5K－5

光バスクラスタ計算機Euphoriaの開発 (1) 概要

福井俊之† 濱口一正† 下山朋彦† 小杉真人‡ 柴山茂樹†
キヤノン（株）情報メディア研究所†/イメージング研究所‡

1. はじめに

キヤノン㈱情報メディア研究所では，従来のワークステーション（WS）に相当するノードを，光波長多重回線を用いて接続することにより，ノード間でのメモリの共有を可能とした「光バスクラスタ計算機」のプロトタイプ“Euphoria”の，設計・試作を行っている．

Euphoria はソフトウェア，ハードウェアの研究開発用プラットフォームとしての利用を目的とし，独自に開発したWSクラスタ型計算機である．オペレーティングシステムとしてはCMUで開発されたMach3.0を独自拡張したシステムソフトウェアを実装している．

本報告ではまず，Euphoria のコンセプト及びハードウェアの概要について解説する．

2. Euphoria コンセプト

近年，コンピュータネットワークの高速化，広帯域化がますます進んでいる．それに伴い，従来のWS内部でのマルチプロセッサから，ネットワークをまたがったクラスタ構成での分散共有メモリ型マルチプロセッサ方式への動きも現われている．

しかし既存のネットワーク技術に依存したクラスタ構成では，各アプリケーションがWSのプロセッサの高速性を十分に活かしてデータを処理するだけの高速データ通信を行うことは難しい．

このような通信ボトルネックを回避し，クラスタ構成の長所を活かせるような大規模アプリケーションを効率よく実行するためには，(1) ハードウェア転送速度を既存LANの10～100倍にする，(2) ソフトウェアのオーバヘッドを極小化し，実効転送速度をハードウェアの速度に近づける，などの高速ネットワーク技術の躍進が必要となる．

その答えの一つとして我々が提案するのが「光バスクラスタ計算機」である．光バスクラスタ計算機では，高速ネットワークの実現法として光を用い，更に光波長多重技術により複数ノード間での広帯域同時通信を可能とした．また，通信プロトコルも，ノードの内部バスプロトコルを基本としたハードウェアに近いレベルに設定し，通信オーバヘッドの低減を図った．これらにより，高速ネットワークによるノード間のメモリ共有を可能とし，より高性能なクラスタ型計算機を提供することを目的とした．

3. システム構成

Euphoriaは従来のWSに相当するノード，ノード間を接続する光回線，及び光回線の分配・調停を行うコンセントレータによって構成される．図1にEuphoriaのシステム構成を示す．

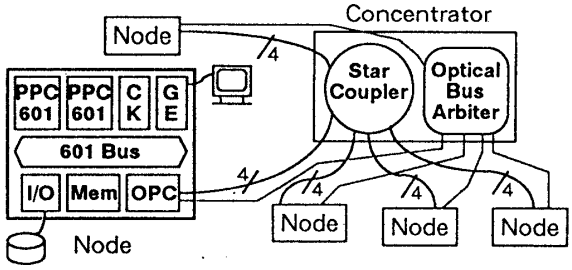

図1　Euphoria システム構成

ノード内部の構成は，先に当研究所で自作したWS Stonehigh[1] の設計を元にしている．機能的に付け加わったのは，光バスを実現する光コネクションコントロール[OPC]，及びノード間キャッシュ機構実現のためのコヒーレンシキーパ[CK]である．

光回線は，ノード間を相互に接続する波長多重データ回線，及び各ノード間のデータ回線利用の調停等を行うアービトレーション回線よりなる．データ回線のトポロジはスター型であり，Single Hop の Broadcast-and-Select方式を採用した．光信号の

Design and Implementation of an Optical Bus Cluster Computer “Euphoria”
(1) Overview
T. FUKUI, K. HAMAGUCHI, T. SHIMOYAMA, K. KOSUGI and S. SHIBAYAMA
Media Technology Laboratory, Canon Inc.

データ転送レートは200Mbpsである．なお，Euphoriaでは光デバイスの関係から，実際には波長多重を複数組の光ファイバ及びカプラを用いることでエミュレートしている．

コンセントレータはデータ回線の波長利用調停を行うための光バスアービタ，及び各ノードからの光信号を再分配するためのスターカプラによって構成される．光バスアービタはStonehighのdaughterボードとして構成される．波長調停処理は種々の方式が実験できるように，ボード上のMPU（68040）によりソフトウェアで実現される．

メモリアーキテクチャとしては，Euphoria はNUMA (Non-Uniform Memory Access) 型メモリアーキテクチャを採る．メモリ資源をアクセスするシステム中の全てのプロセッサは，図2に示すアドレスマップに従ってシステム中の任意のメモリ資源に対するアクセスを区別なく行うことができる．

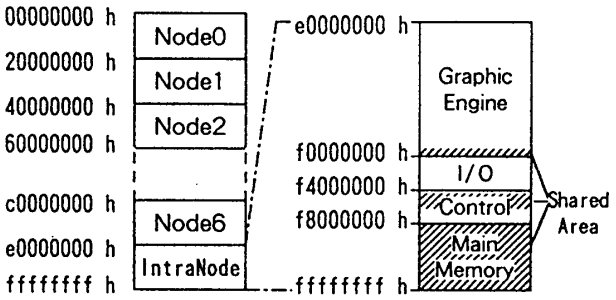

図2 Euphoria Address Map

キャッシュのコヒーレンシ保持はディレクトリ方式により行う．各ノードに自ノード中のメモリデータブロック対応分4Mエントリのディレクトリを備え，このディレクトリを管理するコヒーレンシキーパとコンセントレータ部に存在する光バスアービタがコヒーレンシ保持に必要な動作を執り行う [2]．

4. ハードウェア概要

4.1 プロセッサエレメント（PE）

PEは2つのメインプロセッサ，主記憶，及びメインプロセッサへの割込み制御等から構成される．メインプロセッサにはIBM & MotorolaのPowerPC601（66MHz）を採用した．ノード内部の共有バスである601Busは，PowerPC601のnativeバスを拡張したものであり，スプリットバストランザクションをサポートし，33MHzで動作している．

4.2 光コネクションコントロール（OPC）

OPCは光バスの制御を司る部分である．

光モジュールには1300nm帯のLED／PIN-PDモジュールを用いた．伝送速度は250Mbpsで用いる．なお，光波長多重技術をエミュレートするために，データ回線には4波長分のモジュールを準備した．実際のデータ通信では全2重回線を実現するために各々2波長を上り/下り回線の一組にして利用する．アービトレーション回線は光バスアービタとPoint-to-Pointに接続されるため，1波長分の送受信モジュールにより実現されている．

パラレル／シリアル信号変換器には，Fibre Channel用のICを利用した．符号化に8B／10B方式を採用している．

光回線を流れるFrameはSD (Start Delimiter)，FC (Frame Control)，DT (Data, Addr, Control)，ED (End Delimiter) よりなる．SD・EDには8B／10B方式のコントロールコードを割り当てている．

データ回線のFrameのDT領域では601Busをエミュレートするための情報が転送される．アービトレーション回線では，回線の接続のための情報やキャッシュのコヒーレンシ保持のための情報がやり取りされる．各Frameのヘッダ類（SD・FC・ED）はハードウェアで解釈される．

5. おわりに

光バスクラスタ型計算機のプロトタイプ“Euphoria”を設計・試作した．現在ハードウェアのデバッグ中である．続けてOSの実装，及びネットワークプロトコル，アービトレーション方法などの性能評価を実施していく予定である．

参考文献

[1] 伊達 他，「マルチプロセッサワークステーション“Stonehigh” −コンセプトとハードウェア概要−」，第45回情報処理学会全国大会6L − 02，1992.

[2] 下山 他，「光バスクラスタ計算機Euphoriaの開発 (2) メモリアクセス機構」，第49回情報処理学会全国大会5K − 06，1994.