2K-4

中小規模並列コンピュータ Ships1 の開発

†松尾成志　†岡本恵介　†大谷真
湘南工科大学　情報工学科

１．はじめに

並列計算機技術はこれまで大規模複雑な科学技術計算の高速処理を主目的に研究開発がなされてきた。一方、プロセッサとメモリの高速低価格化、オープンソースを中心としたOSやミドルウェア、ソフトウェアの低価格化または無償化により、並列計算機技術を低価格な小規模システムに応用することが可能になってきている

Ships1(Shonan Institute of Technology Parallel System 1)は、これを実現する目的で開発中の中小規模並列コンピュータである。本論文では、Ships1 の基本部分の設計と開発について全体アーキテクチャおよびノード設計、ノード透過性について中心に述べる。

２．**Ships1** のねらい

これまでの並列計算機技術は高価な計算資源を利用して大型計算を高速で処理することに力を注がれてきた。一方、通常の市場で購入できるCPUやメモリなどのハードウェアの対価格性能比が飛躍的に向上している。また、ソフトウェアも市場の拡大やオープンソース化の進展により対価格機能比が向上している。これらを組み合わせることにより、低価格の並列機を開発できる可能性が生まれている。Ships1 の研究開発の目的は中小規模向けの低価格並列コンピュータの実現である。専用部品や専用ソフトの開発を必要最小限に抑えると同時に中小規模計算に十分使用可能な性能を確保することを開発方針としている。Ships1 は１６個のノードを持つ。また、基本的なハードウェアとソフトウェアに加え、各ノードを使う時にどのノードを使ってもほぼ同一の実行環境が得られる機能(これをノード透過性という)を準備することとした。以降、まずハードウェア全体のアーキテクチャをいかにしたかを述べ、続いて各ノードのハードウェアとOSの設計について記し、最後にノード透過性の実現方法を述べる

３．**Ships1** のアーキテクチャ

Ships1 は図１に示すようノード、管理マシン、主ネットワーク、管理ネットワーク、ノード間通信装置で構成させることとした。

（１）ノード

実際の計算を行う部分で全ての点でPCと同等の機能を持つ。ノードには２種類ありαノードとβノードと呼んでいる。αノードは計算を行う通常のノードである。Bノードは計算の他にノード間通信装置による高速なノード間接続をするノードである。

（２）管理マシン

Ships1 の全ノードの運転を管理･制御するコンピュータである。機能としては各ノードの起動、シャットダウン、動作監視などである。インターネットを介して遠隔から運転制御を可能にした。そのため、セキュリティの視点からShips1から独立した一個のコンピュータとして後述の主ネットワークとは分離した構成とした。

（３）ネットワーク環境

Ships1 のノードは２種類のLANで接続される。それぞれ、主ネットワークと管理ネットワークと呼ぶ。両ネットワークとも1Gbpsとし、専用部品や専用ソフトを必要最小限に抑えるというコンセプトに基づき通常のTCP/IPプロトコルで接続し、両ネットワークは完全に分離したネットワークとした。

・主ネットワーク

Ship1 内の全てのノードを接続する高速のネットワークである。性能を要求されるためPCI Expressを経由して接続とした。

・管理ネットワーク

各ノードの制御と管理のためにノード及び管理マシンを接続するためのものである。

（４）ノード間通信装置

βノードにはPCIバスにノード間通信装置が装着されており、２つのβノード間ではより高速なデータ交換を可能としている。ノード間通信装置のネットワークはポイントツーポイントとしている。

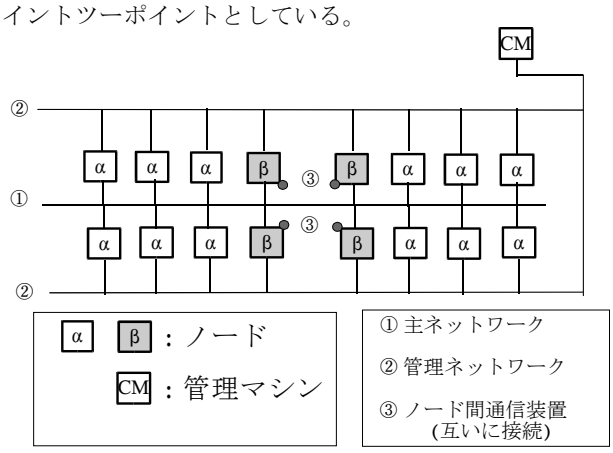

図１ Ships1 の構成

Development of a Small-Mid Range Parallel Computer - Ships1
† Seiji Matsuo , Keisuke Okamoto , Makoto Oya – Shonan Institute of Tecnology

４．ノード設計

PC の部品として通常に広く販売されているものの中からサーバ機に十分かつ比較的廉価なものを調達した。具体的には以下の方針でノードの構成を決めた。表１に開発したノードの概略仕様を示す。

・CPU：対価格性能比などから Intel Pentium 4 3.0GHz を採用した。

・メインボード：チップセットについては PCI Express があり、グラフィック機能が内蔵されている点から Intel 945G とした。また遠隔電源の投入の実装のための WOL 機能を持っていることを条件とした。更に、ネットワーク接続のために、αノードでは１個以上、βノードでは４個以上の PCI スロットを持つことも条件とした。βノードには管理ネットワークの他に３つのノード間通信装置を装備できることが必要だからである。

・メモリ：βノードは負荷が高いことが予想されるため２Ｇバイトとし、αノードを１Ｇバイトとした。メモリの種類は高性能でかつチップセットが Intel 945G なので DDR2 となった。

表１ノード概略仕様

	αノード	βノード
CPU	Intel Pentium 4 (H/T) 3.0GHz	Intel Pentium 4 (H/T) 3.0GHz
チップセット	Intel 945G + ICH7	Intel 945G + ICH7
M/B	GIGABYTE 社　GA-8I945G	MSI 社　945-GNeo2-F
メモリ	DDR2　512MB × 2	DDR2　1024MB × 2
WOL	サポート	サポート
HDD	SATAII　80GB	SATAII　80GB
PCI バス	×3	×4
内部 LAN	1Gbps, オンボード	1Gbps, オンボード
管理 LAN	1Gbps,NIC	1Gbps,NIC
グラフィックス	オンボード	オンボード

４．１ **OS** の選択

Linux と Windows の２種類にすることにした。主となる OS に Linux を採用した。Linux のカーネルのバージョンは 2.6 である。現時点で最新の安定版リリースである。Windows を補助 OS として採用した。Linux のディストリビューションは Fedora core を選択した。理由は無償で入手することができ、市販解説書が他の無償のディストリビューションに比べて多く、大学などでの構成に向いていると判断したためである。

５．ノード透過性

どのノードにログインしても実行環境がほぼ同一であること（ノード透過性）を実現するための主な要素は、(1)すべのユーザのアカウントがどのノードにも準備されていて同じ UID 同じパスワードでログインできること、(2)ログイン後のファイル環境が同一であることである。(1)は全ノードに全ユーザアカウントを直接または NIS などで間接作成することで解決できる。(2)には DFS などの本格的な分散ファイルシステムを使う方法と良く使われるファイル環境（ディレクトリとその下のファイル）に限定して NFS を使う方法がある。Ships1 では中小規模との狙いから後者のアプローチを取った。

ユーザがよく使うディレクトリの中で重要なものは home の下のユーザディレクトリ、とくにその中の Desktop および MyDocument（保存用ディレクトリ）である。これらを透過的にする方法は２つ考えられる。(a)特定の１台のノードに全ユーザのディレクトリの実体を置き、実体に対して NFS マウントをする。(b)各ユーザのディレクトリ実体を各ノードに均等分散配置し、他ノードからは実体をクロスマウントする。(a)では１台に負荷が集中するため Ships1 では(b)の方法を採用した。なおロードバランサが実体の存在するノードを優先的に選択することを想定している。例えば、ユーザ数が 32、利用可能ノード数が 16 の場合、各ノードに対して２人分のユーザのディレクトリの実体を分散配置する。実体が配置されていない他の 15 ノードすべてから実体を NFS マウントする。

上記例の環境で実際にクロスマウントを行い実験を行った。ユーザディレクトリの NFS マウント数は全部で 480(=2×15×16)箇所となった。定常的な状態で任意のユーザが不特定のノードにログインした場合、ほぼストレスなしに作業ができることが確認できた。

図２ハードウェア外観

６．まとめ

本論文に記したアーキテクチャ、ノード設計、OS 選択に基づき、Ships1 基本部を完成させ、動作の確認に成功した。図２に Ships1 の外観を示す。ノード透過性については、前述の実験の結果、簡単な作業を行う範囲ではこの方式により実現可能であり十分な性能も確保できることが確認できた。より複雑な業務には DFS などの分散ファイルシステムを適用することも考えられる。Ships1 の研究開発は 2006 年４月から開始したばかりであり、高速ノード間接続、システム管理などが今後の課題である。

参考文献

[1]Hal stern,Mike Eisler,Ricardo Labiaga,NFS&NIS 第２版, 2001