```

4. インストールが開始されます。インストール中のパッケージ名が画面に表示されます。インストール完了後、root のプロンプトに戻ります。
5. インストールが完了したら System Controller Software 1.1 CD を取り出します。

ソフトウェアのアップグレード

SGI ProPack ソフトウェアのアップグレードが必要になったら、新しい SGI ProPack CD セットを使って、本書 11 ページの「SGI ProPack for Linux のインストール」の手順を実行してください。

システム上の全ソフトウェアを再インストールする場合は、最新の SGI ProPack CD セットを使って、この章で説明されるインストール手順すべてを繰り返してください。

損傷した Root ファイルシステムの修復

Root ファイルシステムが損傷した場合は、次の手順で修復を試みてください。

1. SGI Advanced Linux Environment 2.1 CD1 をシステムの CD-ROM ドライブにセットし、システムを再起動します。
2. システム起動時に、デバイス・マッピング・テーブルを確認します。次のような出力を確認してください。

Device mapping table

```
fs0 : Pci(2|1)/Ata(Primary, Master)/CDROM(Entry1) (この行が CD-ROM)
fs1 : Pci(1|1)/Scsi(Pun0/Lun1)/HD(Part1,Sig00000000)
fs2 : Pci(1|1)/Scsi(Pun0/Lun2)/HD(Part1,Sigg1)
blk0 : Pci(2|1)/Ata(Primary, Master)
blk1 : Pci(2|1)/Ata(Primary, Master)/CDROM(Entry1)
blk2 : Pci(1|1)/Scsi(Pun0/Lun1)
blk3 : Pci(1|1)/Scsi(Pun0/Lun1)/HD(Part1,Sig00000000)
blk4 : Pci(1|1)/Scsi(Pun0/Lun1)/HD(Part2,Sig00000000)
blk5 : Pci(1|1)/Scsi(Pun0/Lun1)/HD(Part3,Sig00000000)
blk6 : Pci(1|1)/Scsi(Pun0/Lun1)/HD(Part4,Sig00000000)
blk7 : Pci(1|1)/Scsi(Pun0/Lun2)
blk8 : Pci(1|1)/Scsi(Pun0/Lun2)/HD(Part1,Sigg10)
```

3. **Shell>** プロンプトが表示されたら、次のように CD-ROM デバイス名を指定します。

```
Shell> fs0: (CD-ROM デバイス名を指定して移動)
fs0:\>
```

4. 次の行に示すように「**elilo linux rescue**」と入力します。

```
fs0:\> elilo linux rescue (ブートローダを起動)
```

5. CD がブートされ、レスキュー・モードが開始します。
6. このレスキュー・プロセスで使用する言語を選びます。**OK** を選択して次に進みます。

メモ：現時点では、コンソールでのアジア言語はサポートされません。SGI では、アジア系言語で使用するシステムの言語として英語 (English) を選択することを推奨します。

7. 次の画面では、レスキュー用イメージがあるメディアを選択できます。現時点で有効な選択肢は [Local CDROM] のみです。**OK** を選択して次に進みます。

8. 次の画面では、root ドライブをマウントするオプションが選択できます。ファイルシステムが損傷を受けている場合は、xfs_repair コマンドを実行する前にマウントを試みることに有効な場合があります。これによって、重要なデータをジャーナルから復元できる可能性があります。

[Continue] を選択してインストーラがハングアップしたりクラッシュした場合は、再起動の後、ここまでのステップをやり直し、[Skip] を選択してください。その後、ファイルシステムをマニュアルでマウントします。この章の「Root ファイルシステムのマニュアル・マウント」を参照してください。

9. これまでの手順が成功した場合は、システムが /mnt/sysimage に検出されたというメッセージが次の画面に表示されます。**OK** を選択してレスキュー・モードのシェルに入ります。
10. この時点で、管理コマンドを使用して破損・削除されたソフトウェアの修復や再ロードが行えます。ここでは、破損した root ファイルシステムを修復する例を示します。
11. ジャーナルからデータを復元するためなどの理由でファイルシステムをマウントしている場合は、xfs_repair コマンドを実行する前にファイルシステムをアンマウントする必要があります。引数を指定せずに mount コマンドを実行すると、現在マウントされているファイルシステムの一覧が表示されます。例を次に示します。

sh-2.05# mount

```
rootfs on / type rootfs (rw)
devfs on /dev type devfs (rw)
/dev/root.old on / type ext2 (rw)
none on /dev type devfs (rw)
/proc on /proc type proc (rw)
/dev/pts on /dev/pts type devpts (rw)
/tmp/cdrom on /mnt/source type iso9660 (ro)
/tmp/loop0 on /mnt/runtime type cramfs (ro)
/dev/xscsi/pci01.03.0-1/target1/lun0/part3 on /mnt/sysimage type xfs (rw)
/dev/xscsi/pci01.03.0-1/target1/lun0/part1 on /mnt/sysimage/boot/efi type vfat (rw)
none on /mnt/sysimage/dev/pts type devpts (rw)
none on /mnt/sysimage/proc type proc (rw)
```

/mnt/sysimage type xfs (rw) で終わる行に、root ディスクに対応するデバイスが表示されています。上の例では /dev/xscsi/pci01.03.0-1/target1/lun0/part3 です。

12. 次のコマンドを実行してファイルシステムをアンマウントします。

```
umount /mnt/sysimage
```

13. ファイルシステムに対して xfs_repair を実行します。例を次に示します。

```
xfs_repair /dev/xscsi/pci01.03.0-1/target1/lun0/part3
```

14. `xfs_repair` が完了したらファイルシステムは修復されています。システムを再起動してください。

Root ファイルシステムのマニュアル・マウント

インストーラが `root` ファイルシステムを検出できない場合は、マニュアルでマウントする必要があります。ここではこの方法について説明します。詳細は『XFS for Linux Administration』を参照してください。まれに、破損したファイルシステムをマウントしようとする `mount` コマンドがハングアップしたりトラブルを起こすことがあります。この場合は、『XFS for Linux Administration』を参照してマウントせずに `xfs_repair` を使用する方法を試してください。

1. 次の例では、`root` ファイルシステムが IX ブリックの先頭のディスク・ベイにあるディスクのパーティション 3 にあることを前提としています。システムからアクセス可能なパーティションの一覧を表示するには、次の例に示すコマンドを実行します。

```
sh-2.05# cat /proc/partitions
```

```
major minor #blocks name
5 0 125470 xscsi/pci01.01.0/target0/lun0/disc
4 0 35843686 xscsi/pci01.03.0-1/target1/lun0/disc
4 1 513008 xscsi/pci01.03.0-1/target1/lun0/part1
4 2 9438208 xscsi/pci01.03.0-1/target1/lun0/part2
4 3 25891840 xscsi/pci01.03.0-1/target1/lun0/part3
4 16 35843686 xscsi/pci01.03.0-1/target2/lun0/disc
4 17 513008 xscsi/pci01.03.0-1/target2/lun0/part1
4 18 9438208 xscsi/pci01.03.0-1/target2/lun0/part2
4 19 25891840 xscsi/pci01.03.0-1/target2/lun0/part3
```

2. システムが問題なく実行中である場合は、次の例に示すような XSCSI パス名が IX ブリックのベイ 1 にあるシステム・ディスクを表します。この例では、`root` ファイルシステムはパーティション 3 に置かれていると仮定しています。

```
/dev/xscsi/pci01.03.0-1/target1/lun0/part3
```

3. ファイルシステムをマニュアルでマウントするには、次のコマンドを実行します。

```
sh-2.06# mkdir /mnt/sysimage
```

```
sh-2.07# mount -t xfs /dev/xscsi/pci01.03.0-1/target1/lun0/part3 /mnt/sysimage
```

サポートされる製品

この章では、SGI Altix 3000 システムでサポートされる製品コンポーネントについて説明します (すべての製品の一覧は 18 ページの 表 3-1 を参照してください)。

この章では、各製品コンポーネントの説明を次のグループに分けて説明します。

- OS 拡張
- I/O サブシステム
- HPC アプリケーション・ツールとそのサポート
- システム管理ツール
- NUMA tools

SGI ProPack for Linux release で SGI から供給されるソフトウェアは、SGI Altix 3000 用カーネル RPM、SGI Advanced Linux Environment 2.1 RPM セット、および SGI システム上での使用

を前提とした SGI ソフトウェアから構成されます。SGI ProPack for Linux 製品の一覧と説明を表 3-1 に示します。

表 3-1 SGI ProPack v2.2 for Linux 製品

製品	説明
Application performance measuring tools	<p>以下のプログラム最適化ツールを含みます。</p> <p>VTUNE - このツールは Intel によって開発、サポートされます。Itanium プロセッサのパフォーマンス測定機能を利用して、消費時間やその他のプロセッサ・アーキテクチャ・イベントをプロファイリングします。プロファイリング・データはアプリケーションのパフォーマンス解析、チューニング、改良に利用できます。</p> <p>pfmon- このツールはオープン・ソースで、GPL ライセンスに沿って配布されます。Itanium プロセッサのパフォーマンス測定機能を制御するコマンド・ライン・インターフェイスを提供します。このツールが出力するデータを加工して、各種のアプリケーション・パフォーマンスのレポートが作成できます。レポートはアプリケーション・パフォーマンスの評価、チューニング、改良に利用できます。pfmon パッケージにはカスタム・パフォーマンス・ツールを作成するためのライブラリ・インターフェイス libpfmon.a も含まれます。</p>
Array Services	SGI システム・クラスタで実行されるシステムと並列アプリケーションの管理を容易にするための、カーネル・サポートを備えたツールを提供します。
CpuMemSets	プロセッサとメモリの配置を実装するインフラストラクチャとカーネル・サポートを提供します。
cpuset	CPU セット指定機能を提供します。
CSA	ジョブ・ベースのアカウンティング機能を提供します。Linux システム上で、特定のログイン・アカウントに対してタスク単位のリソースおよびディスク消費のアカウンティングを行います。
FLEXlm	Floating License のためのランタイム環境です。Floating License を処理するためのデーモン群を含みます。
IOC4 driver	内臓 IDE CD-ROM、NVRAM、リアルタイム・クロック制御用ドライバです。

表 3-1 SGI ProPack v2.2 for Linux 製品 (続き)

製品	説明
kdb	実行中のシステムに対するカーネル・デバッグをサポートします。キーボードからの直接制御とシリアル・コンソールからの制御をサポートします。
Kernel partitioning support	パーティション・システムをサポートするために必要なソフトウェア・インフラストラクチャを提供します。パーティション間通信サポートを含みます。
Kernel performance improvements	コミュニティ・ベースで作成されたパッチを含みます。2.5.xカーネルおよびSGIカーネルでのO(1)スケジューラのサポート、Big Kernel Lock (BKL) でのロック競合を削減するためのパッチ、FRlocks (Fast-Reader Locks) による xtime_lock (gettimeofday が使用) の競合の削減のためのパッチなどが含まれます。
L1 and L2 System Controller firmware	ブリックとシステム・コンピュータ・ラック内の電源、冷却ファン、テスト機能を管理・モニタするためのサポートを提供します。
LKCD	システム・クラッシュ・ダンプ解析ツールを提供します。lcrash と、システム・クラッシュ・ダンプの保存と設定に必要な全ユーザ・レベル・スクリプトが含まれます。
MPT	SGI システム用に最適化された、業界標準のメッセージ交換ライブラリです。
NUMA Tools	NUMA 関連ツールのコレクション (dlook、dplace など) です。
Performance Co-Pilot collector infrastructure	大規模システム用のパフォーマンス・モニタとパフォーマンス解析サービスです。
PROM	システム・ブートとソフトウェア・インストール機能を提供します。
runon	コマンドを特定の CPU または CPU 群で実行できます。
SGI Advanced Linux Environment 2.1 updates	SGI ProPack 同梱の SGI Advanced Linux Environment 2.1 ベースに対するアップデート RPM セットです。
XFS	Linux にハイパフォーマンス・ファイルシステムを導入します。

表 3-1 SGI ProPack v2.2 for Linux 製品（続き）

製品	説明
XSCSI infrastructure	QLogic QLA12160 SCSI ホスト・バス・アダプタと ATAPI CDROM、ディスクをサポートします。QLogic Fibre Channel のホスト・バス・アダプタもサポートされます。サポートされるカードについては『SGI Altix 3000 User's Guide』を参照してください。インフラストラクチャとして、エラー修正ハンドリング、フェイルオーバー、SAN のサポートを含みます。Linux SCSI レイヤと共に、あるいは単独で使用できます。Linux SCSI レイヤと協調動作するように設定すると、一般的なディスク・デバイス名（ <code>/dev/sda</code> など）が使用されます。
XVM	ディスクのストライプ化、ミラーリングなどのソフトウェア・ボリューム管理機能を提供します。

次のソフトウェアは SGI によってサポートされません。

- SGI 以外から供給されるソフトウェア
- Red Hat の発表する他のリリース、アップデート、パッチ
- Linux コミュニティ、あるいは他のベンダーから入手したソフトウェア・パッチ、ドライバ、あるいは他のソフトウェア改変
- SGI の指定する以外のパラメータ設定やその他のモジュールを使って再コンパイルまたは再設定されたカーネル。特に、ext3、ReiserFS、LVM、md カーネル・コンポーネントは品質、機能、パフォーマンスに著しく影響するため、サポートされません。これらのコンポーネントの代わりに、XFS と XVM を使用してください。
- サポート外のハードウェア構成やデバイス

OS 拡張

SGI は、世界中で絶え間なく拡張・改善され続ける Linux OS の商用・企業環境の開発の中でも特に、大規模計算・大規模データに対するハイパフォーマンス・コンピューティング (HPC) 環境に重点を置いています。このために、SGI では IRIX OS と MIPS プロセッサ・ベースのシステムで培った NUMAflex と HPC の技術を活用し、Linux カーネルの中でも HPC 環境において重要な機能の改善に努めています。

CpuMemSets のサポート

CpuMemSets は、複数 CPU のスケジューリングと複数ノードのメモリ・アロケーションを行うシステム・サービスやアプリケーションに必要な Linux カーネル機能を提供します。SGI ProPack のカーネルとライブラリをインストールすることで、CpuMemSets サポートは有効になります。

デフォルト設定では、すべての CPU とメモリがすべてのアプリケーションで使用可能です。CpuMemSets 機能を利用して、任意のプロセス、プロセス・ファミリ、あるいはプロセス仮想メモリ領域を、システム中の特定の CPU 群やメモリ部分に制限されるように指定できます。

runon コマンドは CpuMemSets を利用して、特定のコマンドを特定の CPU 群の上で実行させます。CpuMemSets システム・インターフェイスにアクセスするために、C 共有ライブラリと Python 言語モジュールが供給されます。

SGI ProPack for Linux の将来のリリースでは、CpuMemSets 機能に簡単にアクセスできるその他のシステム・サービスが追加される計画です。

CpuMemSets サポートに関するドキュメント情報やその他の詳細は、『Resource Administration Guide for Linux』の「CPU Memory Sets and Scheduling」という題の章を参照してください。CpuMemSets RPM の一部としてインストールされるファイルの一覧は次のコマンドで参照できます。

```
rpm -ql CpuMemSets
```

また、*cpumemsets* と *runon* のマン・ページにも重要な情報が含まれています。

CpuMemSets は SGI のオープン・ソース・プロジェクトです。CpuMemSets は次のサイトからも入手可能です。

<http://oss.sgi.com/projects/cpumemsets>

Cpuset のサポート

Cpuset システムは、あるプロセスまたはプロセス群が使用できるプロセッサの数を限定するための機能を提供するワークロード管理ツールです。

Cpuset は、特定の CPU の集合に名前を付けたもので、制限付き (restricted) または開放 (open) として定義されます。制限付き Cpuset は、その Cpuset に属するプロセスだけがその CPU 群で実行できます。開放 Cpuset では任意のプロセスが実行できますが、その Cpuset に属するプロセスはその CPU 群内でのみ実行されます。Cpuset は Cpuset 設定ファイルと名前によって定義されます。

大規模システムのシステム管理者は、Cpuset を使って CPU 区画を設定できます。各区画システムでは、特定のプロセス群を特定の CPU 群を使って実行でき、これらのプロセスとシステムの他の処理との間の干渉や競合を軽減します。制限付き Cpuset の場合、Cpuset にアタッチされたプロセスはシステム上の他の処理の影響を受けません。Cpuset にアタッチされたプロセスだけが、Cpuset 内の CPU 群で実行されます。開放 Cpuset は、特定のプロセスをある一定の CPU 群でのみ実行させて、システムの他の部分への影響を最小限に抑えるために使用できます。

完全システム・アカウンティング (CSA: Comprehensive System Accounting)

完全システム・アカウンティング (CSA: Comprehensive System Accounting) ソフトウェア・パッケージは SGI と Los Alamos National Laboratory (LANL) の共同オープン・ソースを IRIX から Linux に移植したもので、ジョブ・ベースのアカウンティング機能を提供します。Linux システム上で、特定のログイン・アカウント群に対してタスク単位のリソースおよびディスク消費のアカウンティングを行います。

高機能システム・アカウンティング機能は超大規模システム、特に複数の組織で共同利用されるシステムで重要になります。CSA はジョブ・コンテナ機能を使って Linux 上でジョブ概念を定義します。ジョブはいくつかのプロセスを内包するコンテナで、任意のシステム・エントリ・ポイント（インタラクティブ・ログイン、cron ジョブ、リモート・ログイン、バッチ処理など）へのリソース配分を把握するために用いられます。

Linux版CSAはSGIのオープン・ソース・プロジェクトです。CSAは次のサイトから入手可能です。

<http://oss.sgi.com/projects/csa>

CSA サポートに関するドキュメント情報やその他の詳細は、『Resource Administration Guide for Linux』の「Comprehensive System Accounting」という題の章を参照してください。

システム・パーティション

SGI Altix 3000 では、単一システムを複数のシステム・パーティションに分割して使用することが可能です。各パーティションはそれぞれ独立した OS カーネルを実行でき、それぞれが個別のシステム・コンソール、root ファイルシステム、IP ネットワークアドレス、および物理メモリを所有します。各パーティション間はパーティション化されない場合と同様に SGI のハイパフォーマンス NUMAlink を使って接続されています。すなわち、パーティション化されたシステムは NUMAlink 接続されたノード群（クラスタ）と見ることもできます。

パーティション化によって、障害耐性を高め、NUMAlink 接続とグローバル共有メモリなどの Altix 3000 機能を活用したハイパフォーマンス・クラスタを構成できます。

システム・パーティションの詳細はリリース・ノート『SGI ProPack v2.2 for Linux Release Notes』を参照してください。

I/O サブシステム

ある種の HPC ワークロードは主に CPU リソースを消費するのに対し、大量のデータを処理し、メモリと外部記憶の間でデータを高速転送するために、または大規模外部記憶システムを効率的に利用するために I/O サブシステムの能力を要求するワークロードもあります。IRIX から移植された SCSI サブシステム、XFS ファイルシステム、XVM ボリューム・マネージャや、その他のデータ転送機能は、安定性と頑健性に富むハイパフォーマンス・ストレージ I/O サブシステムを Linux に導入します。

ここでは、PCI-X バス固定ナンバリング、Ethernet デバイス固定命名、XSCSI サブシステム、XSCSI-SCSI サブシステム、XFS ファイルシステム、XVM ボリューム・マネージャの概要について説明します。

PCI-X バス固定ナンバリング

PCI-X バスの固定ナンバリングにより、システムの障害や再設定のために再起動を行ってもバス番号が維持されます。プラットフォーム初期化中にバスが検出されると、論理バス番号が割り当てられます。それぞれのバスにはシステム全体で一意的な論理番号が割り当てられます。SGI Altix 3000 システムでは、デフォルトで 256 個のバス（バス番号 0-255）がサポートされます。

デフォルトでは、I/O ブリックが接続されている最下位の C ブリック・モジュール ID からバス番号が割り当てられていきます。I/O ブリックとは、SGI Altix 3000 システムの PX ブリック、または IX ブリックです。各 I/O ブリックには 0x10 個のバスが割り当てられますが、実際には、現時点では各 I/O ブリックは 6 個のバスしかサポートしません。このため、システム全体ではバス番号は連続せず、空隙が発生します。

ある I/O ブリックに 6 個のバスが物理的に存在すると仮定します。これらのバスには 0x1 から 0x6 までの番号が、I/O ブリック背面から見て左から右への順序で割り当てられます。システム中に複数の I/O ブリックが存在する場合は、次の I/O ブリックのバスには 0x11 から 0x16 までの番号が割り当てられます。この番号の最下桁は、各 I/O ブリックのバス位置に刻印されている 1 から始まる番号に対応します。

1 個の I/O ブリックしかない場合は、バスの固定ナンバリングは不要です。ただし、複数の I/O ブリックがシステム中に存在する場合は、固定ナンバリングの使用が強く推奨されます。I/O ブリックの 1 個にブート時障害が発生しても、バス番号は変更されません。

バス固定ナンバリングを使用するには、カーネルに対して `ioconfig=` パラメータを指定する必要があります。

`ioconfig=` パラメータはカンマで区切られた I/O ブリック番号のリストを引数に取ります。次の行は、`elilo` によるマニュアル・ブートの例を示しています。この例では、カーネルに対して I/O ブリック 101.01 にバス番号 0x1 から 0x6 までを、I/O ブリック 101.02 にバス番号 0x11 から 0x16 までを割り当てるように指定しています。

```
Shell> elilo vmlinux ioconfig="101.01,101.02"root=/dev/xscsi/pci01.03.0-1/target1/lun0/part3
```

`ioconfig` で指定する数字は、L2 コントローラからシステムを調べることによってわかります。例を次に示します。

```
l2-pumpkin-001-L2>pwr
001c11:
power appears on
```

```
001c27:
power appears on
101i01:
power appears on
101p02:
power appears on
```

上の例で、101i01 は前述の elilo ブート・オプションで示した IX ブリック 101.01 を表しています。同様に 101p02 は PX ブリック 101.02 を表します。

多くの場合、システムを自動ブートできるように設定するのが妥当です。このためには、elilo.conf ファイルで ioconfig パラメータを設定します。実行中の Linux システムでは、elilo.conf ファイルは /boot/efi/EFI/sgi ディレクトリからアクセスできます。elilo.conf ファイルの例を次に示します。

```
prompt
timeout=50
relocatable
default=sgilinux
append="ioconfig=101.01,101.02"

image=vmlinuz-2.4.19-sgi21r4
    label=sgilinux
    read-only
    root=/dev/xscsi/pci01.03.0-1/target1/lun0/part3
```

上の例で、append= 行に注意してください。elilo.conf ファイルでは、ここに ioconfig 情報を記述します。

Ethernet デバイスの固定命名

Ethernet デバイスに対する固定命名は SGI 独自のメカニズムで、SGI Altix 3000 システムにも採用されています。固定命名とは、SGI Altix 3000 システムの IO9 インターフェイスに接続された Gigabit Ethernet カードを常に eth0 に設定するためのメカニズムです。このメカニズムにより、SGI Altix 3000 システムに複数の Ethernet デバイスが存在する場合でも、MAC アドレスを基準にしてベース Ethernet デバイス番号を正しく割当てます。

/etc/sysconfig/networking/eth0_persist ファイルに Ethernet デバイス番号と MAC アドレスの対応が記述されます。このファイルが存在しないとシステムのブート時に /etc/rc.d/init.d/eth_persist スクリプトが実行されて、ファイルが生成されます。実際

に IO9 Ethernet カードの MAC アドレスに `eth0` を確実に割当てするには、SGI Altix 3000 システムを新規インストール後に最初に起動した際に、このファイルの編集が必要な場合があります。

固定命名によって、Ethernet カードを追加しても Ethernet デバイス番号と MAC アドレスの対応が維持されることを保証する以外に、Ethernet デバイスの番号付けを制御できます。たとえばデバイス番号 `ethX` の Ethernet カードが欠けたとき、その機能を Ethernet カード `ethY` で補いたい場合は `/etc/sysconfig/networking/eth0_persist` ファイルを編集することでその Ethernet カードに番号 `ethX` を振り直すことができます。

`/etc/sysconfig/networking/eth0_persist` ファイルの例を次に示します。

```
eth0 08:00:69:13:dc:ec
eth1 08:00:69:13:72:e8
```

このファイルにより、システムは次のように設定されます。

```
[root]# ifconfig -a
eth0 Link encap:Ethernet  HWaddr 08:00:69:13:DC:EC
      inet addr:128.162.246.125 Bcast:128.162.246.255 Mask:255.255.255.0
      UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
      RX packets:843 errors:0 dropped:0 overruns:0 frame:0
      TX packets:1245 errors:0 dropped:0 overruns:0 carrier:0
      collisions:0 txqueuelen:100
      RX bytes:386044 (376.9 Kb)  TX bytes:126741 (123.7 Kb)
      Interrupt:59

eth1 Link encap:Ethernet  HWaddr 08:00:69:13:72:E8
      BROADCAST MULTICAST  MTU:1500  Metric:1
      RX packets:136 errors:0 dropped:0 overruns:0 frame:0
      TX packets:0 errors:0 dropped:0 overruns:0 carrier:0
      collisions:0 txqueuelen:100
      RX bytes:8850 (8.6 Kb)  TX bytes:0 (0.0 b)
      Interrupt:63
```

IO9 のすぐ右のスロット（バス 1: スロット 2）に Ethernet ボードを接続して、ボード接続後に Linux を再インストールした場合または固定命名ファイル `eth0_persist` を削除した場合は、IO9 は `eth0` ではなく `eth1` になります。この設定は製品ライセンスに影響する可能性があるため、推奨されません。Ethernet ボードを他の IX ブリック・スロットに接続すれば、この問題は発生しません。

バス 1 のスロット 2 に Ethernet ボードを接続する必要がある場合は、Ethernet ボードを接続しない状態でまず SGI ProPack を含む OS をインストールします。SGI ProPack のインストール後、Ethernet ボードをバス 1 のスロット 2 に取付けます。

XSCSI サブシステム

SGI XSCSI システムを Linux に付加することで、IRIX システムの頑健なエラー・ハンドリングとフェイルオーバー機能、ストレージ・エリア・ネットワーク (SAN) インフラストラクチャがサポートされます。XSCSI は、IRIX システムで長年にわたって大規模システム向けにパフォーマンス・チューニングされたメカニズムです。XSCSI は、標準的なオープン・ソース・ドライバそのままでは対応できないような SGI アーキテクチャの機能を活用します。

XSCSI デバイス名は次のような形式を持ちます。

```
/dev/xscsi/pci01.03.0-1/target1/lun0/part1
```

この XSCSI デバイス名は次のコンポーネントに分割できます。

pci01	システム・バス番号 0x1
03	バス内でのデバイス番号 3
0-1	論理ユニット 0、ポート 1

デバイス番号 (スロット番号)、論理ユニット、ポート番号は固定です。これらの値が変更されることはありません。ただし、システム・バス番号はハードウェア障害 (I/O ブリックがブートしない等) や再設定によって変更する可能性があります。

PCI-X のバス固定ナンバリング (24 ページの「PCI-X バス固定ナンバリング」参照) が有効になっている場合はバス番号が固定され、ハードウェアの障害や再構成があってもバス番号が変更されることはありません。XSCSI 名をマウントやデバイス指定に使用するときは、バス固定ナンバリングを使用すると、その XSCSI デバイス名は再起動を経ても変更されません。

XSCSI-SCSI サブシステム

XSCSI-SCSI サブシステムは XSCSI デバイスによる SCSI デバイスのエミュレーションをサポートします。XSCSI-SCSI サブシステムは、SCSI ディスク、SCSI テープ、SCSI ジェネリック・デバイスなどのドライバを使用するように書かれたプログラムが、XSCSI ドライバで制御されるデバイスを制御できるようにします。

XFS ファイルシステム

SGI XFS ファイルシステムは、業界トップレベルのハイパフォーマンス・ファイルシステムを Linux に導入します。XFS はオープン・ソースの、高速復帰性を持つジャーナル・ファイルシステムで、ダイレクト I/O サポート、スペース・プリアロケーション、アクセス・コントロール・リスト、quotas、その他の商用ファイルシステム機能を備えています。Linux 上で供給される他のファイルシステムも使用可能ですが、IRIX で長年にわたってパフォーマンス・チューニングと改善を重ねてきた XFS は、HPC 環境で必要になる大規模なデータ I/O ワークロードに特に適しています。

XFS ファイルシステムの詳細は『XFS for Linux』を参照してください。

XVM ボリューム・マネージャ

SGI XVM ボリューム・マネージャは、複数の物理ディスク記憶装置を 1 個の論理的な単位（論理ボリューム）として結合するための論理構造を提供します。論理ボリュームは標準的なパーティションと同じように扱えるため、パーティションが使えるどのような引数にも指定できます。

論理ボリュームを使用することで、単一の物理ディスクよりも大きな raw デバイスやファイルシステムを構成できます。論理ボリュームは複数のディスクをストライプ化することも可能なので、ディスク I/O パフォーマンスの改善にも利用できます。また、他のディスク上のデータのミラー化にも利用できます。

HPC アプリケーションのツールとサポート

SGI IRIX から、HPC ライブラリ、ツール、およびソフトウェア・パッケージが Linux と Itanium 2 をベースにした高性能 HPC 環境のためのシステムへ移植されています。ここではこれらのツール、ライブラリ、ソフトウェアのいくつかについて説明します。

Message Passing Toolkit

SGI Message Passing Toolkit (MPT) は、SGI システム用に最適化された業界標準メッセージ交換ライブラリを提供します。Linux では、MPT は MPI と SHMEM API を含み、メモリ間転送を高速に行うブロック転送エンジン (BTE) などの SGI ハードウェアを低階層でできるように

して、ハードウェア・メモリ・コントローラのフェッチ (fetchop) をサポートします。fetchop は複数の MPT プロセス間の直接通信と同期をサポートし、OS システムコールの使用によるオーバーヘッドを避けます。

MPI ジョブのような並列ワークロードは、SGI Array Services ソフトウェアを使用することでクラスタまたはパーティション・システム上で起動、制御、監視できます。Array Services によって、クラスタの複数ノードや複数のシステム・パーティションで実行できるプロセスのセットを定義するための array セッションを記述できます。Array Services はプロセス・コンテナを提供するカーネル・モジュール Process Aggregates (PAGGs) を使って実装されています。PAGGs は SGI によって Linux 上でオープン・ソース化されています。

Message Passing Toolkit の詳細については『Message Passing Toolkit: MPI Programmer's Manual』を参照してください。

Performance Co-Pilot

SGI Performance Co-Pilot ソフトウェアは大規模システム用に設計されたパフォーマンス・モニタとパフォーマンス管理のサービスのセットで、IRIX から Linux へ移植されています。Performance Co-Pilot は低階層のハードウェア・カウンタや MPT と統合され、CPU、I/O、ネットワークの統計サービス、ビジュアル化ツール、モニタ・ツールなどを提供します。

システム管理

ここではシステム管理ツールについて説明します。システム管理ツールにはシステム起動に必要なハードウェア・ソフトウェア環境、ライセンス管理、システム・コンソール、およびシステム・コントローラが含まれます。

PROM チップ

PROM (Programmable read only memory) チップはシステム・ハードウェアに組み込まれているメモリで、CPU のブート、システム管理、ソフトウェアのインストールなどのために使用されます。PROM チップはディスクや OS の一部ではなく、システムの最下層に配置されます。PROM

は内容を消去したり、実行をバイパスできません。PROM の詳細は `prom(1)` マン・ページを参照してください。

拡張ファームウェア・インターフェイス (EFI: Extensible Firmware Interface)

SGI Altix 3000 には、CPU への入出力をサポートするために Extensible Firmware Interface (EFI) プラットフォームが用意されています。また EFI はサーバのブート設定を管理し、ブート・メニューを不揮発性メモリに保存します。

SGI ProPack v2.2 でサポートされる `elilo-3.3` は EFI 1.10 に準拠して、ブートローダ (`elilo.efi`) が EFI システム・パーティションに置かれます。EFI 1.10 仕様によると、ブートローダ・ファイルは各ベンダの専用ディレクトリに置く必要があります。SGI Altix システムでは、このディレクトリは `/boot/efi/EFI/sgi/` です。`elilo` バージョン 3.3 では、ロードされたこのディレクトリの内部でのみ、設定ファイルが検索されます。カーネルに対しても同様の制限が適用されるため、カーネルは `/boot/efi/` から `/boot/efi/EFI/sgi/` に移動されています。この変更についての詳しい説明は `/usr/share/docs/` にある `elilo` のドキュメントを参照してください。

EFI プロンプトからは、基本的なファイル管理（テキスト編集を含む）、設定の変更、ブート時に実行されるスクリプトの作成などのタスクが実行できます。EFI コマンドの概要を表 3-2 に示します。

表 3-2 EFI コマンド

EFI コマンド	説明
<code>alias [-bdv] [sname] [value]</code>	エイリアスを設定、表示します。
<code>attrib [-b] [+/- rhs] [file]</code>	ファイル属性を設定、表示します。
<code>bcfg</code>	ブート・ドライバの設定とオプションのロードを行います。
<code>cd [path]</code>	カレント・ディレクトリを変更します。
<code>cls [background color]</code>	画面をクリアします。
<code>comp file1 file2</code>	2 つのファイルを比較します。
<code>cp file [file] ... [dest]</code>	ファイルやディレクトリをコピーします。
<code>date [mm/dd/yyyy]</code>	日付を設定、表示します。

表 3-2 EFI コマンド (続き)

EFI コマンド	説明
<code>dbl device [Lba] [blocks]</code>	ブロック I/O デバイスの 16 進ダンプを出力します。
<code>dh [-b] [-p prot_id] [handle]</code>	ハンドル情報を出力します。
<code>dmpstore</code>	変数ストアをダンプします。
<code>echo [-on -off] [text]</code>	テキストを標準出力にエコー出力したり、スクリプトのエコー出力のオン・オフを制御します。
<code>edit [file name]</code>	ファイルを編集します。
<code>endfor</code>	(スクリプト内のみ) ループの末尾を示します。
<code>endif</code>	(スクリプト内のみ) IF THEN 構文の末尾を示します。
<code>err [level]</code>	エラー・レベルを設定、表示します。
<code>exit</code>	終了します。
<code>for var in set</code>	(スクリプト内のみ) ループ構文を始めます。
<code>getmtc</code>	単調カウンタ値を取得します。
<code>goto label</code>	(スクリプト内のみ) ラベル位置にジャンプします。
<code>guid [-b] [sname]</code>	既知 guid ID 一覧を表示します。
<code>help [-b] [internal command]</code>	コマンド・ヘルプを表示します。
<code>if [not] condition then</code>	(スクリプト内のみ) IF THEN 構文を始めます。
<code>load driver_name</code>	ドライバをロードします。
<code>ls [-b] [dir] [dir] ...</code>	ディレクトリの内容を表示します。
<code>map [-bdvr] [sname[:]] [handle]</code>	デバイス・パスに対する短縮名をマッピングします。
<code>mem [address] [size] [;MMIO]</code>	メモリやメモリ・マップド I/O をダンプします。
<code>memmap [-b]</code>	メモリ・マップをダンプします。
<code>mkdir dir [dir] ...</code>	ディレクトリを作成します。
<code>mm address [width] [;type]</code>	Mem、MMIO、IO、PCI のメモリ設定を変更します。
<code>mode [col row]</code>	テキスト・モードを設定、表示します。
<code>mount BlkDevice [sname[:]]</code>	ブロック・デバイス上のファイルシステムをマウントします。

表 3-2 EFI コマンド（続き）

EFI コマンド	説明
<code>mv <i>sfile dfile</i></code>	ファイルを移動します。
<code>pause</code>	（スクリプト内のみ） 終了 / 継続を尋ねるプロンプトを出力します。
<code>pci [<i>bus dev</i>] [<i>func</i>]</code>	PCI デバイス情報を表示します。
<code>reset [<i>cold/warm</i>] [<i>reset string</i>]</code>	リセットの設定 (Cold/Warm) を示します。
<code>rm <i>file/dir</i> [<i>file/dir</i>]</code>	ファイルやディレクトリを削除します。
<code>set [<i>-bdv</i>] [<i>sname</i>] [<i>value</i>]</code>	環境変数を設定、表示します。
<code>setsize <i>newsize fname</i></code>	ファイルのサイズを設定します。
<code>stall <i>microseconds</i></code>	<i>x ms</i> 間の遅延を作ります。
<code>time [<i>hh:mm:ss</i>]</code>	時刻を設定、表示します。
<code>touch [<i>filename</i>]</code>	ファイル属性を設定、表示します。
<code>type [<i>-a</i>] [<i>-u</i>] [<i>-b</i>] <i>file</i></code>	ファイル内容を出力します。
<code>ver</code>	バージョン情報を表示します。
<code>vol <i>fs</i> [<i>volume label</i>]</code>	ボリューム・ラベルを設定、表示します。

FLEXlm

FLEXlm は Macrovision から提供される、頑健性と柔軟性を備えたライセンス管理システムです。サードパーティは開発する製品に簡単にライセンス機能を追加でき、システム管理者は最低限の労力でライセンスのインストールと管理が行えます。FLEXlm では、単純なノードロック・ライセンスと冗長サーバで有効なフローティング・ライセンス、それらに伴う幅広いライセンス・オプションがサポートされます。

ライセンス付きのソフトウェアをビルドするには Macrovision からキーのセットを購入する必要があります。システム管理者は任意の場所にライセンス・サーバをインストールできます。SGI が販売する製品は通常、FLEXlm を使ってライセンスが設定されています。

詳細については下記サイトを参照してください。

<http://www.macrovision.com/solutions/esd/flexlm/flexlm.shtml>

SGIconsole

SGIconsole は、IRIX OS および SGI ProPack for Linux を実行する SGI サーバ群をコンソールから管理、モニタするためのハードウェアとソフトウェアです。サーバには SGI Altix 3000 サーバなどの、SGI パーティション・システムや 単一システム・イメージ・サーバなどを含められます。

SGIconsole は 1U ラックマウント・タイプの Intel Pentium ベース SGI サーバ、シリアル・マルチプレクサまたは Ethernet ハブ、および Console Manager パッケージと Performance Co-Pilot（ここからハードウェア / ソフトウェアのリモート管理ツール群にアクセスします）を含むソフトウェア・パッケージから構成されます。

Console Manager は SGIconsole ツールの GUI となり、複数の SGI サーバを制御するために使用されます。SGIconsole はコマンドライン・インターフェイスからも使用できます。SGIconsole の詳細については『SGIconsole Start Here』を参照してください。

システム・コントローラ・ファームウェア

L1/L2 コントローラは SGI システムのシステム・コントローラ・ファームウェアです。

SGI Altix 3000 システム と SGI Origin / Onyx 3000 Series システムの各ブリックには L1 コントローラが内蔵されています。L1 コントローラは各ブリックの電源投入と制御のシーケンス、温度と電源供給のモニタを行います。

L2 コントローラはラック・レベルのコントローラで、ラック内のブリックを監視、制御します。システム内の L2 コントローラ同士はネットワーク接続されて、各ブリックの制御と監視の情報を交換し、システム全体の制御と監視に利用します。

L1/L2 システム・コントローラ・ファームウェアについての詳細は『SGI L1 and L2 Controller Software User's Guide』を参照してください。

NUMA Tools

ここでは、NUMA 関連ツールとして現在供給されているコマンド群について説明します。

dlook コマンド

dlook コマンドは指定したプロセスのメモリ・マップと CPU 利用状況を表示します。プロセス仮想アドレス空間の各ページについて、次の情報が出力されます。

- ページを所有するオブジェクト（ファイル、SYSV 共有メモリ、デバイス・ドライバなど）
- ページ・タイプ（RAM、FETCHOP、IOSPACE など）
- RAM メモリでは次の情報が出力されます。
 - メモリ属性（SHARED、DIRTY など）
 - ページが配置されているノード
 - ページ物理アドレス（オプション）

オプションとして、プロセスが使用した各システム物理 CPU の 消費 CPU time も表示できます。

dplace コマンド

dplace コマンドは関連するプロセス群を特定の CPU（ノード）のグループにバインドして、プロセスの移行が発生するのを防ぎます。このツールはメモリ・アクセスをローカル・ノードに集中させることにより、ある種のソフトウェアのパフォーマンスを改善できます。

パフォーマンス・チューニング

この章では、シングルプロセッサまたはマルチプロセッサによるプログラミングでのチューニングに関する話題を取り上げます。本章の内容は次の各節に分かれます。

- シングルプロセッサ・プログラムのパフォーマンス・チューニングについては 36 ページの「シングルプロセッサ・コードのチューニング」を参照してください。
- マルチプロセッサ・プログラムのパフォーマンス・チューニングについては 45 ページの「マルチプロセッサ・コードのチューニング」を参照してください。
- シングルプロセッサ、マルチプロセッサの双方のプログラミングのパフォーマンス・チューニングで利用できるツールについては 50 ページの「プロファイラとパフォーマンス・ツール」を参照してください。

xvi ページの「関連ドキュメント」に挙げた参考文献も参照してください。

システム構成の確認

使用中のシステムの詳しい状態は `/proc` 疑似ファイルシステムで調べられます(詳細は `proc(5)` マン・ページを参照してください)。たとえば、次のような情報が得られます。

- `/proc/cpuinfo` は各プロセッサの状態を出力します。クロック・スピードとプロセッサのステッピング番号がわかります。
- `/proc/meminfo` はシステム・メモリの全体的な使用状況(総容量、空き容量、スワップ容量など)を表示します。
- `/proc/discontig` はメモリ利用をページ単位で表示します。
- `/proc/pal/cpu0/cache_info` は L1/L2/L3 キャッシュ構造についての詳細(サイズ、レイテンシ、連想度、ライン・サイズなど)を表示します。`/proc/pal/cpu0` 内のその他のファイルは TLB (Translation Lookaside Buffer) 構造、クロック比、その他の詳細を表します。

- `/proc/version` は現在インストールされているカーネルについての情報を表示します。
- `/proc/perfmon` ファイルが `/proc` に存在しない（エクスポートされていない）場合は、カーネルはパフォーマンス・カウンタを開始しておらず、カウンタを使用するパフォーマンス・ツールは動作しません。
- `/proc/mounts` は現在マウントされているファイルシステムの詳細を表示します。
- `/proc/modules` は現在インストールされているカーネル・モジュールについての詳細を表示します。

`uname` コマンドはカーネルのバージョンその他のシステム情報を出力します。また、`topology` コマンドは `/dev/hw` の情報を元にシステム構成を表示する Bourne シェル・スクリプトです。詳細は `topology(1)` マン・ページを参照してください。

SGI Altix 3000 システムで現在推奨されるコンパイラは `efc` および `ecc`（Intel Fortran コンパイラおよび Intel C/C++ コンパイラ）です。他のコンパイラ（GNU コンパイラなど）も使用できますが、`efc` と `ecc` が Linux システム上での最善のパフォーマンスを導きます。この章の説明では、特に注意がない限り `efc` / `ecc` コンパイラを前提とします。

シングルプロセッサ・コードのチューニング

シングルプロセッサ・コードのパフォーマンスをチューニングするときは、次の基本ステップに沿ってください。

- 正しい結果が得られていることを確認してからチューニングを開始します。詳細は 37 ページの「出力の正確性の確認」を参照してください。
- 数値演算ライブラリ、科学計算ライブラリなどのパッケージに含まれるチューニング済みの既存コードを使用します。詳細は 41 ページの「チューニング済みコードの使用」を参照してください。
- チューニングが必要な部分を決定します。詳細は 41 ページの「チューニング必要部分の決定」を参照してください。
- コンパイラを適切に実行します。詳細は 42 ページの「コンパイラ・オプションの選択と適用」を参照してください。

- キャッシュ・パフォーマンスのチューニングの可能性を検討します。詳細は 43 ページの「キャッシュ・パフォーマンスのチューニング」を参照してください。
- メモリ管理モードを変えるための環境変数を設定します。詳細は 43 ページの「メモリ管理」を参照してください。
- スタックとデータ・セグメントを変更します。詳細は 44 ページの「chatr によるスタックとデータ・セグメントの変更」を参照してください。

出力の正確性の確認

チューニングの最初にまず行うのは、正しい出力が得られていることを確認することです。出力の正確性が確認できたら、チューニングを開始できます。まず特別な最適化はすべて無効にし、デフォルトの最適化に絞って出力をチェックします。これは、コンパイラ・オプションとデバッグ・ツールを使って行います。

次のコンパイラ・オプションはパフォーマンスのトレースと移植に関して役立ちます。

- `-O: -O0` はすべての最適化をオフにします。デフォルトは `-O2` です。
- `-g: -g` オプションはデバッグ・シンボルを有効にします。
- `-mp: -mp` オプションは浮動小数点最適化を制限して、宣言された精度を守ります。
- `-IPF_fltacc: -IPF_fltacc` オプションは精度に影響する浮動小数点関連の最適化を無効にします。
- `-r: -i: -r8` および `-i8` オプションを指定するとデフォルトの実数、整数、および論理値のサイズが 8 バイトになります。Cray コードを移植するときに効果があります（このオプションは `intrinsic/external` ライブラリ関数を明示的に宣言します）。

デバッグ・ツールは、出力の正確性を確認するために使用できます。次のデバッグが使用できます。

- `gdb`: GNU プロジェクトのデバッグです。C、C++、Fortran 95 で書かれたプログラムのデバッグに使用できます。C や C++ でコンパイルするときは、コンパイラ・コマンド・ラインに `-g` オプションを指定して `gdb` の使用する `dwarf2` シンボル・データベースが生成されるようにします。

Fortranのデバッグにgdbを使用するときは-gおよび-O0オプションを指定します。Fortranプログラムで-O1またはそれ以上のオプションを指定する場合は、gdbは使用しないで下さい。

Fortran 95 コード用のデバッガは

http://sourceforge.net/project/showfiles.php?group_id=56720 からダウンロードできます（標準のgdbコンパイラはFortran 95コードには対応していません）。正しいバージョンのgdbを使用しているかどうか確認するにはgdb -vを実行します。出力が次のようになることを確認してください。

```
GNU gdb 5.1.1 FORTRAN95-20020628 (RC1)
Copyright 2002 Free Software Foundation, Inc.
```

gdb コマンドは http://sources.redhat.com/gdb/onlinedocs/gdb_toc.html からダウンロードできるユーザーズガイド、またはgdbコマンドのhelpオプションで一覧が参照できます。現時点のバージョンのgdbはar.ecレジスタの状態を正しく報告しません。アセンブリ・コード・レベルで、ソフトウェア・パイプラインを使用した、ローテーションする、レジスタ・ベースのループをデバッグする場合は、idbを使用してください。

- idb:Linuxプラットフォーム用の完全な機能を備えたIntel製のシンボリック・デバッガです。このデバッガはC、C++、FORTRAN 77、Fortran 90で書かれたプログラムのデバッグを幅広くサポートします。現時点では、idbはItaniumシステムで動作するマルチスレッド・プログラムやマルチプロセッサ・プログラムのデバッグには使用できません。

シェル・コマンドでidbに-gdbオプションを付けて実行すると、使用できるユーザ・コマンドとデバッガ出力がgdb風になります。

- ddd: コマンドライン・デバッガに接続するGUIです。gdbとidbがサポートされます。詳細は38ページの「dddの使用」を参照してください。

dddの使用

dddは任意のコマンドライン・デバッガに接続できるGUIツールです。デフォルトではgdbに接続されます。新たにdddを起動するときは--debuggerオプションでデバッガが指定できます（例:--debugger "idb"）。

起動後、画面の各ペインに次の情報が表示されます。

- 配列表示
- ソースコード

- 逆アセンブル・コード
- デバッガ・エンジンへのコマンドを入力するウィンドウ

各パネルのオン・オフは [View] メニューから制御できます。

頻繁に使うコマンドはメニューから実行できます。また、次の操作も覚えておくと便利です。

- アセンブラ画面でアドレスを選択してマウス右ボタンをクリックして [lookup] を選択すると、コマンド画面で `gdb lookup` コマンドが実行されて対応するソース行が表示されます。
- ソース画面で変数を選択してマウス右ボタンをクリックすると、現在の値が表示されます。配列は配列表示ウィンドウに表示されます。配列は [Menu]> [Print Graph] オプションを選択すると PostScript 形式で出力されます。
- レジスタ・ファイルの内容 (general, floating-point, NaT, predicate, application) は [Status] メニューの [Regsisters] コマンドを選択すると表示されます。[Status] メニューからは、スタック・トレースの表示やスレッドの切り替えも行えます。

ヒープ破壊問題対策

`glibc` の `malloc/free` を使用してメモリを動的に管理するプログラムで発生する可能性のあるヒープ破壊問題には、2 つのチェック方法があります。1 つは環境変数を使用する方法、もう一つは `Electronic Fence` を使う方法です。

`MALLOC_CHECK_` 環境変数の値を 1 にすると診断メッセージが表示されます。2 にすると、ヒープ破壊が検出された瞬間に実行が中断されます。

`ElectricFence` は Linux および Unix 用の `malloc` デバッガです。このデバッガを使用すると、`malloc()` メモリ・バッファがオーバーラン (またはアンダーラン) した瞬間にプログラムを停止して、ソースコード中のバグ発生個所を表示します。このデバッガはメモリ割当て用ブロックの先頭または末尾を不正なページに隣接させて、エラー発生時のバッファオーバーランまたはアンダーランに対してセグメンテーション・エラーを起こします。また、すでに開放されたメモリ領域がアクセスされたときも検出されます。

アンダーラン / オーバーランは同時に検出できません。デフォルトの動作では、割当てられたメモリの直後にアクセス不能なページを配置しますが、環境変数 `EF_PROTECT_BELOW` を設定する

ことで逆の状況も設定できます。Electric Fence を使用するには libefence ライブラリをリンクします。次の例を参照してください。

```
% cat foo.c
#include <stdio.h>
#include <stdlib.h>
int main (void)
{
    int i;
    int * a;
    float *b;

    a = (int *)malloc(1000*sizeof (int));
    b = (float *)malloc(1000*sizeof (float));

    a[0]=1;
    for (i=1 ; i<1001;i++)
    {
        a[i]=a[i-1]+1;
    }
    for (i=1 ; i<1000;i++)
    {
        b[i]=a[i-1]*3.14;
    }

    printf("answer is %d %f \n"a[999],b[999]);
}
```

このプログラムのコンパイル、実行は次のようになります（ライブラリを指定してコンパイルしたときのエラー表示に注意してください）。

```
% ecc foo.c
% ./a.out
answer is 1000 3136.860107
% ecc foo.c -lefence
% ./a.out

Electric Fence 2.2.0 Copyright (C) 1987-1999 Bruce Perens
Segmentation fault
%
```

巨大コア・ファイルの生成を回避するために推奨される方法は、デバッガから Electric Fence を使用することです。詳細は efence マン・ページを参照してください。

チューニング済みコードの使用

可能な場合は、ハードウェア・パフォーマンスに対して最適化されたチューニング済みコードを使用してください。

次の数値演算関数ライブラリを必要なときに使用することで、最善の結果が得られます。

- MKL: Intel の Math Kernel ライブラリです。BLAS、LAPACK、および FFT ルーチン群が含まれます。
- VML: Vector Math ライブラリは、MKL パッケージの一部 (libmkl_vml_itp.so) として提供されます。
- Standard Math library: 標準数値計算ライブラリ関数は Intel コンパイラの libimf.a で提供されます。-lm オプションが指定されると、glibc の libm ルーチンが優先的にリンクされます。

MKL と VML のドキュメンテーションは次の Web サイトで入手できます。

http://intel.com/software/products/perflib/index.htm?iid=ipp_home+software_libraries&

チューニング必要部分の決定

チューニングでは、コードの中で改善効率の高い部分を特定することが重要です。次のツールを活用してください。

time	ユーザ・タイム、システム・タイム、および経過時間を測定します。
gprof	プログラムの実行プロファイル (pcsamp プロファイル) が得られます。-p コンパイラ・オプションを指定すると gprof が使用可能になります。
VTune	Intel のパフォーマンス・モニタ・ツールで、Linux サーバ、Windows クラウドのアプリケーションです。複数の Itanium/Linux システムのリモート・サンプリングをサポートします。
pfmon	Itanium と Linux のために設計されたパフォーマンス・モニタ・ツールです。Itanium の Performance Monitoring Unit (PMU) を使ってバイナリを加工せずに直接カウント、サンプリングします。

VTune と pfmon の詳細は、50 ページの「プロファイラとパフォーマンス・ツール」を参照してください。

コンパイラ・オプションの選択と適用

コンパイラ・オプションには、パフォーマンス最適化に使用できるものがあります。efc や ecc のオプションの概要を見るには、-help オプションを付けてコンパイラをコマンドライン実行してください。-dryrun オプションを指定すると、efc や ecc が生成するドライバ・ツール・コマンドが表示されます。このオプションは実際にツールを実行しません。

次のオプションは、パフォーマンス・チューニングに利用できます。

- | | |
|-----------------------|---|
| -ftz | アンダーフローをゼロにしてカーネル・トラップを回避します。-O3 最適化レベルではデフォルトで有効になります。 |
| -fno-alias | ポインタ・エイリアスがないものと見なします。エイリアス関連オプションにはこの他に -ansi_alias と -fno_fnalias があります。エイリアスを使用すると不正なコードが生成される可能性があることに注意してください。 |
| -ip | 単一ファイル内でのプロシージャ間の最適化を行います。 |
| -ipo | 複数ファイル間のプロシージャ間の最適化を行います。 |
| -O3 | -O2 レベルの最適化に加えて、ループ変換やプリフェッチなど、より積極的な最適化を行います。レベル 3 最適化は、必ずしもすべてのプログラムでパフォーマンス改善に寄与するとは限りません。 |
| -opt_report | 最適化に関するレポートを生成し、-opt_report_file で指定されるファイルに保存します。 |
| -override_limits | 公式にドキュメント化されているオプションではありませんが、コンパイラの内部制限を越えて最適化処理を継続することを許します。 |
| -prof_gen / -prof_use | プロファイリング情報を生成、利用します。これらのオプションを使用するときは、次の 3 ステップの手順が必要です。 <ol style="list-style-type: none">1. -prof_gen を指定してコンパイルします。2. 試験データを使ってプログラムを実行します。 |

- S -prof_use を指定して、試験実行のデータを元にコンパイルします。コンパイルを実行して、アセンブル・リストを .s ファイルに生成しますが、リンクは行いません。アセンブル・リストは -opt_report オプション出力と照合して、ループ最適化の成否をチェックするために使用します。

キャッシュ・パフォーマンスのチューニング

キャッシュのパフォーマンスをチューニングするには、次のような方法があります。

- キャッシュのスラッシングを引き起こすような、大きな 2 の指数幅のストライドや次元指定は避けます。
- 可能な限りストライド 1 を使用します。
- 同じ 16 バイト幅バンクへのアクセスが 2 つ同時に発生すると、キャッシュ・バンクのコンフリクトが発生する可能性があります。pfmon -e L2_OZQ_CANCEL1_BANK_CONF コマンドの出力と pfmon -e CPU_CYCLES コマンドの出力に、トータル CPU サイクル数に比べて多数のバンク・コンフリクトが検出される場合は、配列パディングの方法を変えてみてください。2 つのコマンドは、次の行で 1 つにまとめることができます。

```
% pfmon -e CPU_CYCLES,L2_OZQ_CANCEL1_BANK_CONF a.out
```

パフォーマンスのためのイベント監視では、最大で同時に 4 つのイベントがカウントできません。

- 同時に使用するデータはグループ化し、コード中では可能な限りベクトルを使わないようにします。
- テンポラリ配列の使用は避け、データ・コピーをできるだけ減らします。

メモリ管理

glibc の malloc/free でメモリの割当と開放を頻繁に行うコードは、メモリ管理オーバーヘッドによってシステム時間を消費する傾向があります。デフォルトでは、glibc はシステム全体のメモリ効率をパフォーマンスよりも優先します。この話題の詳細は <http://www.linuxshowcase.org/ezolt.html> で解説されています。

ハイパフォーマンス指向のメモリ管理モードを有効にするには、次の環境変数を設定してください。

```
% setenv MALLOC_TRIM_THRESHOLD_ -1
% setenv MALLOC_MMAP_MAX_ 0
```

efc では組み込みのメモリ割当てでも内部的に glibc の malloc コマンドを使用するので、これらの環境変数は malloc/free で管理される Cray ポインタ等を使った Fortran コードでも原則的に有効です。ただし Fortran 90 のアロケータブル配列は Fortran ライブラリ呼出しで直接管理されているため、効果はありません。

chatr によるスタックとデータ・セグメントの変更

chatr コマンドは、プログラムの ELF ヘッドを変更して、実行時にスタック・セグメントとデータ・セグメントにある命令が実行可能かどうかを切り替えます。

デフォルトでは、スタック・セグメントとデータ・セグメントの両方が実行可能です。特定のワークロードでは、デフォルトの設定をオフにしてスタックとコード・セグメントを実行不可にすることでパフォーマンスを改善できます。この手法は、多数の fork() 呼出しを伴うワークロードで特に効果があります。fork() 呼出しをあまり含まないワークロードでは、パフォーマンスの大幅な改善は見込めません。

sysctl コマンドを使って、この動作をシステム全体で変えることができます。次のコマンドは、スタックおよびデータ・セグメントを常に実行不可にします。

```
% sysctl vm.executable_stacks=0
```

次のコマンドは、スタックおよびデータ・セグメントを実行可能にすることを許可します。

```
% sysctl vm.executable_stacks=1
```

この変更で、プログラムによっては従来問題なかったものが SEGV でエラーを起こす可能性があります。問題を起こすのは通常、スタックまたはデータ・セグメントに命令を保存していて、そこへ分岐するプログラムです（たとえば古いバージョンの X サーバには、グラフィックス・ドライバをデータ・セグメントへロードするものがあります。Java JIT コンパイラも同様の問題を持ちます）。このようなプログラムも、chatr コマンドによって vm.executable_stacks の値に関係なく正しく実行できるようになります。

詳しい使用法については、chatr(1) マン・ページを参照してください。

マルチプロセッサ・コードのチューニング

マルチプロセッサ・コードのチューニングを行う前に、まずシングルプロセッサのチューニングを行ってください。多くの場合、マルチプロセッサ・コードでも良い結果が得られます。詳細は 36 ページの「シングルプロセッサ・コードのチューニング」を参照してください。

マルチプロセッサ・チューニングは大別して次のステップに分かれます。

- コードの並列化手法を選択します。詳細は 45 ページの「コードの並列化」を参照してください。
- コードが適切に並列化されているかどうかを解析します。詳細は 47 ページの「コード中のボトルネックの検出」を参照してください。
- 無効な共有がないかどうかをチェックします。詳細は 47 ページの「無効な共有の修復」を参照してください。
- データの配置をチューニングします。詳細は 48 ページの「dplace と runon の利用」を参照してください。
- チューニングを支援する環境変数を設定します。詳細は 49 ページの「環境変数によるパフォーマンス・チューニング」を参照してください。

コードの並列化

マルチプロセッサ環境でのパフォーマンス・チューニングで最初に行うことは、チューニングに使用する並列化手法を選択することです。次の中から選べます。

- SGI Message Passing Toolkit (MPT)によるメッセージ交換インターフェイス(MPI: Message Passing Interface) を使う方法。この MPI は一般的な MPI ライブラリに比べて、SGI Altix 3000 Series システム用に最適化されています。SGI Altix 3000 アーキテクチャと SGI Linux NUMA の特徴が活用できます。

MPI を使用するには `-lmpi` コンパイラ・オプションを使います。サポートされる環境変数の一覧は `mpi(1)` マン・ページを参照してください。環境変数の一部は IRIX システムでのみ有効です（マン・ページに注記されています）。

SGI MPT version 1.6.1 以降の `MPIO_DIRECT_READ` と `MPIO_DIRECT_WRITE` は Linux でローカル XFS システムに対して使用できます。

- OpenMP を使用する方法。C、C++、Fortran コードで OpenMP ディレクティブを使用する場合、次のコンパイラ・オプションが利用できます。

`efc -openmp / ecc -openmp`

Intel コンパイラに組み込まれている Intel OpenMP フロントエンドを利用します。生成される実行ファイルは Intel OpenMP ランタイム・ライブラリ `libguide.so` への呼出しを行います。

`guide`

Intel コンパイラで OpenMP コードを使用するもう一つの方法です。`guidec` (`ecc` の代わりに)、`guideefc` (`efc` の代わりに)、または `guidec++` によって、OpenMP ディレクティブを含むコードを `libguide` 呼出しにコンパイルします。詳細は 53 ページの「その他のパフォーマンス・ツール」を参照してください。

`-openmp` オプションによるコンパイルが、Intel が推奨する Linux OpenMP コンパイラ手法です。この方法でパフォーマンスに問題がある場合は、`guide` によって改善される可能性があります。

- コンパイラの自動並列化機能を利用する方法。`efc / ecc` コンパイラで `-parallel` オプション、`-par_report` オプションを使用します。これらのオプションは並列化されたループを表示し、また、並列化されなかったループについてはなぜされなかったのかが説明されます。ソース・ファイルに多数のループが含まれるときは、`-override_limits` フラグを使って自動並列化を促進する必要があります。`-parallel` によって生成されるコードは OpenMP API を利用しており、標準的な OpenMP 環境変数や Intel 拡張環境変数が有効です。

コードの並列化度を見積もります。並列化度は次の式で計算できます。

$$p = N(T(1) - T(N)) / T(1)(N-1)$$

ここで $T(1)$ はコードがシングル CPU で実行される場合の実行時間、 $T(N)$ は N 個の CPU で実行される場合の時間です。速度比は $speed-up = T(1)/T(N)$ で求められます。

$speed-up/N$ が 50% 未満の場合 ($N > (2-p)/(1-p)$) の場合)、CPU 数の追加は中断してスケーラビリティ改善のためのチューニングを検討してください。

CPU の活動度は `top`、`vmstat` などのコマンドや、PCP ツールを利用（たとえば `pmval kernel.percpu.cpu.user`）するか、あるいは Performance Co-Pilot のビジュアル化機能を使ってリモート IRIX ワークステーションから確認できます。ハードウェア情報は `/proc/cpuinfo` や `/proc/meminfo` で確認できます。キャッシュ等の詳細情報は `/proc/pal/cpuX` (X は CPU 番号) で確認できます。

コード中のボトルネックの検出

コード中のボトルネックを検出できる各種ツールについての詳細は、50 ページの「プロファイラとパフォーマンス・ツール」を参照してください

無効な共有の修復

並列化後のプログラムが並列化前よりも遅い場合は、無効な共有（偽共有、False Sharing）が発生している可能性があります。無効な共有は、プロセッサのデータ・キャッシュの同じキャッシュ・ラインに対応する 2 つまたはそれ以上のデータ・アイテムが、共有メモリ・アプリケーションの複数のスレッドによってアクセスされない場合に発生します。異なる CPU 上で実行されている 2 つのスレッドが同じキャッシュ・ラインを変更しようとする、と、キャッシュ・ラインは正確な内容を保ったまま複数の CPU 上で常駐できなくなり、一貫性を保つためにメモリ・サブシステムに移動させられます。これはアプリケーションのパフォーマンスとスケーラビリティの低下を招きます。データ・アイテムの書き込みがなく読出しのみの場合は、キャッシュ・ラインは関与するすべての CPU 間で共有状態のまま残ります。無効な共有は、共有アレイの中で隣接する要素が異なるスレッドによって変更されたときに発生することがあります。

無効な共有が発生しているかどうかは、次のいくつかの方法で検出できます。

- パフォーマンス・モニタ pfmon の出力で、BUS_MEM_READ_BRIL_SELF イベントと BUS_RD_INVALID_ALL_HITM イベントを監視します。
- pfmon を使って DEAR イベントを監視し、キャッシュ・ラインを追跡します。
- PCP pmshub ユーティリティを使って、キャッシュ・トラフィックと CPU 利用率をモニタリングします。

メモ：pmshub ユーティリティは初期リリースには同梱されない可能性があります。最新情報はリリース・ノートを参照してください。

無効な共有が問題を引き起こしている場合は、次の方法で解決を試みます。

- HW カウンタを使ったプロファイルを実行して、キャッシュ・ライン共有のための記憶領域をモニタリングします。これによって、問題が発生している場所がわかります。
- データ構造またはアルゴリズムを変えます。

- 共有オブジェクトのプライベート / パブリック変数、共有データ、静的変数、コモン・ブロックを調べます。
- クリティカル・リージョンを使って問題を引き起こしているコード部分を特定します。

dplace と runon の利用

dplace コマンドは、複数のプロセスを指定された CPU 群にラウンドロビン方式でバインドします。バインドされたプロセスは移行されることはありません。この動作は IRIX システムの DSM_MUSTRUN と同様です。dplace のナンバリングは現在の CPU メモリ・セットのコンテキストで決まります。

runon コマンドは、指定されたリスト中の CPU にプロセスの実行を制限します。ただし、プロセスはリストされた CPU 間で自由に移動できます。

これらのコマンドの詳細については、以下の各節を参照してください。

MPIコードに対する dplace の使用

CPU とデータの親和性を高める最善の配置方法は、次のいずれかの手法を用いることです。

- Cpuset で MPI_DSM_MUSTRUN を使用します。
- MPI_DSM_CPULIST 環境変数を設定します。
- dplace-s1 コマンドを使用します。

SGI の MPI コンパイルされたコードを使用するときは、dplace のフル・パスと -s1 オプションを指定します。例を次に示します。

```
% mpirun -np 32 /usr/bin/dplace -c16-47 -s1 ./a.out
```

-s1 オプションで、dplace が軽量 MPI 管理プロセス (MPI sepherd process) を配置する手順をスキップします。

OpenMPコードに対する dplace の使用

Intel が提供する OpenMP コードを使用する場合は -x6 オプションを使用します。例を次に示します。

```
% setenv OMP_NUM_THREADS 4
% dplace -x6 -c5-8 ./a.out
```

-x6 オプションで、dplace が 2 つの軽量 OpenMP 管理プロセスを配置する手順をスキップします。

dplace と runon の違いに注意してください。

```
% setenv OMP_NUM_THREADS 4
% dplace -x6 -c5-8 ./a.out
% runon 5-8 ./a.out
```

dplace コマンドは管理プロセスの配置をスキップできるのに対して、runon はスキップできません。dplace コマンドはまた、各スレッドを特定の CPU に固定するのに対して、runon ではスレッドを CPU 5-8 の間に限定しますが、これらの CPU の範囲内では自由に移動できます。

dplace の詳細と使用法については、『Linux Resource Administration Guide』を参照してください。

環境変数によるパフォーマンス・チューニング

パフォーマンス・チューニングのためにいくつかの環境変数が利用できます。MPI の動作を制御する環境変数の詳細については、mpi(1) マン・ページを参照してください。

OpenMP 環境変数は、OpenMP ライブラリの動作に影響します。たとえばいくつかの環境変数は、アプリケーション中であることがないスレッドや、他のスレッドとの同期のため待機中のスレッドの動作を制御します。他の変数は、OpenMP ライブラリのスレッド間でのループ繰返しスケジュールを設定します。次の環境変数は OpenMP 標準で定義されます。

- OMP_NUM_THREADS(デフォルトはシステム中の CPU 数)
- OMP_SCHEDULE(デフォルトは static)
- OMP_DYNAMIC(デフォルトは false)
- OMP_NESTED(デフォルトは false)

上記の環境変数以外にも、Intel による OpenMP 拡張が提供されています。次の 2 つは KMP_LIBRARY 変数による制御の例です。

KMP_LIBRARY 変数はランタイム実行モードを次のように設定します。

- serial に設定すると、シングルプロセッサ実行が適用されます。
- throughput に設定すると、処理待機中は CPU を他のプロセスに渡します。これはデフォルトの動作で、マルチユーザ環境でのシステム全体のパフォーマンスを維持します。IRIX での _DSM_WAIT=YIELD と同等の設定になります。
- turnaround に設定すると、待機処理中も CPU を明け渡しません。IRIX での _DSM_WAIT=SPIN と同等の設定になります。KMP_LIBRARY を turnaround に設定すると、他のユーザとの CPU リソース競合がない専用システムで実行されるベンチマークのパフォーマンスを改善できます。

プログラムの実行直後にセグメンテーション違反が発生するときは、KMP_STACKSIZE の値を増やしてみてください。これはスレッドのプライベート・スタック・サイズです。デフォルトは 4MB です。場合によってはシェルのスタック・サイズ制限も上げる必要があります。

プロファイラとパフォーマンス・ツール

パフォーマンス改善のポイントを特定するために使用できるツールがあります。以下にいくつかのツールの概要を説明します。

pfmon を使ったプロファイリング

pfmon ツールは Linux 用に設計されたパフォーマンス・モニタ・ツールです。Itanium Performance Monitoring Unit (PMU) を使って、バイナリを加工せずにカウント、サンプリングします。次のタスクが実行できます。

- 未加工のバイナリのプロセス単位でのモニタリング。
- システム全体のモニタリング・セッションの実行。任意の CPU で実行されている全プロセスのモニタリングが可能です。
- 専用 CPU 上でのシステム全体セッションの起動。複数 CPU 上で並列実行することも可能です。
- ユーザ・レベルまたはカーネル・レベルを指定したモニタリング。
- 基本イベント・カウントの集計。

- プログラムまたはシステムのサンプリング。同時に最大 4 個のイベントをモニタリングできます。

可能なオプションの一覧を見るには、`pfmon -help` コマンドを実行してください。

profile.pl スクリプト

`profile.pl` スクリプトは `dplace` の実行を含む、ユーザ・プログラム全体のプロファイリング処理を操作します。典型的な利用法は次の通りです。

```
% profile.pl -c0-3 -x6 command args
```

このスクリプトはプロセッサ 0-3 を指定します。-x6 は OpenMP コードを使用する場合のみ必要です。

CPU_CYCLES PMU イベントに対するプロファイルが `profile.out` に出力されます。このスクリプトでは、IA64_INST_RETIRED や L3_MISSES など他のイベントに対するプロファイリングも行えます。PMU イベントの一覧は `pfmon -l` の出力を参照してください。このスクリプトはパフォーマンス・モニタの下でのコマンド実行、実行ファイルとダイナミック・ライブラリにあるシンボル名とアドレスの対応づけをするマップ・ファイルの生成、プロファイル・アナライザの実行などもサポートします。

詳細は `profile.pl(1)`、`analyze.pl(1)`、`makemap.pl(1)` のマン・ページを参照してください。

MPI プログラムでの profile.pl の使用

MPI プログラムでは `profile.pl` スクリプトに `-s1` オプションを付けて実行してください。例を次に示します。

```
% mpirun -np 4 profile.pl -s1 -c0-3 test_prog </dev/null
```

`/dev/null` を指定することで、MPI プログラムは TTY 入力を待たずにバックグラウンド実行されます。

VTune を利用したリモート・サンプリング

Intel VTune パフォーマンス・アナライザはリモート・サンプリングによる試験を行います。VTune データ・コレクタは Linux システムで実行され、これに対応する GUI は IA32 Windows マシンで結果を解析するために用いられます。

VTune の詳細は下記 URL を参照してください。

<http://developer.intel.com/software/products/vtune/vtune61/index.htm>

メモ：VTune は初期リリースには同梱されない可能性があります。最新情報はリリース・ノートを参照してください。

GuideView の利用

GuideView はプログラムの並列実行に関するパフォーマンス情報をグラフィック表示します。GuideView は KAP/Pro Toolset の一部です。KAP/Pro Toolset はこの他に Guide OpenMP コンパイラと Assure Thread Analyzer を含んでいます。GuideView はデフォルトのソフトウェア・インストールではシステムにインストールされません。

GuideView は並列処理の問題点となるパフォーマンス・ボトルネックを、色を使って視覚的に示します。各プロセッサの動作状況は、階層化されたサマリによって詳細度別に報告されます。

統計データは対応するサマリを展開すると表示されます。サマリには、注目すべき場所に関する情報が含まれています（たとえば、全体のパフォーマンスに最もインパクトを与えるコード部分が示されます）。

プログラミングに有効な統計データを収集するには、コンパイラ・オプション `-O3`、`-openmp`、`-openmp_profile` を指定します。これによってデフォルト `libguide.a` の代わりに `libguide_stats.a` がリンクされます。次の例は、`swim` というファイルを生成するコンパイラ・コマンド行を示しています。

```
% efc -O3 -openmp -openmp_profile -o swim swim.f
```

プロファイリング・データを収集するには、このプログラムを次のように実行します。


```
% export OMP_NUM_THREADS=8
% ./swim < swim.in
```

プログラムが終了すると、GuideView で使用できる *swim.gvs* というファイルが生成されます。GuideView をこのファイルと共に起動するには、次のコマンドを実行します。

```
% guideview -jpath=your_path_to_Java -mhz=998 ./swim.gvs.
```

GuideView のグラフィック部分を実行するには Java が必要です。Java 1.1.6-8 および Java 1.2.2 がサポートされており、これ以降のバージョンでも動作することが予想されます。Java がない場合は GuideView の機能は大幅に制限されますが、テキスト出力によって情報を得ることができます。テキスト出力の一部を次に示します。

```
Program execution time (in seconds):
cpu                :          0.07 sec
elapsed           :          69.48 sec
  serial          :          0.96 sec
  parallel         :          68.52 sec
cpu percent        :          0.10 %
end

Summary over all regions (has 4 threads):
# Thread           #0          #1          #2          #3
Sum Parallel       :   68.304    68.230    68.240    68.185
Sum Imbalance      :    1.020    0.592    0.892    0.838
Sum Critical Section:    0.011    0.022    0.021    0.024
Sum Sequential     :    0.011    4.4e-03    4.6e-03    1.6e-03
Min Parallel       :  -5.1e-04  -5.1e-04    4.2e-04  -5.2e-04
Max Parallel       :    0.090    0.090    0.090    0.090
Max Imbalance      :    0.036    0.087    0.087    0.087
Max Critical Section:  4.6e-05    9.8e-04    6.0e-05    9.8e-04
Max Sequential     :    9.8e-04    9.8e-04    9.8e-04    9.8e-04
end
```

その他のパフォーマンス・ツール

次のパフォーマンス・ツールも、コードの最適化に有用です。

- *Guide OpenMP Compiler* は C、C++、Fortran に対する Intel の OpenMP 実装です。
- *Intel の Assure Thread Analyzer* はスレッド化アプリケーションのプログラミング・エラーをレコーディング無しで検出します。

これらの製品の詳細については、下記 URL を参照してください。

<http://developer.intel.com/software/products/threading>

メモ：これらの製品は、SGI 製品上での完全検証は行われていません。SGI は、サード・パーティ製品の動作や合目的性に対する保証は行いません。

索引

A

Application performance measuring tools 18

Array Services 18

C

CD からブート 7、14

CD の内容 2

CpuMemSets のサポート 21

CSA 22

CSA (Comprehensive System Accounting) のサポート 22

D

dlook コマンド 34

dplace コマンド 34

E

EFI (Extended Firmware Interface) のサポート 30

Electric Fence デバッガ 39

ELF ヘッダの変更 44

Ethernet デバイス名 25

F

FLEXlm のサポート 32

G

gdb ツール 37

GNU デバッガ 37

H

HPC サポート

Message Passing Toolkit 28

Performance Co-Pilot (PCP) 29

ライブラリとツール 28

I

idb ツール 38

I/O サブシステム

HPC システムのための 23

XFS ファイルシステムのサポート 28

XSCSI-SCSI のサポート 27

XSCSI のサポート 27

XVM のサポート 28

IP アドレスの設定 9

L

L1/L2 コントローラ 33

Linux カーネル・クラッシュ・ダンプ 6

M

Message Passing Toolkit のサポート 28

N

NUMA Tools 33

 dlook コマンド 34

 dplace コマンド 34

O

OS 拡張

 CSA(Comprehensive System Accounting) 22

 CpuMemSets 21

 cpuset のサポート 22

 HPC 環境のための 21

 システム・パーティション 23

OS 構成の設定 1

P

PCP (Performance Co-Pilot) のサポート 29

PROM のサポート 29

R

Rescue モード 14

root パスワード 9

S

SGIconsole のサポート 33

System Controller Software
 CD 13

X

XFS ファイルシステムのサポート 28

XSCSI-SCSI のサポート 27

XSCSI のサポート 27

XVM のサポート 28

い

インストール

 CD 11

 Open/Free Source Software 11

 Proprietary Software 11

 SystemController Software 13

 アップグレード 13

 インストール手法の選択 8、14

 概要 4

 再起動 12

 タイムゾーン 9

 追加ユーザ 9

 パスワード設定 9

 パッケージ・グループの選択 10

 パッケージの選択 12

 ベース OS SGI Advanced Linux Environment 7.2 7

 ユーザの作成 9

ログ 12
SGI ProPack for Linux 11

お

オープン・ソース・サイト 4

か

画面

Complete 12
Installation to Begin 12
Package Group Seleciton 10、12
Package Installation 12
SGI ProPack Welcome 11

き

旧リリースからのアップグレード 10

け

言語の選択 8、9、14

こ

固定命名

Ethernet デバイス 25
XSCSI デバイス 27
バス番号 24
コンパイラ 36

さ

サポートされる製品 17
サポート対象外 20

し

システム管理

EFI 30
FLEXlm 32
PROM チップ 29
SGIconsole 33
システム・コントローラ・ファームウェア 33
ツール 29

システム・コントローラ・ファームウェアのサポート 33
自動パーティション・オプション 8

せ

製品の一覧 17

そ

ソフトウェア

概要 1
計画 5
インストール 5

ソフトウェアのアップグレード 13

ち

チューニング

chatr の使用 44

ddd ツール 38
dplace 48
Electric Fence 39
環境変数 49
偽共有 47
キャッシュ・パフォーマンス 43
コンパイラ・オプション 42
システム構成 35
出力の正確性の確認 37
シングルプロセッサ・コード 36
推奨コンパイラ 36
数値演算関数の利用 41
デバッグ・ツール
 Electric Fence 39
 gbd 37
 ibd 38
ヒープ破壊 39
プロファイリング
 dplace コマンド 51
 GuideView 52
 pfmon 50
 profile.pl スクリプト 51
 VTune アナライザ 52
並列化 45
ボトルネックの検出 47
マルチプロセッサ・コード 45
無効な共有 47
メモリ管理 43

て

ディスク・パーティション表 6

は

パーティションの消去 8
パーティションの選択 8
ハードウェア・プラットフォーム 1
バス固定ナンバリング 24
バス番号 24
パスワードの必要性 5
パッケージ・グループの選択 10、12
パフォーマンスのチューニング 35

ふ

ファイアウォールの設定 9
ファイルシステムの修復 14
ファイル設定と LKCD 6
ファイルの設定 6
ファイルの配置 6

ほ

ボタン
 戻る 7

り

リリースの内容 1