CP-PACS システムの開発を省みて

和田　英夫
(株) 日立製作所エンタープライズサーバ事業部
RISC 開発部(当時)
第 1 サーバ本部第 3 部(現在)

2005 年 9 月に CP-PACS が約 10 年間にわたる稼動を完了しました. 私は, 初期検討から稼働後の事故対策まで, ハードウエア設計者として, CP-PACS を担当させていただきました. 至らない技術者でしたが, 筑波大の先生方は, よく付き合っていただいたと思うとともに, 10 年もの長い間 CP-PACS を使っていただいたことに感謝いたします.

2005 年 8 月に, 開通したばかりの, つくばエクスプレスに乗ってみようということで, 数年ぶりにつくばを訪れました. 地下の改札口から外に出てみると, そこは, 見慣れたつくばバスセンターでした.「ここがつくば駅になったのか」と初めて知りました. 駅周辺もずいぶん様変わりして, 華やかな街になったように思えました. 私は, CP-PACS の件で, つくばに 5 年くらい定期的に通っていたので, 郷愁に似た感覚を覚えました.

さて, その懐かしい CP-PACS の開発ですが, 以下, 年度毎に, 私が経験したこと, 思ったこと等, 意の趣くままに記します.

(1) 平成4年度

最初にぶつかった難問は, 長大データを処理するときに起きる, キャッシュ溢れによる, 実効性能低下でした. CP-PACS は, 従来のようなベクトル型スーパーコンピュータでなく, 汎用 RISC チップを並べた並列型スーパーコンピュータであり, 汎用 RISC チップは, キャッシュにデータがあるときに高性能が出るような構造になっているので, この問題が発生したわけです.

ここで, 考案された機能が, かの名高い「擬似ベクトル処理機構」です. この機構によって, 後続命令を止めることなく, メモリーからデータをパイプ

ライン的に取り込むことにより，ベクトル型スーパーコンピュータと同様に，メモリー上の大規模データを演算器へ高速に供給します．

また，PU(Processing Unit)間の通信の高速化が問題となり，Remote DMA 転送が考案されました．Remote DMA 転送では，送信側/受信側ともにユーザの仮想アドレス空間の一部をリアルメモリー上に固定的にマッピングしておき，それらのメモリー間でデータ転送を行ないます．これにより，異なるPU 上のユーザ仮想アドレス空間の間で直接データ転送ができます．カーネル空間とユーザ空間の間でデータコピーが発生しないため，高速な転送ができます．

さらに，PU 間ネットワークとしては，「3 次元ハイパークロスバネットワーク」を採用しました．「3 次元ハイパークロスバネットワーク」は，以下の構造を有します．

1. PU を，x 方向，y方向，z 方向に，n1 個，n2 個，n3 個，直方体状に並べる．(すなわち，総 PU 数=n1 x n2 x n3)
2. x 方向，y 方向，z 方向に並んだ各列の PU を各列ごとに完全クロスバで結ぶ．
3. 各 PU はスイッチ(router)を持ち，転送方向を変える(たとえば，x方向→y方向)ことができる．

「3 次元ハイパークロスバネットワーク」の特長は，以下の通りです．

a. 短距離通信：最大 3 回のクロスバスイッチ乗り換えで任意の PU と通信可能．
b. 柔軟なトポロジ：演算プロセスの PU へのマッピングに対する自由度が高い．

(2) 平成 5 年度

CP-PACS の構造も固まりつつあり，QCD 計算の性能評価を行なうために，メモリーシミュレータを作成しました．このシミュレータは，筑波大の先生方に，重宝がられたようで，光栄でした．

次に，大問題が起こりました．2^{nd} cache のアクセスに関する制御の問題

で，プリロード命令（メモリー先読み命令）が，1サイクルピッチで実行できないことがわかりました．どうやって，この問題を解決するか，長時間かつ激しい議論が続きました．ユーザデータをアンキャッシャブル領域(キャッシュに格納しないメモリー領域)に置こうか，2nd cache を使用しないモードを作ろうか，等々，種々の案が出されましたが，結局，ストアイン方式をストアスルー方式に変更することによって，解決しました．

また，ディスクに関して，信頼性の問題から，通常のディスクを並べるのではなく，RAID(Redundant Array of Independent Disks)の採用が決まりました．

(3) 平成 6 年度

メモリー素子の件で問題が起きました．当初，SDRAM（Synchronous DRAM）の採用を考えていましたが，供給が困難になりました．そこで，通常の DRAM に変更になりました．SDRAM は現在では，多くのマシンで使われ，popular になりましたが，この時点では，時期早尚だったようです．SDRAM にチャレンジしたかったのですが，残念に思いました．

次に，CP-PACS と FCS(センタフロントシステム：ユーザの CP-PACS 使用を支援するシステム)の間の実効性能が問題になりました．CP-PACS と FCS はピーク転送速度 100MB/s の HIPPI で接続しますが，通常のプロトコルで転送すると実効数 MB/s の転送性能しか出ません．これでは，大量のデータを扱う，CP-PACS ユーザには不足です．筑波大側から，「実効 50MB/s」という要求が出ました．これに対し，メモリー間転送の回数を減らした HFTP という転送プログラムを開発しました．システム完成後，平成 8 年に，実際に実機で実効 66MB/s の性能が出ることを確認しました．

また，ネットワークレイテンシ，メモリーレイテンシといった基本性能諸元や，バリア同期，ブロードキャスト，分割運転，ブロックストライド転送といった PU 間ネットワーク転送の詳細仕様が固まったのもこのころです．

実装構造が固まったのも，このころです．最大 1GB のメモリーをもったプロセッサ(0.3GFLOPS)を約 15cm×20cm の面積に凝縮し，45.6cm x 62.5cm のパッケージ当たり8台のプロセッサを搭載するという，世界最高ク

ラスの実装密度を実現しました.

(4) 平成 7 年度

CP-PACS の製造/評価が行なわれました. 毎 WG(ワーキンググループ)で, どの工程まで進んだかを報告しました.

結局, 設備は, センターの用意した設備の容量内で収まりました.

平成 8 年 3 月に, 1024PU 機の検収も無事に終わり, 達成感を味わいました.

(5) 平成 8 年度

システムが稼働開始しました. いくつか問題も起き, また, いくつか喜ばしいこともありました.

まず, 電源ノイズにより誤動作が発生しました. これは, PU 間ネットワークのクロスバ部の電流の変動が, 予想以上に大きかったためでした. 我々は, Bi-polar のマシンの経験はありましたが, CMOS の並列機は, CP-PACS が最初ということもあって, 電流変動量を読み誤りました. Bi-polar は論理回路の動作率にかかわらず, ほぼ同じ電流が流れますが, CMOS の電流量は論理回路の動作率に正の相関があります. CP-PACS は並列機であり, そのためにクロスバの部分が極端な論理回路の動作率の変動を起こすということを予想していませんでした. この点については, 平成 8 年 9 月の 2048PU 化の時に対策させていただきました.

この 2048PU 化によって, CP-PACS は, ピーク性能 614.4GFLOPS の, 世界に冠たるスーパーコンピュータシステムになりました.

さらに,「CP-PACS が 1996 年 11 月の TOP500 で1位をとる」という偉業を成し遂げました. この Linpack の測定は, 1996 年 9 月下旬に行なわれました. 朝から測定を開始したのですが, OS パニックは起きる, ディスクはパンクする, 等々で, 結局, 測定値が得られたのは, 翌日の明け方でした. マシン室で「世界記録達成！」と仲間で喜びをわかちあったのを記憶しています. このときの記録(Rmax)が 368.2GFLOPS でした. 今は 500 位以内に入るのに, Rmax 20.05TFLOPS が必要です. 技術の進歩の速さを感じま

す.

しかし, 2048PU 化以降, インターミッテントに, 原因不明のネットワークエラーが発生し, 悩まされました.「再現性なし. 発生位置も, 発生プログラムも一定しない. 」ということで, 約半年, 原因究明に苦しみました. 結局, ケーブルの製造不良(接触不良)が原因でした. この不良が解決して, CP-PACS は安定稼働に入りました.

この CP-PACS で培った技術(失敗も含めて)が, 日立のその後のスーパーコンピュータの開発に生きています.

私は, 最後の WG で, 中澤先生から「いろいろあったけれど, 全体としてみれば, よくやった. 」というお言葉を聞き, 胸が熱くなりました.