



Myri-10G

第4世代のMyrinet と 10Gigabit Ethernet の連携

住商情報システム株式会社
エンジニアリングソリューション事業部
グリッドソリューション部

2006/3/3

PC Cluster Consortium ワークショップ in 関西

Myricom

www.myri.com

© 2006 Myricom, Inc.

SCS



Myrinet Technology Roadmap

- **Myrinet-10G**

- 4th-generation Myrinet, products to be announced in 2Q05
- 10+10 Gb/s data rate links
- Interoperable with 10-Gigabit Ethernet
- Fully compatible with Myrinet-2000
 - Same architecture, same software, same APIs, same applications

- **Myrinet Express (MX) software**

- Message-passing system for initial Myrinet-10G products
- MX-2G-Beta is already available for Myrinet-2000 D, E, & F cards
 - Myricom software support has always spanned two generations of NICs

We expect Myrinet-2000 products to continue to sell well through 2006, and to co-exist with the higher priced Myrinet-10G products. Myrinet-2000 is well positioned as a superior alternative to Gigabit Ethernet for clusters, while Myrinet-10G offers performance and cost advantages over 10GbE for clusters.



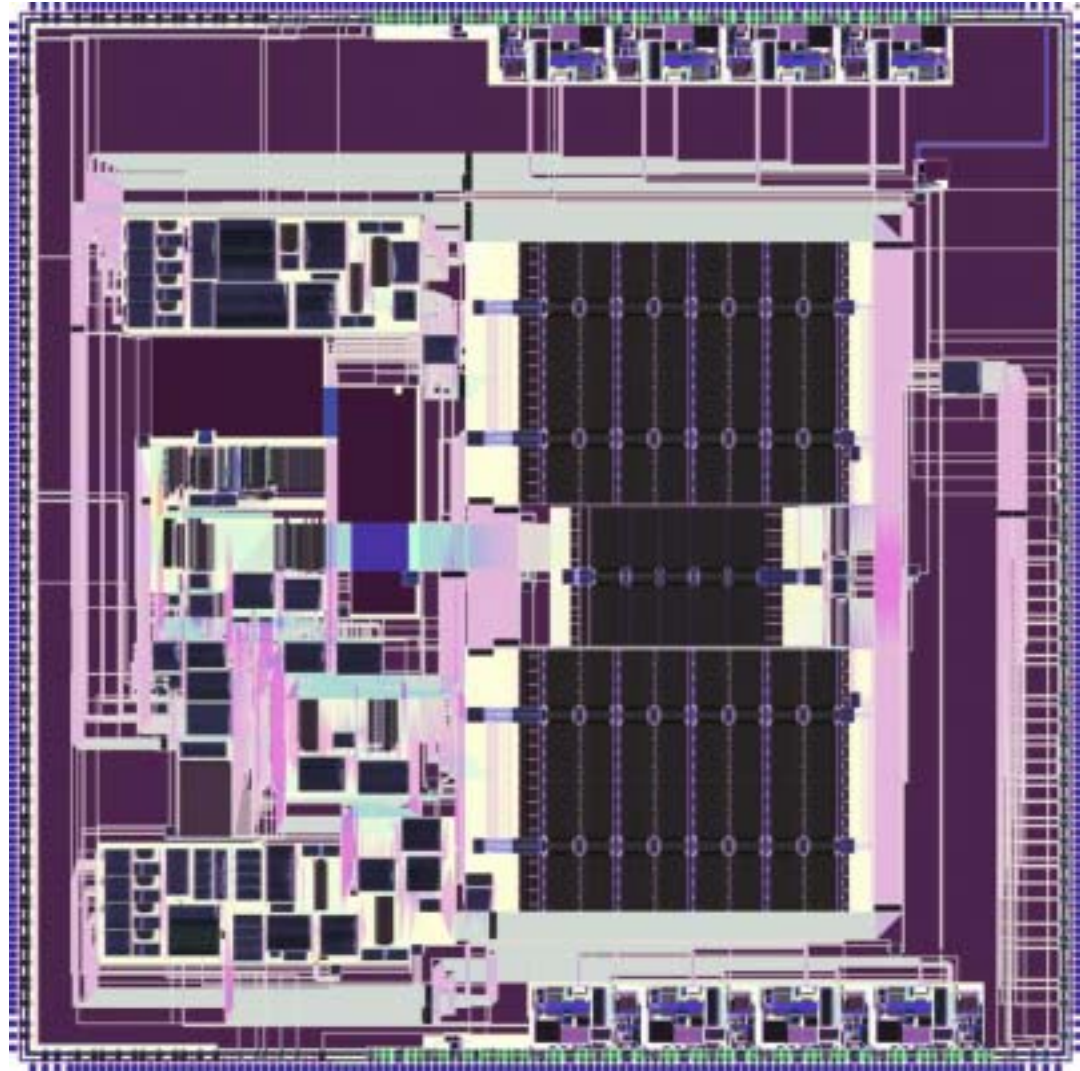
Myrinet-10G Links, NICs, and Switches

- Myrinet-10G links
 - **Any PHY (cables, connectors, signaling) that can be used for 10-Gigabit Ethernet (10GbE) can be used for Myrinet-10G**
 - The ports of Myrinet-10G chips are XAUI, per IEEE 802.3ae
- Myrinet-10G NICs
 - Even more highly integrated than Myrinet-2000 PCI-X NICs
 - PCI Express, HyperTransport, and other fast host buses
 - **Any Myrinet-10G NIC is also a 10GbE NIC**
 - The standard firmware supports both Myrinet-10G and 10GbE modes
- Myrinet-10G switches
 - Similar except for data rates to Myrinet-2000 switches, based initially on a Myrinet-10G 16-port crossbar-switch chip



The Initial 10G-Ethernet/10G-Myrinet NIC

- PCI Express
- Based on the Lanai ZE chip (on right)
 - Internal packet buffers as well as external SRAM
- 333MHz processors and external memory
- Self-initializes to 10GbE mode, but operates also in Myrinet-10G mode
- 9.5 Gbits/s expected TCP/IP rate (GbE mode)
- 2 μ s expected MPI latency with MX-10G, and 1.2 GBytes/s one-way data rate (Myrinet mode)





Myricom

Pioneering Higher Performance Computing

Myri-10G

*10-Gigabit Ethernet
with a Supercomputing Heritage*





Myri-10G ?

- **第4世代の Myricom プロダクト**, 10-Gigabit Ethernet テクノロジをHPC の世界に、HPC のテクノロジーをEthernet の世界に反映
- **10-Gigabit Ethernet プロダクト**
 - ハイパフォーマンス、ローコストでIEEE 802.3ae に適応、他社の10G Ethernet 製品群と親和性を持つ
- **第4世代の Myrinet (HPC)**
 - クラスタインターコネクトソリューションとして低レンシテンシを持つNIC
 - Myrinet 2000のユーザアプリケーションと完全互換
 - 10G Ethernet 物理層 (layer 1)、10 Gbit/s data rates を基本
 - NIC はEthernet と Myrinet プロトコル の両方をデータリンクレベル (layer 2)でサポート
 - スイッチは高い効率性と拡張性を持ち内部的にはlayer-2 のスイッチとして動作、外部に対してはMyrinetとEthernetのポートを持つ

Myri-10G NIC



10GBase-CX4



10GBase-R



XAUI over ribbon fiber

全てのNICは8-laneのPCI-Express スロット対応と Myricom Lanai-Z8E chipを実装

10-Gigabit Ethernet NICもプロトコルオフロードサポート 上記NICとMyricom から提供されるドライバ、10G Ethernet スイッチ(汎用品)を使いnetperf テストで9.6 Gbits/s (TCP/IP data rate、Linux 2.6.9, Opteron hosts)を実現。ジャンボフレームにも対応。

10-Gigabit Myrinet NIC 上記NICとMyricomから提供されるMyrinet Express (MX) ソフトウェア、Myrinet 10G スイッチにより、

- 2 μ s MPI latency
- 1.2 GBytes/s data rate
- 非常に低いCPU 利用率を実現



Myri-10G Software

- 10-Gigabit Ethernet モードのためにNIC のドライバとファームウェアが標準で提供されます。
 - Myri-10G Ethernet ドライバは(将来)OS に含まれます(予定)。
 - 最適化されたMyri-10G の iSCSI ドライバは現在開発中であり、ドライバと同様に(将来)OS に含まれます(予定)。
- Myrinet Express (MX)
 - MX-10G はMyri-10G プロダクトをローレベルのメッセージパッシングシステムとして動作させるソフトウェア
 - 2005年6月にMyrinet 2000 PCI-X NIC のためにMX-2G をリリース
 - Myricom のソフトウェアは常に2世代前の製品までサポート
 - Ethernet のエミュレーション(TCP, UDP/IP)、MPICH-MXとSocket - MX(DAPL-MXとMPICH-2-MXも開発中
 - MX-2G と MX-10G はユーザアプリケーションレベルで互換性を保つ



Myri-10G Switch

- Myri-10G のためのスイッチ
- 物理的には10-Gigabit Ethernet テクノロジーを持ち、高い効率性と拡張性を持つlayer-2スイッチ
- 10-Gigabit Myrinet スイッチはデータレートを除いては Myrinet 2000 スイッチと同様であり、16または32ポートクロスバー・スイッチ・チップを実装します。初期のスイッチは高さ14U でClos 64+64とClos128 をサポートします。
- 10-Gigabit Myrinet スイッチケーブルは10-Gigabit Ethernet ポートと親和性を持ち、シンプル、layer-2、プロトコルコンバージョン機能を有します。
 - 2006年下期にラインカードをリリース予定

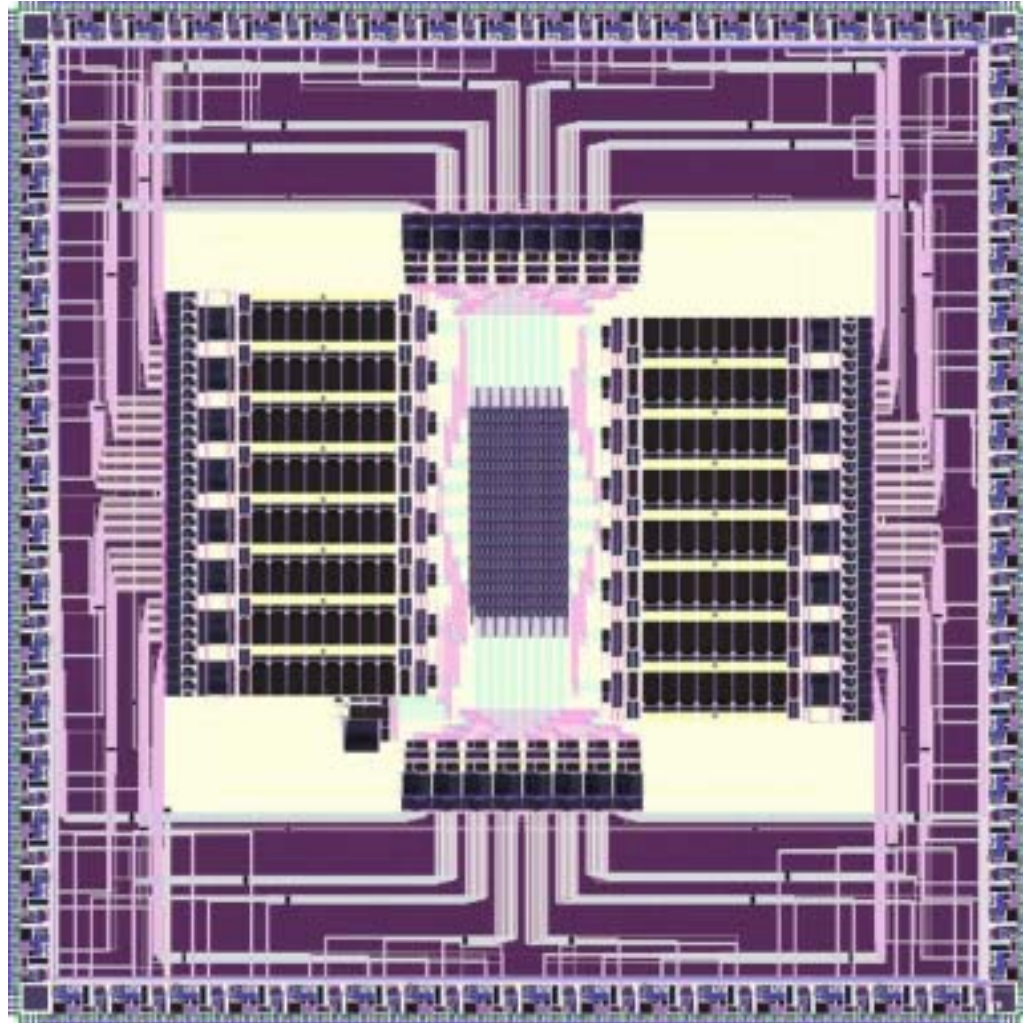


10G Myrinet と 10G Ethernet との比較

- フローコントロール:
 - 10G Myrinet: byte-level フローコントロール
 - 10G Ethernet: packet-level、MAC-layer-PAUSE フローコントロール
- スイッチングルーティング:
 - 10G Myrinet: source-based ルーティング、cut-through
 - 10G Ethernet: destination-based ルーティング、store-and-forward
- 拡張性:
 - 10G Myrinet: 事実上無制限に近い。Clos トポロジを採用
 - 10G Ethernet: 全体のケーブル長で制限あり。複数台のスイッチ構成ではスイッチ間でLayer-3のIPルーティングを利用
- Discovery:
 - 10G Myrinet: あるホストは他のホストをmapper により認識
 - 10G Ethernet: スイッチがネットワークトラフィックを見ることでホストのロケーションを認識

10 Gigabit Myrinet Switch Chip

- 10G_Xbar-16 chip
 - 16 XAUIポート
 - 消費電力 ~15w
- Myrinet 2000 用のスイッチチップと同様のパッケージング
- Myrinet 2000 X-Bar32 と同様のチップも開発予定



Myri-10G Switch 製品例

- 128ホストまで対応のClosネットワーク
- 64ホストポートと相互接続のための64リーフスイッチ構成64
- 物理層を混在した構成
- エンタープライズのための機能
 - 冗長性のためのホットスワップ対応の電源とファン
 - ホットスワップ対応のラインカード
 - ネットワーク状態の監視とファンクション用の二つの10/100 Ethernet ポートとモニタリング用のTFTディスプレイ
- より高密度実装のスイッチが2006年後半にリリース予定





Myri-10G の最新状況

- 10-Gigabit Ethernet アプリケーションのためにNICを2005年12月よりリリース
 - http://www.myri.com/Myri-10G/10gbe_solutions.html にパフォーマンスと規格をアップ中
- Myri-10GのためのNICとスイッチは2006年2月より順次出荷中
 - MX-10G ソフトウェア使用時に於いて2 μ s MPI レイテンシと1.2 GByte/s MPIデータレート
- 2006年2月8日にIBM Blade Center HのためのMyri-10G high-speed daughter card (HSDC) と switch modules (HSSM) を発表
 - Myrinet 2000と同様の協業



Myri-10G まとめ

- 圧倒的なパフォーマンス
 - 9.6 Gigabits/s (netperf) TCP/IP (Ethernet モード)
 - 2 μ s MPI レイテンシと1.2 GBytes/s スループット (Myrinet mode)
- 各種H/Wに対応したソフトウェア環境
 - Linux, Windows, Solaris, AIX, Mac OS X, FreeBSD
- 業界標準の10-Gigabit Ethernet ケーブルを採用
 - copper と fiber
- 10-Gigabit EthernetとMyrinet モードでの親和性
- 拡張性に優れ低価格、低レイテンシのスイッチ
- 標準規格
 - IEEE 802.3ae (10GbE) と ANSI-VITA 26-1998 (Myrinet)
- 常に Price / performance リーダ