



White Paper

SGI® ICE X

非常に高い柔軟性を備えた世界最速のスーパーコンピュータ

目次

SGI ICE X: どのようなワークロードにも対応可能な究極の柔軟性	3
高密度と電源の柔軟性を目指した設計	4
統合バックプレーンを備えたモジュール型エンクロージャ	4
単独で拡張が可能な電源シェルフ	5
高性能な計算ブレード・オプション	7
ノード、スイッチ、トポロジ・レベルでの FDR InfiniBand インターコネクットの柔軟性	10
FDR InfiniBand メザニンカード・オプション	11
FDR InfiniBand スイッチ・ブレード・オプション	12
InfiniBand トポロジの幅広いオプション	14
スケーラブルなアウトオブバンド管理	15
ノード、およびラック・レベルでの冷却の柔軟性	16
標準的なホットアイル/コールドアイル環境に対応した SGI ICE X D-Rack	16
大規模 HPC システム用の SGI ICE X Cell	17
革新的な SGI コールドシンク技術	19
まとめ	20

SGI ICE X: どのようなワークロードにも対応可能な究極の柔軟性

人々が科学、技術、ビジネス分野における複雑な問題を解明しようとするなかで、現代のテクニカル・コンピューティングにおける課題では非常に高い計算能力が求められています。こうした需要に応えるため、高密度で性能の高い計算ノードを集めたスケールアウト型のクラスタが多くの業界で広く導入されるようになってきました。スーパーコンピューティング・クラスタは今や、スーパーコンピュータ・サイトのTOP500リスト(www.top500.org)を席巻しています。しかし、こうした相互接続されたシステムが効果的であるためには、特に制限なくシステムを拡張する事ができて、施設や電源装置、冷却方法においてしばしば見られる制約にも対応できるシステムである事が必要です。

業界標準のコンポーネントをベースにした大規模なスーパーコンピューティング・クラスタへの流れが目立っているにもかかわらず、ハイパフォーマンス・コンピューティング(HPC)では、万能サイズのソリューションは存在しません。HPCの技術を適用して新しい問題を解決、あるいは古くからの問題を大きなスケールで解決するなかで、システムの柔軟性は全体的な性能と同様の重要性を持つようになってきました。アプリケーションが異なれば、特定の計算に特化した構成、およびインターコネクト・トポロジの必要性が高まります。また物理的環境が異なれば、設置面積、電源装置、冷却方法の面でもそれに適したソリューションが必要です。スーパーコンピュータ・アーキテクチャが効果的であるためには、最新の技術が利用可能であると同時に、インターコネクト、電源装置、冷却方法の観点で最大限の柔軟性が実現されていなければなりません。

SGIは、既存の最大規模のInfiniBandクラスタの導入で長年の実績があり、豊富な専門知識を有しています。なかには自社の制約に基づいたソリューションを提供するベンダーもありますが、SGIは最良のインフラを選択する事で、アプリケーションのニーズや企業のニーズに応えるための柔軟性と最高の性能を提供できる事をよく理解しています。次世代システムSGI ICE Xは、このような考え方を採用しており、世界最大かつ最速のInfiniBand計算クラスタを提供するとともに、どのようなワークロードもこなす究極のスケラビリティと柔軟性を実現しています。

- 次世代インテル® Xeon® プロセッサ E5 ファミリーに基づくSGI ICE Xシステムは、従来世代のSGI ICE 8400システムと比べ、性能密度が約5倍向上しています。
- ペタスケール・クラスであり、エクサスケールの性能までの明確なロードマップを持つ当システムは、数十テラフロップスから数百ペタフロップスまでシームレスなスケラビリティを実現し、しかもそれは業界標準のコンポーネントに基づくシステムで完結します。
- SGI ICE Xは多くの競合他社の製品と異なり、電源を入れればすぐ使える容易さを実現しており、導入には数週間、数カ月ではなく、数時間、数日単位しかかかりません。このため、クラスタにダウンタイムを起さずにアップグレード、および拡張することが可能です。

本ホワイトペーパーでは、柔軟性を提供するアーキテクチャのイノベーションに特に重点を置いて、電源装置、冷却方法、FDR InfiniBandインターコネク、トポロジの観点からSGI ICE Xシステムの機能を説明します。

高密度と電源の柔軟性を目指した設計

SGI ICE Xシステムは、同一の技術世代および異なる世代間の拡張性を考慮して設計されています。SGI ICE Xでは、シャーシにとりわけ注意を払い、電源の柔軟性と高性能計算ブレードの選択を可能にする統合バックプレーンを備えたモジュール型シャーシを設計しました。

統合バックプレーンを備えたモジュール型エンクロージャ

図 1 に示す通り、SGI ICE X は 9.5 ラック・ユニット(9.5U)モジュール型エンクロージャをベースにしています。従来世代のシステムと違い、このエンクロージャでは、すべてのシステム構成で単一の統合バックプレーンが提供されており、個々の計算ブレード・スロットに電力が供給されるとともに、計算ブレード・スロットが FDR InfiniBand スイッチ・ブレード・スロットに接続されています。

エンクロージャはモジュール型のブロックとして 2 つ一組(エンクロージャ・ペア)で提供されます。エンクロージャ・ペアには、下記のものがあります。

- ブレード・スロット 36 個(次世代インテル® Xeon® プロセッサ E5 ファミリーを搭載したシングルまたはデュアルノードの SGI ICE X 計算ブレードを搭載)
- FDR(Fourteen Data Rate) InfiniBand スイッチ・ブレード向けスロット 4 個
- シャーシ管理コントローラ・スロット 4 個

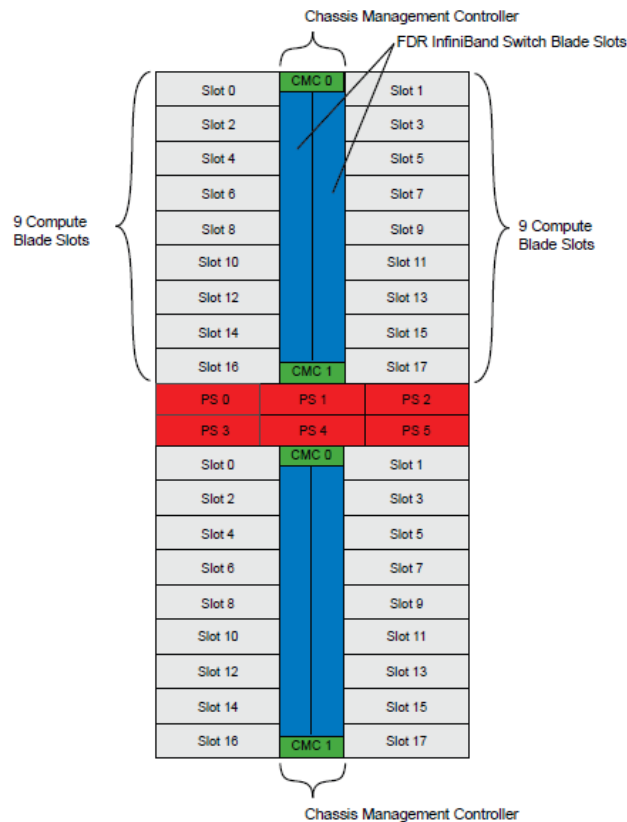


図 1. それぞれの SGI ICE X エンクロージャは、FDR InfiniBand を介して 36 個のブレード・スロット(最大 72 個の計算ブレード)と接続

単独で拡張が可能な電源シェルフ

ブレード・シャーシのパワーサプライは通常、特定のコンポーネントの電源ニーズを満たすとともに、個々の電力供給に障害が起きた場合にも継続的な運用ができるキャパシティや冗長性を備えるよう設計されています。このようなアプローチは一般的ですが、とりわけ大規模 HPC においてシステムが冗長性を持って導入される場合には不利な点や制約があります。

- シャーシの電力供給能力により、導入できるコンポーネントやプロセッサ技術が最終的に制約を受けます。例えば、新世代のプロセッサがリリースされても、より高性能なそれらのプロセッサを搭載したブレードがサポートできない可能性があり、シャーシまたはラックのレベルでのアップグレードを強いられます。
- パワーサプライは通常、冗長性のある形態で構成されます。高可用性のためにはエンクロージャ・レベルの冗長性が必要であるものの、そうすると相当な数のパワーサプライの

導入が必要となります。またN+1の冗長性で構成した複数のブレード・シャーシをラック毎に搭載する場合は、余分な(使用されない)パワーサプライが搭載される事になります。

- 余分なパワーサプライは、コンポーネントに電力を供給するように作動していても電力を使い続けるため、電力が浪費され余計な熱を発生させる事になります。

ブレード・シャーシとは独立して拡張可能な SGI ICE X の電源シェルフは、エンクロージャに電力供給されるパワーサプライの数を増やすと同時に、冗長性に必要な予備のパワーサプライの数を減らすことによって、こうした問題に対処しています。複数のエンクロージャが単独で拡張できる電源シェルフを共有し、接続されたパワーサプライはすべて、共通の 12V 電力バスに通電します。このシェルフにより SGI は、ラックの構成や導入コンポーネントの電力要求に基づいた大きな柔軟性を実現しています。その結果、エンクロージャはパワーサプライと無駄を低減して、いっそうの利用可能電力を得ることができます。例えば 5+1 のパワーサプライは、二重の 2+1 構成に比べて大量の利用可能電力を供給できる上、予備のパワーサプライを半分に減らせます。こうした規模による経済性は、大規模構成になればなるほど向上します。

この革新的なアプローチにより、利用可能電力を柔軟に拡大してノードごとに要求される電力レベルに対応するとともに、より強力な電力を必要とするであろう将来世代の技術のサポートも可能にします。SGI ICE X はノードあたり 400W から 1,440W まで拡大できるよう設計されており、新たな電源テクノロジーが登場した際には、電源シェルフを単独でアップグレードできます。電力をあまり必要としない場合は、設置済コンポーネントへの電力供給に必要なだけの電源シェルフを装着すれば済みます。

図 2 はそれぞれ、SGI ICE X D-Rack および M-Rack 向けの電源シェルフ構成を示しています。D-Rack では、9.5 ラック・ユニット(9.5U)のラック・マウント・エンクロージャ 2 台が、単独で拡張可能な共有の電源シェルフ 2 個と組み合わせられており、この電源シェルフは共有の 12V 電力バスを介して両方のエンクロージャに電力を供給します。これとは対照的に、M-Rack では 4 個の電力シェルフ(最大 12 個のパワーサプライ)が、エンクロージャの両側に取り付けられた 12V のバス・バーに電力を供給するため、その電力供給能力と柔軟性は相当なものになります。

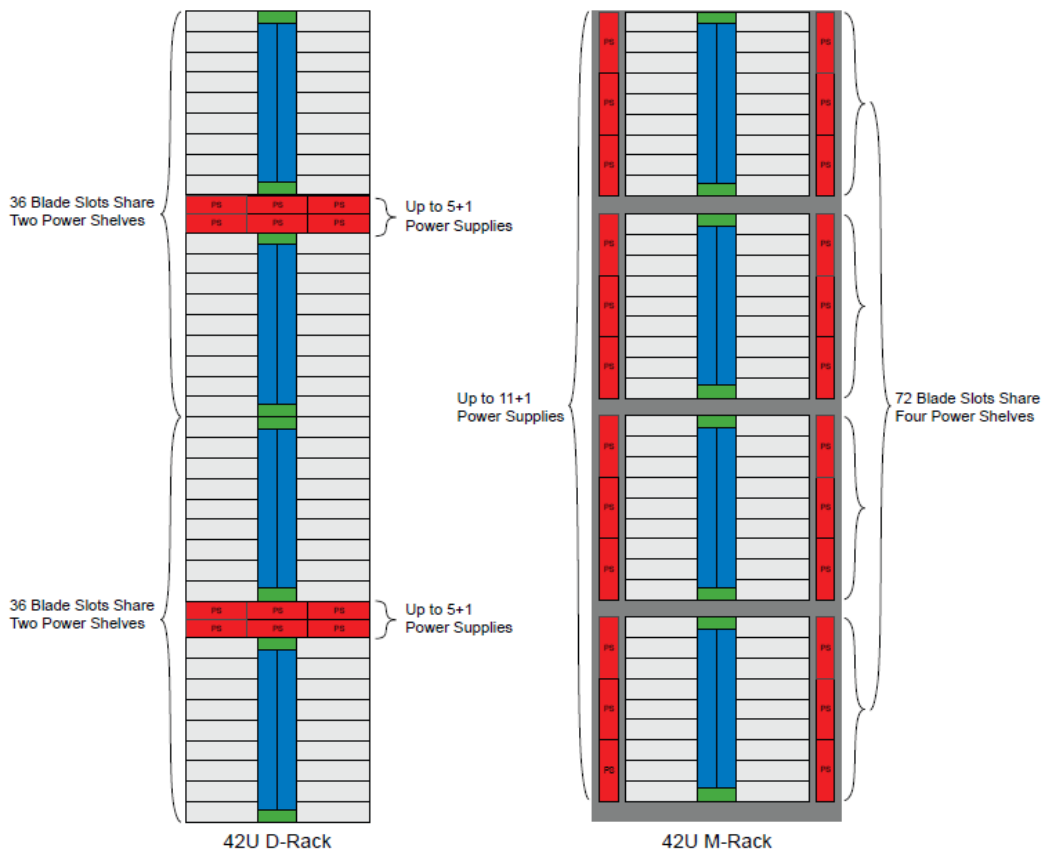


図 2. 単独で拡張が可能な電源シェルフを使い、電力容量とラックの柔軟性を向上

高性能な計算ブレード・オプション

アプリケーションの幅広いニーズに対応するため、SGI ICE X は次世代インテル® Xeon® プロセッサー E5 ファミリー搭載の計算ブレードをサポートしており、幅広いアプリケーションのニーズに対応可能です。インテル® Xeon® プロセッサー E5 ファミリーでは、インテル® Advanced Vector Extensions、インテル® Trusted Execution Technology、およびインテル® AES New Instructions をサポートしています。これらの高性能な計算ブレードはノードごとに 2 個のソケットを提供し、各ソケットは 8 コア、16 スレッドをサポートします。

最大のインターコネクト・バンド幅を提供するため、SGI はインテルその他の企業と緊密に協力し、ICE X に業界のトップの性能を持つ FDR InfiniBand 機能を実装しました。SGI ICE X では従来世代の SGI ICE システムとは異なり、InfiniBand 機能をメザニンカードオプションとして提供します。計算ブレードがサポートする InfiniBand コントローラと冷却方法の選択肢については後述します。

SGI ICE X IP-113 シングルノード計算ブレード

SGI ICE X IP-113 計算ブレードは標準の 19 インチ・ラックマウント向けに設計されています。図 3 に示すように、この計算ブレードは最新の高性能インテル® Xeon® プロセッサを採用しており、仕様は下記の通りです。

- 次世代インテル® Xeon® プロセッサ E5 ファミリー向けのソケット 2 個
- CPU ソケットあたり 8 個の DIMM ソケット(ソケットあたり 4 個または 8 個の DIMM を装着し、1600 MT/s を提供)
- 各ノードで 2 個の 2.5 インチ SATA HDD または SSD をサポート

計算ノード上のインターフェースには FDR InfiniBand メザニンカードも選択可能です。これには 2 個の PCI Express3.0 x8 接続が備わっており、それぞれの接続が、2 つある CPU ソケットの片方に通じています。FDR InfiniBand の選択肢については後述します。

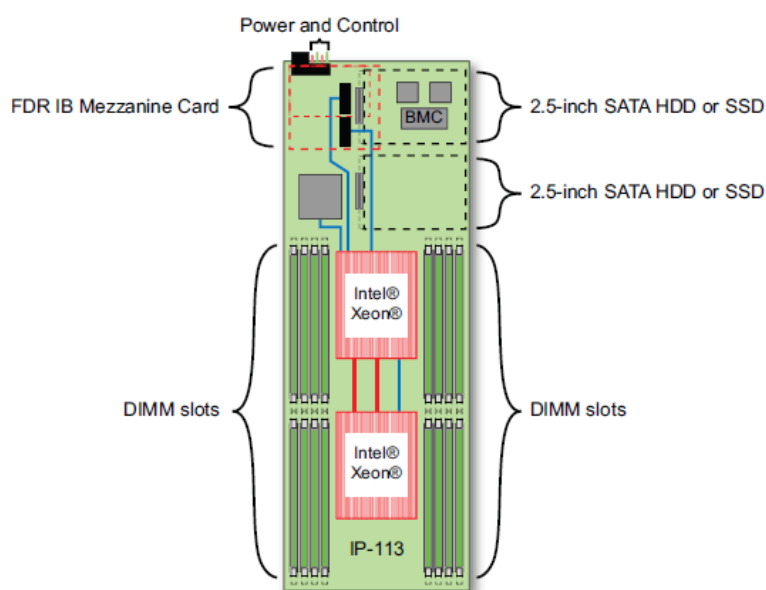


図 3. SGI ICE X IP-113 計算ブレードの概略図

各エンクロージャには最大 18 個の SGI ICE X IP-113 計算ノードが設置され、SGI ICE D-Rack のラックあたりの最大構成は下記の通りです。

- 72 の計算ノード
- 144 個の次世代インテル® Xeon® プロセッサ E5 ファミリー
- 1,152 コアおよび 2,304 スレッド

SGI ICE X IP-115 デュアルノード計算ブレード

SGI ICE X IP-115 デュアルノード計算ブレードは、M-Rack、M-Cell を使用した高密度、かつ特殊用途の HPC 向けに設計されています。図 4 に示すように、この計算ブレードには、2 つの計算ノードが格納されます。このデュアルノード構成により、物理的スペースはそのままに、IP-113 ブレードに比べて計算密度を効果的に 2 倍にすることができます。

SGI ICE X IP-115 各計算ブレードの仕様は下記の通りです。

- ブレードあたり、デュアルノード構成
- 次世代 Intel® Xeon® プロセッサ E5 ファミリー向けソケット合計 4 個 (1 ノードにつき 2 個)
- メザニンカードを介して統合バックプレーンにつながる、2 つのシングルポート FDR InfiniBand 接続 (1 ノードにつき 1 つのポート)
- 計算ノードあたり 8 本の DDR3-1600 DIMM (ソケットあたり 4 本、バスあたり 1 本)

各計算ノード上のインターフェースは 2 つのシングルポートを持つ InfiniBand メザニンカードにつながっており、各 FDR InfiniBand HCA は各ノードのプロセッサの 1 つに PCI Express 3.0 x8 で接続されています。

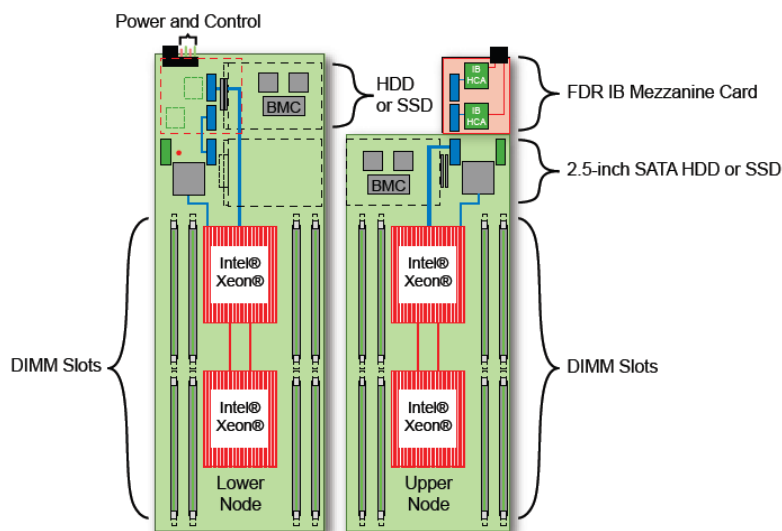


図 4. SGI ICE X IP-115 計算ブレードの概略図

SGI ICE X IP-115 計算ブレードの構成は、プロセッサの消費電力、熱出力、その他導入の際の考慮事項などにより、従来型のヒートシンクにするか、オプションの SGI コールドシンクにするか選

択可能です。SGI コールドシンク技術では、図 5 で示すように、ヒートシンクのかわりに上下ブレードのプロセッサ間に置かれた液冷式の冷却プレートを使用します。構成にもよりますが、SGI コールドシンク技術により、個々のデュアルソケット・ノードが放散する熱の約 50~60%が直接吸収されます。残りの熱は、M-Cell におけるクローズド・ループ方式で空冷します(後述の冷却方法の項で説明します)。



図 5. SGI コールドシンク技術では、より冷却が必要なプロセッサに対して従来型のヒートシンクのかわりに液冷式の冷却プレートを使用します。

ノード、スイッチ、トポロジ・レベルでの FDR InfiniBand インターコネク の柔軟性

企業にとっては、ベンダー側の制約に基づいた決して理想的ではない構成を採用するのではなく、トポロジを柔軟に選択できる事が理想です。アプリケーションに適したシステム構成を選択するにあたり、インターコネクの柔軟性はとりわけ重要です。アプリケーションが異なれば必要とされるバンド幅も異なり、トポロジの選択も変わります。トポロジが柔軟に選択できる事により、適切なコストで性能ニーズを満たすインフラストラクチャを構築する事が可能となります。

- 小規模クラスタまたは左程バンド幅を必要としないアプリケーションは、ファブリック接続をあまり多く必要としないため、ニーズに見合った手頃なソリューションが提供されなくてはなりません。
- 大規模なトポロジまたは要求の厳しいアプリケーションでは、メッセージ・パッシングの負荷を分散させるため、あるいはメッセージ・パッシングを I/O から切り離すために、接続に冗長性を持たせる必要があります。

SGI ICE X は、ノード、スイッチ、トポロジ・レベルで幅広い選択ができるよう設計されており、非常に柔軟なインターコネクをデザインできます。

FDR InfiniBand メザニンカード・オプション

SGI ICE X システムに不可欠な FDR (Fourteen Data Rate) InfiniBand は、IBTA (InfiniBand Trade Association) が開発および仕様決定した次世代の InfiniBand 技術です。FDR InfiniBand は 1 レーンあたり 14 Gbps のデータ転送速度を実現しており、これと肩を並べることができるのは、他社の多くがサポートしている 1 レーンあたり 10 Gbps のデータ転送速度を持つ QDR (Quad Data Rate) InfiniBand だけです(注意 1)。計算ノードそのものに InfiniBand が含まれる従来世代の SGI ICE と違い、SGI ICE X では FDR InfiniBand メザニンカードを選択できるようにし、幅広いアプリケーションのバンド幅の要求に対応しています。サポートされる FDR InfiniBand メザニンカードは表 1 と図 6 で紹介しています。

(注意 1: QDR、FDR IB の実効転送レートは、それぞれ 8Gbps と 13.6Gbps です。QDR は、8b/10b 変換をするので、80%の効率になりますが、FDR は 64b/66b 変換を用いるので、97%の効率があります。)

- シングルポートの FDR InfiniBand メザニンカードはシングルポートの FDR InfiniBand HCA を提供し、SGI ICE X IP-113 計算ブレードのプロセッサが提供する PCI Express3.0 x8 インターフェースに接続しています。
- デュアルポートの FDR InfiniBand メザニンカードはデュアルポートの FDR InfiniBand HCA を提供し、SGI ICE X IP-113 計算ブレードのプロセッサが提供する PCI Express3.0 x8 インターフェースに接続しています。
- 2 つのシングルポートを持つ FDR InfiniBand メザニンカードはシングルポートの FDR InfiniBand HCA を 2 つ提供し、それぞれが別の PCI Express3.0 x8 インターフェースに接続しています。SGI ICE X IP-113 計算ブレードでは、2 つのシングルポートの HCA のそれぞれが 1 つの Intel® Xeon® プロセッサ・ソケットに接続しています。SGI ICE X IP-115 計算ブレードでは、シングルポートの HCA のそれぞれが、2 つある計算ノードの 1 つに接続しています。

表 1. SGI ICE X ブレードでサポートされる FDR InfiniBand メザニンカード

SGI ICE X 計算ブレード	シングルポート FDR IB カード	デュアルポート FDR IB カード	2 つのシングル ポートを持つ FDR IB カード
SGI ICE X IP-113 シングルノード 計算ブレード	X	X	X
SGI ICE X IP-115 デュアルノード 計算ブレード			X

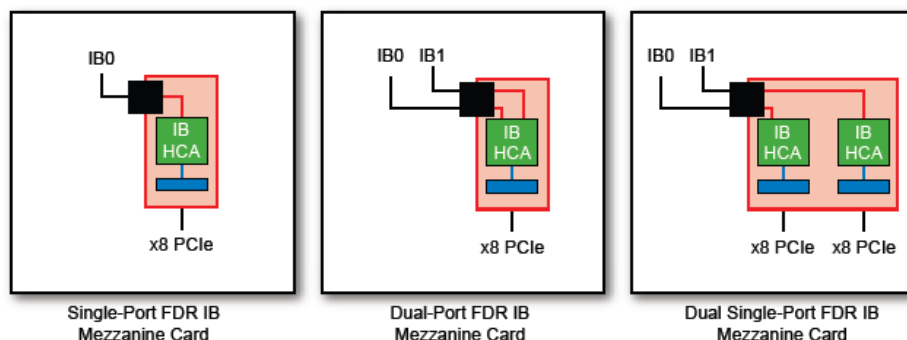


図 6. SGI ICE X システムでは、アプリケーションやトポロジのニーズに応じて FDR InfiniBand メザニンカードを選択できます。

FDR InfiniBand スイッチ・ブレード・オプション

トポロジおよびシステム導入について幅広い選択肢を提供するため、SGI ICE X では FDR InfiniBand スイッチ・ブレードを選択できるようになっています。2 個のスイッチ・ブレードが各エンクロージャの正面に差し込まれて、統合バックプレーンを介して 18 個の計算ブレード・スロットに接続しています。スイッチ・ブレードでは、FDR InfiniBand スイッチ・チップおよび、他のシャーシとの接続に使われる外部 QSFP (Quad Small Form Factor Pluggable) ポートを多様に構成できます。

SGI ICE X の Standard FDR IB スイッチ・ブレード

SGI ICE X IB Standard スイッチ・ブレードは Fat Tree、および All-to-All のトポロジを構築するのに最適で、Hypercube または小規模な Enhanced Hypercube トポロジの構築にも利用できます。図 7 に示す通り、Standard ブレードは統合バックプレーンに接続された 18 個のポートと、外部バルクヘッドに接続された 18 個のポート (スイッチ・ブレードがエンクロージャに差し込まれたときにアクセス可能) を備えるシングルの 36 ポート Mellanox® FDR InfiniBand スイッチ ASIC を提供します。システムがシングル・プレーンかデュアル・プレーンかによって、最大 2 個のスイッチ・ブレードがエンクロージャに装着可能です。2 個のスイッチ・ブレードが装着された場合は、各ブレードはエンクロージャの計算ブレード・スロットにある 2 つの接続のどちらかに接続します。

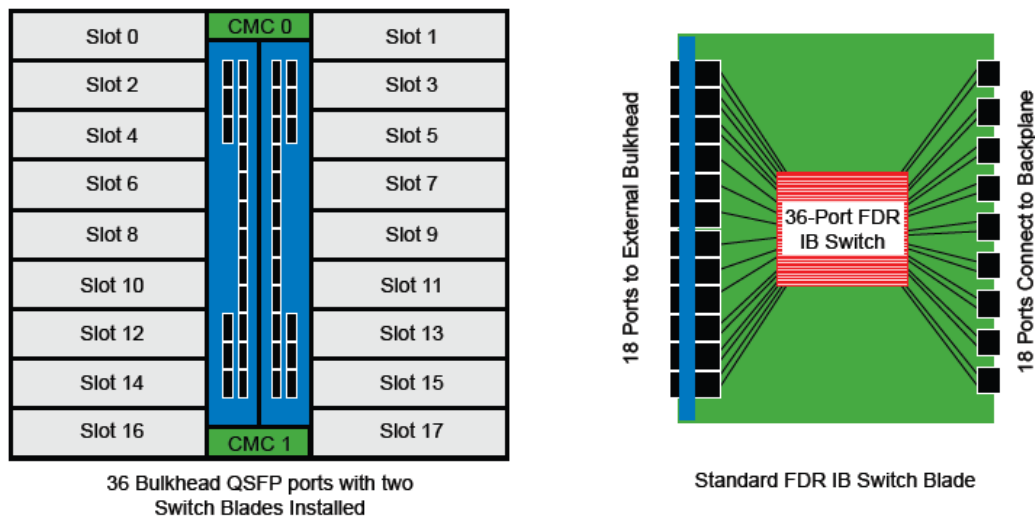


図 7. SGI ICE X IB Standard スイッチ・ブレードは、18 個のエンクロージャ・スロットと 1 つの 36 ポート Mellanox FDR InfiniBand スイッチ・チップを接続し、バルクヘッド上で外部接続用に 18 個の QSFP ポートを提供します。

SGI ICE X の Premium FDR IB スイッチ・ブレード

SGI ICE X IB Premium スイッチ・ブレードは、高バンド幅の All-to-All トポロジや、大規模な Enhanced Hypercube トポロジの構築に最適です。図 8 に示す通り、プレミアム・ブレードは、デュアル 36 ポートの Mellanox FDR InfiniBand スイッチ ASIC を介して、相互接続を提供します。各 ASIC の接続は下記の通りです。

- 統合バックプレーンに接続した各スイッチ・チップの 9 ポート(計 18 ポート)が、18 個の計算ノード・スロットに接続
- 各チップの 3 ポートがチップ間を接続
- 各スイッチ・チップの 24 ポート(計 48 ポート)が、外部バルクヘッドに接続

導入の条件により、最大 2 個のスイッチ・ブレードが取り付け可能です。2 個のスイッチ・ブレードが取り付けられた場合、各ブレードは、エンクロージャのそれぞれの計算ブレードにある 2 個のポートのどちらかに接続します。

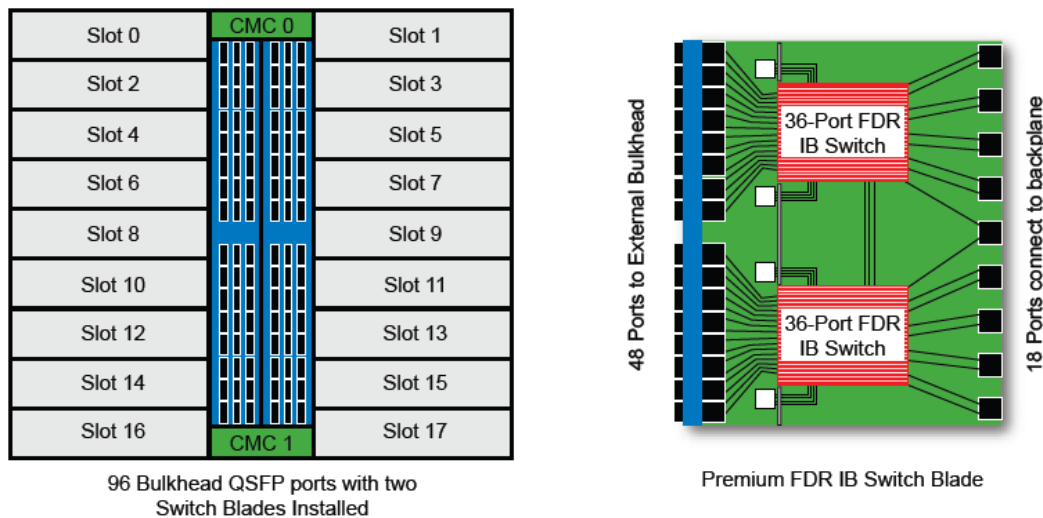


図 8. SGI ICE X IB Premium スイッチ・ブレードは 18 個のエンクロージャ・スロットと 2 個の 36 ポート Mellanox FDR InfiniBand スイッチ・チップを接続し、バルクヘッド上で外部接続用に 48 個の QSFP ポートを提供します。

InfiniBand トポロジの幅広いオプション

SGI には、非常に大規模な InfiniBand クラスタを設計、導入した豊富な経験があります。当社はその専門知識を活用し、最適化された柔軟な InfiniBand トポロジ構成のために設計されたシステムを提供しています。アプリケーションの要件と、低レイテンシ、高いバンド幅のニーズに幅広く応えるため、SGI ICE X は下記のような複数の InfiniBand トポロジの選択肢をサポートしています。

- All-to-All:** All-to-All トポロジは、ホップ数の点で遅延が最も少ないため、MPI (Message Passing Interface) のレイテンシに大きな影響を受けるアプリケーションに最適です。All-to-All トポロジはノン・ブロッキングのファブリックと高いバイセクション・バンド幅を提供できますが、スイッチ・ポート数が限られるため、比較的小規模なクラスタに限定されています。
- Fat Tree:** Fat Tree トポロジは、ノード数が少ない MPI ジョブに適しています。Fat Tree トポロジはノン・ブロッキングのファブリックと一貫したホップ数を提供でき、MPI ジョブのレイテンシが予想可能なものになります。しかしながら、クラスタの規模が大きくなればなるほど、非常に大きなコアのスイッチが必要になり、ケーブルリングとスイッチングが困難になるとともにコストも高くなります。
- Standard Hypercube:** Standard Hypercube トポロジはノード数が多い MPI ジョブに適しており、高いバンド幅を提供します。また、小規模なクラスタから非常に大規模なものまで拡張が容易です。Hypercube はシステム規模が大規模になっても、クラスタ内のローカ

ルおよびグローバル通信が容易に最適化できます。Standard Hypercube トポロジは、それぞれの次元のリンクは 1 本のケーブルで接続されるため、非常に軽量のファブリックを最少のコストで提供することができます。

- **SGI Enhanced Hypercube:** Standard Hypercube トポロジの利点に加えて、SGI Enhanced Hypercube トポロジは、ハイパーキューブの低次元において複数のケーブルによる冗長リンクを追加して利用可能なスイッチ・ポートを活用する事で、インターコネクタ全体のバンド幅を向上させます。

トラス・トポロジは、現在 OFED (Open Fabrics Enterprise Distribution) に入っておらず、拡張性の問題や、クラスタ規模が増大するに従いホップ数やレイテンシが発生するため、現時点では SGI でサポートしていません。Hypercube トポロジはスケーラビリティと共に、トラス・トポロジの利点の多くを提供します。

スケーラブルなアウトオブバンド管理

計算環境、計算リソース、オペレーティング・システム、および全般的なエコシステムの大量データを解析することは、スーパーコンピューティング・クラスタなどの大規模リソースの維持と管理において非常に重要です。効果的運用を行うには、スケジューリング、電源効率、障害やコストの入念な調整が必須になります。SGI ICE X では、システム管理者にエラー状況やシステムの状態に関する適切かつ簡潔なデータを提供する SGI Management Suite のサブシステムである Failure Analysis や Power Management によって、こうしたモニタリング活動が容易になりました。障害解析のためのパネルや計測パネルにグラフィックで表示されるこのデータは、ハードウェアのモニター、修正、隔離、交換に活用されます。

SGI ICE X は、階層型ギガビット・イーサネット・ネットワークを介して、こうした管理業務のサポートを提供します。このアウトオブバンド・ネットワークは従来型の TCP/IP およびイーサネットのネットワーク・モデルを活用しており、InfiniBand ネットワークに負荷がかかる事はありません。階層型アーキテクチャにより、非常に大規模のスーパーコンピュータ・クラスタであっても非常に効率よくシステムを管理できます。

- **ブレード管理コントローラ (BMC: Blade Management Controller):** 各計算ノード上の BMC はボード・レベルでハードウェアを制御するとともに、計算ノード環境をモニターします。
- **シャーシ管理コントローラ (CMC: Chassis Management Controller):** 各 BMC は 1 対のブレード・エンクロージャあたり最大 2 つまたは 4 つの CMC にレポートします。CMC はすべての計算ノードのマスター電源を制御し、電源およびブレード・エンクロージャ環境をモニターします。

- **ラック・レベル・コントローラ(RLC: Rack Level Controller)**: CMC は、2つのブレード・エンクロージャ・ペアに対して提供される RLC に集約されます。RLC はブレードのブート・イメージの保持、ファブリック管理ソフトウェアの実行、ラックのクラスタ管理データの収集を行います。
- **システム管理コントローラ(SAC: System Administration Controller)**: SAC は、各 ICE システムに1つ提供されます。SAC は RLC に対してソフトウェアをプロビジョニングし、RLC からクラスタ管理データを引き出します。

ノード、およびラック・レベルでの冷却の柔軟性

アプリケーションおよびトポロジの要件が様々であるのと同様に、スーパーコンピュータ・クラスタの導入においては様々な物理的制約があります。密度の向上は明らかな目標の1つですが、効率的かつ効果的な冷却方式も、大規模 HPC データセンターには不可欠です。こうしたニーズに対応するため、SGI ICE X は冷却方式に大きな柔軟性を持たせています。

- 従来のホットアイル/コールドアイル空冷システム用の SGI ICE X D-Rack
- クローズド・ループ冷却ソリューション、およびホットアイル・コンテインメント(熱気囲い込み)をサポートする SGI ICE Cell
- ブレード・スロットあたりのノード密度の向上をサポートする SGI コールドシンク技術

標準的なホットアイル/コールドアイル環境に対応した SGI ICE X D-Rack

多くの企業では、24 インチのフロア・タイル、およびホットアイル/コールドアイル環境に適した標準の 19 インチ・ラックのシステム導入が望まれています。このニーズに応えるため、SGI ICE X のエンクロージャは SGI D-Rack にマウントされ、SGI ICE X IP-113 計算ブレードを搭載できる設計になっています。この構成では、それぞれの 42U ラックが2つのエンクロージャ・ペアを搭載でき、それぞれのエンクロージャ・ペアは独立した2個の電源シェルフにつながっています。1本の D-Rack は72個のデュアル・プロセッサ・ブレードを収容でき、従来世代の SGI ICE システムに比べて1.4倍の密度向上が実現しました。

図 9 に示すように、D-Rack はオープン・ループの空冷方式と、オープン・ループ・エアフローを併用するオプションの冷水コイルをサポートします。当製品の設計では、エンクロージャ・レベルの冷却ファンではなく、ラック・レベルの大型で効率の良い冷却ファンを使用しています。オプションの冷水コイルは、シャーシの背面に取り付けるドアとして提供され、ラック内にたまった熱気がデータセンター内に流れ出す前に冷却します。

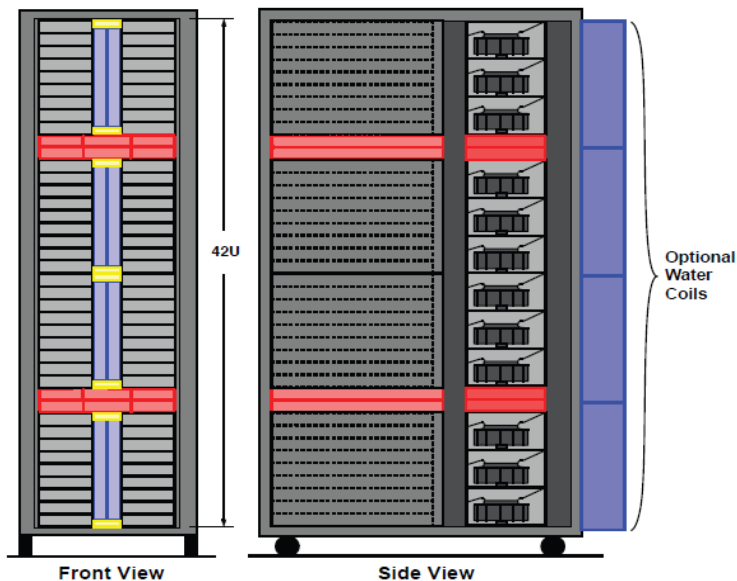


図 9. SGI ICE X IP-113 計算ブレードは、オープン・ループ空冷装置とともに SGI D-Rack 搭載され、オプションのウォーター・コイルとの併用が可能です。

大規模 HPC システム用の SGI ICE X Cell

大規模 HPC システムではしばしば、標準の 19 インチ・ラックでは対応が難しい規模や密度の要件に遭遇します。また、40 フィート ISO コンテナなどのモジュラー式データセンターは、様々な環境での計算能力集約型システム向けに、ますます人気を集めるようになってきました。SGI ICE X Cell は、こうした要求の厳しいシステムに対応するよう設計されており、非常に効率の高いクローズド・ループの冷却ソリューションと、密度の向上を実現します。図 10 に示す通り、1 個の SGI ICE X Cell には計算ラック 4 個と、空冷と水冷を組み合わせたクーリング・ラック 2 個が備わっています。アクセスは、Cell の両側に取り付けられたドアから行います。

SGI ICE X Cell には、下記のような明確な利点があります。

- クローズド・ループ・エアフロー:** 各 Cell は完全に密閉された空間となり、Cell 内の空気はデータセンターの空気と混ざることはありません。Cell 内の空気はすべて水冷により冷却されます。このクローズド・ループ・エアフローはオープン・ループ・システムに比べて振動(アコースティック・エミッション)が軽減できるという利点もあります。
- 温水による冷却:** Cell では、施設から提供される摂氏 7.2~30 度という幅広い温度帯の冷却水をサポートします。温度上限が高いため、年間でみて冷却プラントを稼働させずに冷却水が供給されるフリー・クーリング時間が増えます。この効率性により、非常に高密度のスーパーコンピュータ・クラスであってもコスト削減を図る事が可能です。すべての

Cell で air-to-water の熱交換器が提供され、SGI コールドシンクを利用する際は liquid-to-water の熱交換器が導入されます。

- **統合クーリング・ラック:** スタンドアロン型ラックとは異なり、SGI ICE X Cell に導入されている計算ラックは、それ自体ではラック・レベルの冷却装置を持たず、Cell 内のクーリング・ラックに依存しています。クーリング・ラックは水冷された熱交換器の空気を吸い出し、それを再循環させることで Cell 内の計算ラックを冷却します。この方法は、単一の水源を利用して冷却するのにかかる電力コストを削減できるので、ラック・レベルの冷却に比べて高い効率を得ることができます。

SGI ICE X Cell について、図 10 に平面図を、表 2 に詳細を示しました。2 機種の SGI ICE X Cell は、D-Rack を補完し、導入の様々な要件に対応します。

- **SGI ICE X D-Cell:** D-Cell は、SGI ICE X IP-113 計算ブレードを導入する企業に対し、冷却に関するスケールメリットを提供します。D-Cell は電力コストの削減に役立つほか、温水冷却用共有ソースを利用する場合に必要な付属品を最小限にする事ができます。
- **SGI ICE X M-Cell:** M-Cell は、デュアルノードの SGI ICE X IP-115 計算ブレードの搭載により、D-Cell に比べて計算密度を 2 倍にできます。また M-Cell は、将来の大消費電力テクノロジーへのアップグレードをサポートするよう設計されています。M-Cell 内の M-Rack は、最大 12 個のパワーサプライが 1 個の 12W 電力バス・バーを共有しているため(図 2)、共有のパワーサプライからより多くの電力を提供します。また、SGI コールドシンク技術は高温のコンポーネントをより効率的に冷却するので、より発熱量の多いコンポーネントをサポートする事が可能です。

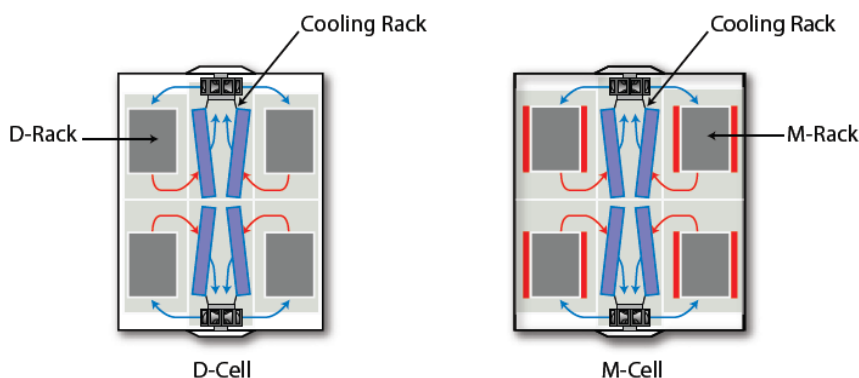


図 10. D-Cell および M-Cell はクローズド・ループ・エアフローと温水冷却を使用し、Cell 内にホットアイル・コンテインメント(熱気の囲い込み)を作り出します(この Cell 図は上から見たもの)。

表 2. SGI ICE X D-Rack、D-Cell、M-Cell の機能

展開オプション	ブレード スロット	設置面積 (平方フィート)	計算ノード	冷却方法	SGI ICE X ブレード
ICE X D-Rack	72	6.67	72	オープン・ループ・エアフローによる空冷、およびオプションで水冷	IP-113 計算ブレード
ICE X D-Cell	288	72	288	クローズド・ループ・エアフローと温水による水冷	IP-113 計算ブレード
ICE X M-Cell	288	80	576	クローズド・ループ・エアフローと温水による水冷、およびオプションで SGI コールドシンク	IP-115 計算ブレード

革新的な SGI コールドシンク技術

高性能な HPC システムの構築および導入を 20 年近くにわたって展開してきた SGI は、効果的な冷却システムの設計に必要なとされる十分な知識を有しています。SGI コールドシンク技術はそうした技術面での優れた強みの 1 つであり、必要な冷却装置を最小限にしつつ、計算密度の向上を促進するために温水冷却インフラを活用可能な洗練された、かつ効率も良い冷却システムです。

IP-115 計算ノードを使った SGI ICE X M-Rack システムでは、クローズド・ループ・エアフローによる冷却に加えて、SGI コールドシンク技術が利用可能です。導入プロセッサの消費電力に応じて、SGI コールドシンク技術では、計算ノード上の従来型ヒートシンクのかわりに上下のノードにあるソケット間に置かれた液冷式の冷却プレートを用います。冷却液は対になった冷却プレートの列を流れ、ブレード・エンクロージャと一体化した液体供給管の上を通ります(図 11)。構成にもよりますが、SGI コールドシンク技術は、個々のデュアル・ソケット・ノードが放出した熱の約 50~65%を直接、吸収できます。残りの熱は M-Cell 内のクローズド・ループ・エアフローによって空冷されません。SGI コールドシンク技術を使用する場合は、air-to-water の熱交換器に加え、liquid-to-water プレート熱交換器と二次ポンプ・ループを備えます。

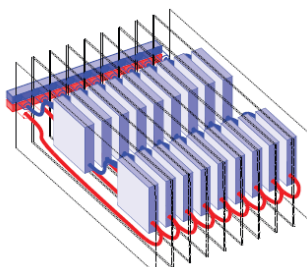


図 11. SGI コールドシンク技術は、M-Cell での高消費電力プロセッサを搭載した SGI ICE X IP-115 計算ブレード用に、プロセッサ間に水冷式冷却プレートを提供します。

まとめ

SGI ICE Xは、当初から非常に広範囲なHPCコンピューティングの要件に対応する、究極の柔軟性を念頭に設計されました。単独で拡張が可能な電源シェルフは将来の技術の進歩に備え、シャーシに対してより多くの利用可能な電力と電源の柔軟性を提供すると同時に、冗長パワーサプライの削減や、コストのかさむ電力浪費の排除を実現しています。革新的なFDR InfiniBandメザニンカードと、スイッチおよびInfiniBandトポロジのオプションは、比類のないインターコネクットの柔軟性を提供しています。冷却ソリューションには空冷および温水水冷が選択可能であり、現在、近い将来、そして未来のシステム・ニーズ、および施設の要件に対応します。


SGI ICE Xでは、次世代インテル® Xeon® プロセッサ E5 ファミリーに基づくデュアル・プロセッサ計算ノードの選択により、優れた計算性能、メモリ容量、スループットが実現するとともに、従来型の19インチ・ラック、およびSGI高密度M-Cellにも対応します。SGI コールドシンク技術をはじめとする革新は、過去に成功を収めてきた設計やHPC導入の経験を活用し、HPCコンピューティングの幅広いニーズに対応するシステムとして結実しています。とりわけ、大規模なSGI ICE Xシステムであっても迅速な導入が可能であり、クラスタが稼動中であっても拡張およびアップグレードをする事が可能です。

(C) 2011 SGI Japan, LTD. All rights reserved.

SGI、Altix、NUMalink、XIOのロゴマークは米Silicon Graphics, Inc./日本SGI 株式会社の登録商標です。Intel および Xeon は Intel コーポレーション、またはその子会社の商標、または登録商標です。その他の商標については商標の所有者に所有権が属しています。

日本SGI株式会社

〒150-6031 東京都渋谷区恵比寿4-20-3 恵比寿ガーデンプレイスタワー31階

 TEL:0120-161-086 FAX:0120-161-087 <http://www.sgi.co.jp>

本 社	TEL : 03-5488-1811(大代表) FAX : 03-5420-7201
西 日 本 支 社	TEL : 06-6479-3918(代表) FAX : 06-6479-3919
中 部 支 社	TEL : 0565-35-2561(代表) FAX : 0565-35-2189
つくば・東北事業所	TEL : 029-858-1551(代表) FAX : 029-858-1071
東 北 営 業 所	TEL : 022-221-2301(代表) FAX : 022-221-2304
北 海 道 営 業 所	TEL : 011-708-1511(代表) FAX : 011-758-2789