１９９９年１１月５日

サイエンティフィック・システム研究会
科学技術計算分科会　第２回会合資料

KEKB 計算機における大規模分散計算システムの運用

高エネルギー加速器研究機構 計算科学センター
真鍋 篤
Atsushi.Manabe@kek.jp

1．はじめに

高エネルギー加速器研究機構（KEK）では、素粒子物理学の重要な問題である CP 不変性の破れを解明するために、B ファクトリー加速器および BELLE 測定器を建設し、今年からデータ収集を開始した。KEKB コンピューターシステム（以下 KEKB システム）は実験準備、実験データ解析を目的に 1997 年より導入された 1,000 SPECint95 以上の計算処理能力、4TB のディスクと 150TB のテープライブラリーを有する大規模な分散計算機システムである。このような大規模な計算機資源をひとまとまりのものとして実現するには分散コンピューティング環境が必須である。分散システムでは、ハードウエアが分散され構成要素が多くなり、また構成要素間の複雑な相互依存関係が生ずる。　結果、ユーザーにとって名も知らぬ一台の計算機に問題が生じただけで、その影響が全体におよぶこともある。このため、大規模なシステムにおいては要素機器毎の信頼性には大きな問題がなくても、全体として信頼性・可用性の低いものとなる危険が少なからずある。これを避けるためには構成要素の冗長度をあげ障害時にできるだけバックアップ可能とすることが有効である。しかし、予算に限りがあるため、どこの冗長度をあげるかを見極める必要があるし、また、せっかくのバックアップ機構も肝心なときに働かないこともある。

2. KEKB コンピューターシステムの概要

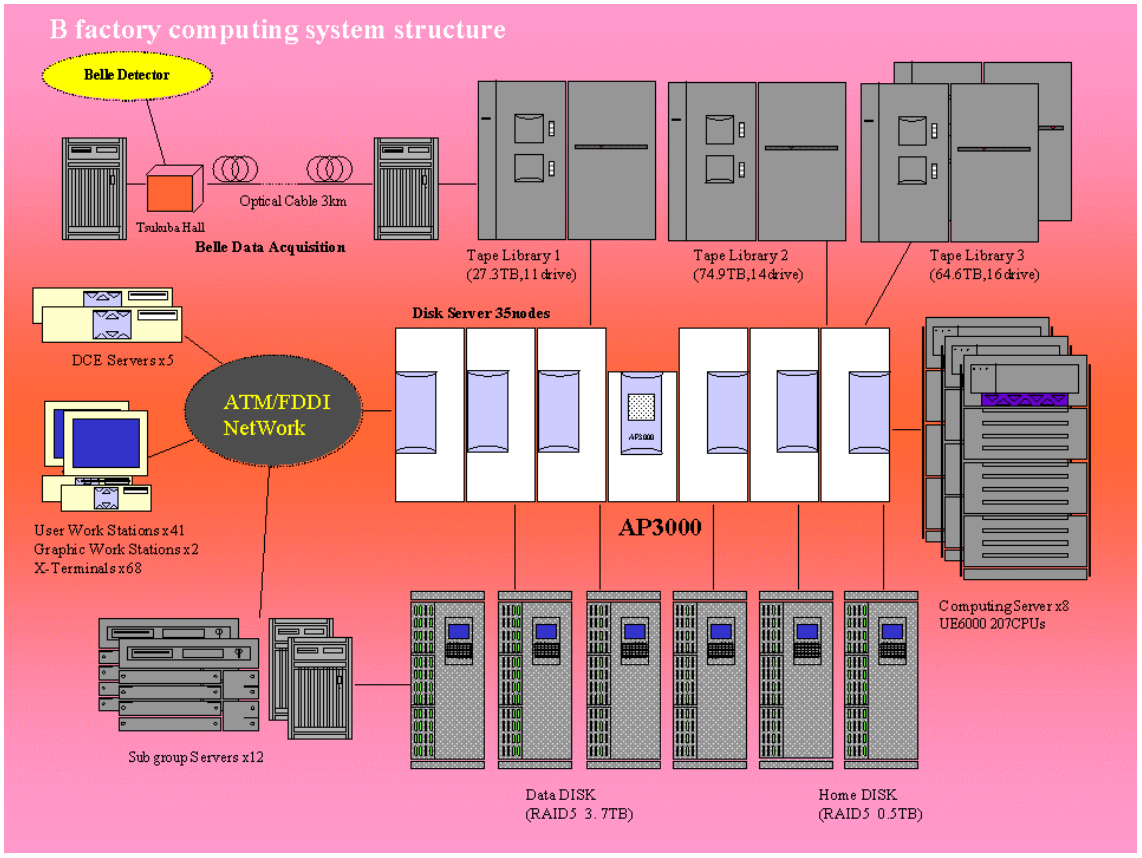
KEKB システムはすべて Solaris 2.5.1 を OS とする UNIX システムであり、中核のファイルサーバ（36 台）と計算サーバを富士通 AP3000 システムによりまかなっている。7 台の計算サーバは 28CPU を有する SMP(Shared Memory Processor)マシンであり、これが KEK の AP3000 を特徴づけている。 ファイルサーバーは一台あたり約 0.1TB の RAID5 ディスクに SCSI2 で接続され NFS サーバとして運用されている。NFS サーバのうち 2 台は４CPU をもつ SMP で後述する階層型ファイルシステム（HSM: Hierarchical Storage Management）サーバとしても機能している。

KEKB 加速器で電子、陽電子が生成、加速され、BELLE 実験装置内で衝突させられる

。この衝突で生じた二次粒子は BELLE 実験装置で検出され、検出データは光ケーブルで 3km 離れた計算機室に転送されテープライブラリーに記録される。最高 15MB/秒のデータを終日滞りなく記録する必要があるため、高速のディジタルテープが使用されている。

データは同じテープライブラリの計算サーバに直結された別のテープドライブから読み出され解析が行われる。計算サーバでは表１に示される種々のジョブが実行される。

これらの主なジョブは、1 ジョブで fork()と shared memory プロセス間通信による比較的単純な並列プロセッシングをおこない、28CPU を無駄無く使用している。電子-陽電子の一回の衝突事象を「イベント」というが、高エネルギー実験物理学のデータはイベント間の依存性が殆どなく、また多数のイベントを解析する必要があるため、本質的に並列計算が容易である。しかしながら、28CPU まで台数効果を上げることができるのは SUN Ultra Enterprise System のメモリ設計、バス設計の優秀さによるところが大きい。



KEKB システムにおいては、AP3000 のセールスポイントである AP net は並列計算の計算機間通信には使われておらず、ジョブが走る計算サーバとファイルサーバの間で NFS プロトコルを運んでいるだけである。

ユーザーは、X 端末や WS(work station)経由で 13 台のワークグループサーバにログインしてプログラム開発やグラフィック処理を行う。多量のデータや計算資源を使用するジョブは Platform Computer 社の LSF(Load Sharing Facility) によるバッチキューシステムで負荷分散をとりつつコントロールされる。

表１

JOB の種類	必要計算能力 (SPECint95)	必要 I/O 能力 (MB/s)	同時 実行数	データ 生成量	記憶媒体
実験データ収集	-	15MB/s	1	10TB	tape
実験データ解析（再構成）	120	6MB/s	1~3	15TB/年	tape
実験データ解析（分類）	20	6MB/s	1	1TB/年	disk
グループによる シミュレーション	120	5MB/s	1	8TB/年	tape & disk
ユーザーによる解析１（抽出）	8	5MB/s	2~3		disk
ユーザーによる解析２	2	2MB/s	~20		disk
ユーザーによるシミュレーション	20	＜１MB/s	~10		disk

3. 分散コンピューティング技術の適用

つぎに、KEKB システムで使用している分散コンピューティングシステム技術についていくつか紹介する。

ユーザーアカウント管理

分散コンピューティング環境において、ユーザーの認証、資源アクセス権の認可、 uid, gid などの属性管理は基本的なものである。このような場合 NIS(Network Information System)が多く使われてきたが、KEKB システムではセキュリティの観点などから kerberos 認証を用いる TOG(The Open Group)の DCE(Distributed Computing Environment)を使用している。 DCE、DFS(Distributed File System)によりユーザー/バッチジョブはどのホストにログインしても/ディスパッチされても、ホスト名以外については同一の環境を得ることができる。DCE はサーバーとして Cell Directory (Directory サービス) サーバ、セキュリティサーバ（認証・権限認可サービス）、時間サーバ（時間同期サービス）等によりサービスを提供するが、すべてのサービスにおいて複数台サーバによる負荷分散と障害時に備えた冗長度を上げる設計がなされている。例えばセキュリティサーバは、通常複数サーバの任意の一台をマスターとし、マスターのみがセキュリティ情報を書き換えることができる。 他のサーバはクライアントの読み出し要求にのみ答える。定期的にサーバ間のデータベースは同期がとられる。マスターサーバに異常が起こった場合にはオペレータの介入によりマスターサーバを変更できる。自動的に変更されないのは、ブリッジの故障などによるネットワークの分割によりマスターが 2 つそれぞれ独自にデータベースを変更し相互に矛盾が生じるのを避けるためである。マスターに異常があり、マスターが不在となってもユーザー認証には支障がない。新規のユーザ登録などができなくなるだけである。

計算処理能力の負荷分散 （バッチシステム）

長時間の計算処理は LSF によるバッチキュー処理により処理を行う。 LSF ではキュー毎に実行ジョブの資源制限、ジョブスケジューリングポリシー等を設定することができ、ユー

ザ間、グループ間の公平な資源の分配を大型汎用機なみに行うことができる。また、キューのパラメーターは運用を続けながら動的に変更できるのでチューニングが容易である。
LSF は主に、バッチ実行全体を制御するマスターサーバ、ジョブを実行するホスト（バッチサーバ）の資源使用状況などの情報を定期的に収集し、集中して管理する負荷情報サーバ、ジョブが実行されるバッチサーバ、ユーザが実際にジョブをサブミットするクライアントからなる。負荷情報サーバのホスト情報によりマスターバッチサーバーは次にジョブをディスパッチするサーバを選択する。バッチサーバに異常が起こった場合、実行中のジョブは失われるが、他のバッチサーバで実行されているジョブや新規にサブミットされるジョブに影響はない。次に負荷情報サーバやマスターバッチサーバ機に異常が起こった場合は他のバッチサーバ機がそれらの機能を代替する。したがって理屈上ではバッチサービス自体はすべてのバッチサーバがダウンしない限り継続することができる。

4. ファイルサービス

KEKB システムにおけるファイルサービスはその用途により大まかに次の 5 種類分類され、それぞれ異なる方法により提供されている。また、システムファイルを除き実際にファイルを保持するディスクには活性保守が可能な RAID5 ディスクが使用されている。ディスクユニットの総数が 600 を超えるため、平均寿命を単純に割り算すると週に１度程度のユニット故障が予期される。実際には初期故障を除き、月に１ユニット程度の故障に収まっている。運用時間が平均寿命を過ぎるころから故障頻度が急増しないことを祈っている。

<u>システムファイル</u>： OS やシステム設定ファイル、ログなど。このファイルにアクセスできないと計算機が停止してしまうため、計算機に直接接続された独立な SCSI ディスクに 2 重化して保持している (Solaris Disk Suite 使用)。２つのうち１つに異常が生じても運用続行が可能であり、ほとんどが読み出しだけなのでオーバーヘッドもほとんどない。しかし、<u>立ち上げの際に</u> Solaris Disk suite の問題で<u>立ち上がらなくなってしまうこ</u>とがあった。

<u>ホームディレクトリー</u>： ユーザー開発のプログラムソースやメイル、ドキュメントなどのファイルをこの場所に保持する。ファイルのサイズがデータファイルなどに比べ小さく、読み書きの比率はほぼ同等である。KEKB システム以外のホストからマウントされることを考慮してセキュリティー上の観点から DFS が使用されている。DFS では、１日に一回スナップショットのバックアップがとられるので、ユーザーはうっかり消してしまったファイルなどを自分で簡単に復旧することができる。

<u>アプリケーション実行ファイル</u>：ユーザー開発または、市販のアプリケーションソフトウエアのファイルを保管するためのものであり、DFS で提供されている。ほぼ読み出しのみで、多数のクライアントから同時にアクセスされるため、クライアントのファイルキャッシュが有効に機能する。また、ファイルの実体を複数自動的にコピー・保持して

負荷分散や可用性を向上するレプリカサーバ機能も有効につかえる。

データファイル（1）： 解析結果などのデータで、多くのユーザが同時に使用する可能性の高いファイルが置かれる。階層型ファイルシステムと NFS (Network File System Ver.3) によりサービスされる。階層型ファイルシステムは簡単にいえば使用頻度が下がったファイルが自動的にディスクからテープなどのアクセス性は低いが保存コストの低い媒体に移動(migrate)され、再び必要になったときテープなどからディスクに戻される仕組みである。KEKB システムではディスク階層が 3.5TB、テープ階層を含めると最大約 60TB がこの方法で提供されている。ファイルの典型的サイズは 1 GB オーダーで、読み書きはほぼ同頻度である。一人のユーザが次々と一連のファイルをスキャンしていくことが普通であるため、ファイルサイズの大きさと共にクライアント側のキャッシュが大きなオーバヘッドとなるため DFS は使われず、NFS を AP net を通じて使用している。AP net 自体は 200MB/秒 の転送速度を有するが、プロトコルオーバーヘッドと DISK の性能の上限（～10MB/秒）により NFS を介する使用では 8MB/秒程度の性能である。

データファイル（２）： 実験の生データで、データ収集などの定型ジョブにより使用される。ファイル最大容量は約 80TB である。ファイルの平均サイズは 40GB、読み書き頻度はほぼ同程度。かぎられたユーザからのみ利用されること、転送速度を保証する必要があること（データ収集で 15MB/秒）から、テープへの直接読み書きによりおこなっている。テープのボリューム管理やテープライブラリーのコントロールには富士通特製のソフトウエアを使用している。

DFS と NFS

ファイルシステムとして、主に DFS と NFS を用途により使い分けていることがおわかりいただけたと思う。DFS と NFS の主な相違は

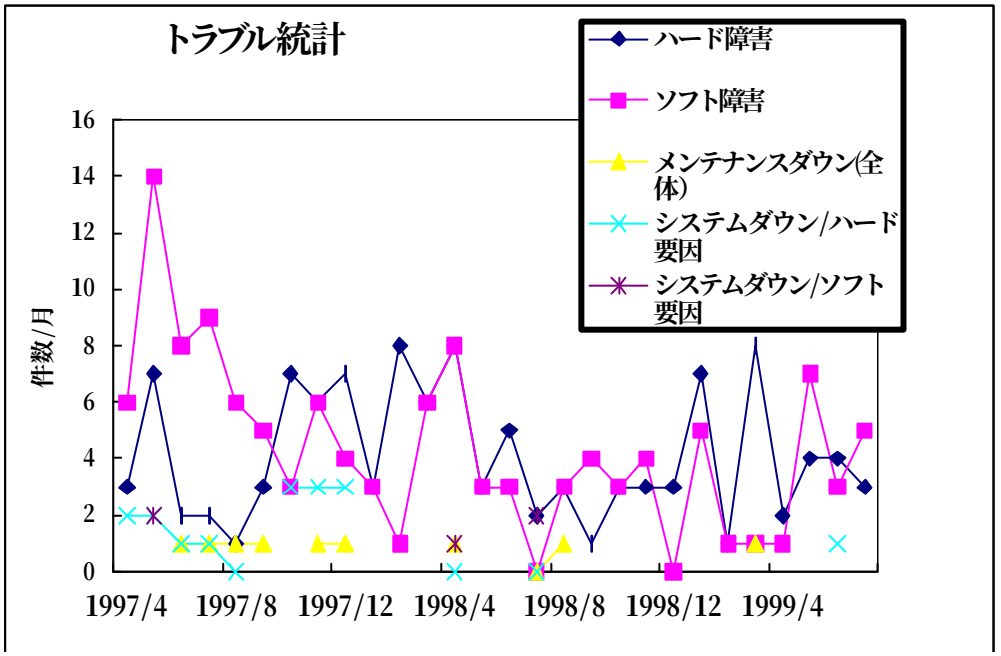
(1) クライアントにおける、ファイルキャッシュの有無。
(2) ファイルシステムのマウント情報を、クライアント側に（NFS）持つか、ネットワーク上の独立したファイル情報サーバ（DFS）に持つか。
(3) レプリカサーバの有無。

等である。DFS では、各クライアントに必要に応じた量のキャッシュを持つことが出来、複数のキャシュと実体ファイルの無矛盾についてもよく考慮されている。これにより、クライアント数がふえても、サーバ側の過負荷がおこりにくい。一方、キャッシュにヒットしなかったデータについては、オーバーヘッドのために NFS より転送速度をあげることができない。キャシュの中身の更新が頻繁になると、このオーバーヘッドがひどく大きくなる。したがって、キャシュのサイズやユーザのファイルアクセスパターンなどを考慮しながらのチューニングや適用が必要である。

上記(2)の特徴のために、DFS はファイルの物理位置（サーバの場所）をクライアントユーザには透過的に移動することができるため、パーティションあたりのユーザの配置につい

てあまり悩む必要がない。（移動中のファイルへのアクセスは、移動終了まで待たされる。）
サーバが増えた場合でも、クライアント側に一台ごとにマウント情報（/etc/fstab や auto_xxx）を用意する必要がないため、クライアント側の設定が随分容易である。
DFS では ファイルの複写をもつ複数の読み出しのみ可能なレプリカサーバを持つことができる。これは読み出しが主なアプリケーションソフトの供給には非常に有用である。読み書き可能なレプリカサーバは開発が待たれる技術であるが、ファイルの同期と coherency(一貫性)に困難があり、いまだ実現されていない。

5. トラブル



BELLE 実験は本夏より本格的にデータ収集を始めている。システムは 1997 年に導入されたので、約 2 年間は実験データなしに、シミュレーションデータの生成やソフトウエアの開発に使用されてきた。この段階で計算サーバの CPU 使用率が平均で 70～90%、使用ファイル量 6TB(うち４TB がテープにある)、ファイルサーバの平均 CPU 使用率は 3～20%程度である。今まで述べてきた各種の分散コンピューティング環境を適用した結果はどうであったろうか。上図にトラブルの発生状況を描いたグラフを示す。ただし、運用に全く影響の出なかった冗長部分の故障などは含まれていない。最近の UNIX システムワークステーションはかなり安定で何か月もダウンせずに動作するが、このような大規模なシステムでは、また次元の違う話がある。

グラフよりシステムを運転してから一年半ほどは非常にトラブルが多かったことがわか

る。主な原因はソフトウエアおよびハードウエア（ファームウエアの問題や不適切な部品の使用等）のバグであり、SMP 上の DFS クライアントの不具合、RAID コントローラの不具合、ファイルサーバのハードウエア、テープライブラリ・ドライブの不具合等が含まれる。図中システムダウンとあるのは、フォルトトレランス機能がなかったり、バグで機能しなかったために、ユーザーがなにも仕事ができない状況におちいった回数である。これらは障害全体の 1/10 以下ではあるが当初は半月に一回程度も発生した。また、問題を解決するために頻繁なメンテナンスダウンも必要であった。しかしながら、最近では、ほぼ仕様（3 ヶ月に 1 回以下）程度に安定運用できるまでこぎつけている。これもひとえに、このような複雑なシステム見捨てず、ご努力いただいた富士通および関連メーカ各位のご協力の賜である。この場を借りて深く感謝するとともに、今後の変わらぬサポートをお願いする次第である。

最後まで残っている問題としては、テープドライブの問題（最近解決した）、CPU 故障による障害、キャッシュメモリーのソフト放射線による問題とあといくつかのソフトウエアに関する問題がある。

6. 分散システムの憂鬱

システムの管理者からすると、本質的に事情が複雑になる分散システムなどはやらなくてすめばそれに越したことはないが、計算機需要を満たすためにはそうもいってはいられない。もう一度、分散環境の問題点を整理すると、

1） 物の量が増えるため、管理の対象も増える。
2） それに付随して、システムが複雑になる。
3） 相互に依存関係が生ずる。この依存関係が普段は意識されないため、問題が起こったときに予期せぬ事態になる。
4） 関係する業者の数が増える。

などである。

分散環境用ソフトウエアで、1）の問題を解決すると唱うもの (TCO 削減)が市場に多くあるが、

i) 実際使ってみるとたいして効果がないものがある。一番大変なのは、普段と違ったことを行わなければいけないとき（トラブル時）であるが、それは想定外の場合が多い。
ii) 信頼性をあげるより、市場でアピールするように新機能を付け加えることが優先になりがちである。普通使われる方法とちょっと違った方法で使っていたりすると、そのトラブルを直すメリットはベンダーには少ない。古いバージョンが早々に切り捨てられることがある。
iii) マルチスレッドや並列処理などで動作するものは、負荷が高くなると顕在化するバグが入りやすく、この種のバグは修正することが非常に困難である。

iii) 分散環境対応のソフトウエアは海外ベンダー製品であることが多く、国内の中間代店の技術力や、やる気で受けられるサポートが段違いであることもある。

などの問題があるが、これは分散環境が普及するまでの過渡的な問題なのかもしれない、というかそう考えたい。

B計算機システム３年間の運用から得た教訓としては、以下のようなものが挙げられるだろう。

A) クライアント、サーバーホストはできるだけ多量に用意し、できるだけ各ホストを単用途かつシンプルにする。サーバーの独立性を向上させる。

B) 冗長度を高くする。たとえ冗長にできなくても、細分化して被害が局所化するようにする。すなわち、大きいもの少数より、小さいものを多数用意する。ただし、安物のマシンはすぐ壊れるのでそのことを注意する。高いものは伊達に高いわけではない。

C) 沢山のものをうまく管理する方法をもっともっと考える。

これは、数を増やすと面倒なことが増えるということに関して逆説的ではある。100 のものが 10 に減るなら、それを真剣に考えるべきだが、減らすことができないのであれば 100 のものが 150 になっても労力は同じで問題を減らす可能性が生まれるということである。

分散環境の大変さに世間も懲りたためか、最近は一台の計算機システムに 100 以上の CPU を載せて、これらを分割しても一体としても運用可能で管理が容易で、冗長構成も可能なスーパーサーバが多くみられるようになってきた。これらが思った通りに動作し、価格も安くなれば嬉しいことであるが、今後はどうなっていくのであろうか。

［参考文献］

(1) KEKB コンピューターシステム仕様書： http://corvus.kek.jp/~manabe/KEKB

(2) 真鍋 篤：分散システム・インターネット運用技術シンポジウム '99 論文集情報処理学会シンポジウムシリーズ Vol.99 No.3 p.35

(3) DCE/DFS: http://www.transarc.com

(4) LSF: http://www.comtec.daikin.co.jp/lsf/index.html

(5) Sony PetaServe (HSM): htp://corvus.kek.jp/~manabe/