TSUBAMEの飛翔: ペタスケールへ向けた「みんなのスパコン」の構築

松岡 聡
東京工業大学学術国際情報センター
matsu@is.titech.ac.jp

1. はじめに – TSUBAMEへの道

Tom Sterling, Don Becker提唱のBeowulf型［1］PCクラスタが1994年に生まれてから12年余りが経過した。最初のWiglafdeの性能は数百MFlopsに満たなかったが、この十年で10万倍近い性能の進歩を果たした。しかしながら、地球シミュレータのような専用設計のスパコンと比較すると、単純に性能だけでなく、サイズ的なスケーラビリティ・運用性などに関して、今後の100TFlops/万単位のプロセッサ時代のトップマシンをクラスタ技術で構築し、センター運用の多数のユーザの様々なアプリケーションの実行環境を安定に供給できるかは未知数であった。

TSUBAME (Tokyo-tech Supercomputer and Ubiquitously Accessible Mass-storage Environment)は、アクセラレータ部分を含めると100TFlops近いピーク性能・20TByte以上のメモリ・Fat Nodeおよび高速な間接網を基本とした共有メモリスパコン形アーキテクチャ・並びに1PByte以上のオンラインストレッジを備え、2006年春の時点で、あらゆるパラメタで現在でわが国最高性能・最大規模の新型スパコンとして東工大・学術国際情報センターに構築された。TSUBAMEは最高性能のスパコンであると同時に、超大規模のPCクラスタとして、従来の専用スパコンとPCクラスタ両者の利点を兼ね備え、かつそれぞの欠点を補うマシンとして設計されており、スパコンとしての柔軟かつ強力なハイエンドの大規模並列計算やデータ処理の能力と、エンドユーザのデスクトップ・ノートブックやPCクラスタ環境との高い親和性（同じバイナリが動くなど）との両立を実現している。これらを同時に満たすことが、TSUBAMEが標榜する「みんなのスパコン」の世界であり、従来の限定されたユーザのみを対象としたスパコンとはその基本的な発想を異にしており、その点で今後の大学や研究所運用のスパコンおよびセンターのあり方に関して、大きな方向性を示すことができたと信じている。

特に今回TSUBAMEの大きな目標は、従来のスーパーコンピュータの既成概念を打ち崩すことにあった。数十年前のメインフレーム時代は、一台の計算機を多くの人間が閉じた環境で共有して使っていたため、逆に使用が可能な端末が置いてある箇所からは全て同一の環境---いわゆるSSO(Single Sign-On)---が自然に実現しており、業務系の処理からシミュレーションまで、同一の環境で行われていた。その後、70年代後半から80年代よりマイクロプロセッサ/PC時代に本格的に移行し、同時期にCray等のスーパーコンピュータが提供されるようになって、計算環境は大幅に異機種・異環境のものが混在することとなった。しかしながら90年代以降マイクロプロセッサの性能向上やビジネス利用のための信頼性が著しく高くなり、かつLAN, WANともに劇的な高速化・コモディティ化が急速に進行して、さらにソフトウェアのOSや並列プログラミングの環境がLinux,Windowsや、MPI/OpenMPIなど、寡占化や高機能化が進むにつれて、PCクラスタ等の最新のPCやそれを中心としたコモディティながらエンタプライズ系の最新技術を用いられ、劇的なコストパフォーマンス向上とが可能となった。実際、エンタプライズ系でもLinuxやWindowsは大幅に採用され、x86系の性能・機能向上は、多くのハイエンドRISCプロセッサをほぼ絶滅に追いやった。

ところが、特にわが国を中心として、スパコンは80年代の異機種・異環境のレガシーが、ほとんどの大学等のセンターに導入されるスパコンにおいて今でも残り続けている。結果、として、以下に代表されるような状況となっている：

- Window, Linux等と異なるOSやミドル、コマンド体系、ツール群
- ファイル共有やその他の様々なサービスにおけるデスクトップパソコンとスパコンの非共有・費接続性
- 特殊なマシンでの重要なISVアプリの欠如、非対応、高コスト化
- プログラム言語や最新の開発環境ツール(Visual Studioや)などの欠如

当然これではシステム全体のコストパフォーマンスは低くなり、またユーザベースが寡占状態にある通常の環境から移行しにくいので、結果としてスパコンというマーケット全体の縮小を生んでいる。

逆に一般のエンタープライズインフラ、特にサーバを鑑みると、すでにメインフレームの時代は過ぎ、 x86 + Linux/Windowsは50%以上の金額ベースを占める[2]多大なる成功を収め、かつビジネス用途にとっても十分な信頼性をもたらしており、デスクトップ環境との高い親和性を実現している。しかも、近年に並列処理においては、いかに効率のよい並列演算ユニットを安価に並べ相互結合することが性能を決定しており、x86のハードからアプリケーションに至るまでの大量生産・活用による「枯れた」プラットフォームの活用は大変技術的に有用となる。

逆に、IT投資の観点から鑑みると、各センターのスパコンの年間予算をスーパーコンピュータに、「共有設備」の名目で実際には数十名程度のユーザベースに成果のチェック機能なしに投入し続ける合意を今後得るのは難しい。むしろ、学内の多数のユーザに供し、かつデスクトップと連続した環境で、多数の一般サービスやアプリケーションをサポートしていく必要がある。これは(技術世代は全く違うが)メインフレーム時代と共通する思想でもある。

以上により、東工大における新世代のスパコンは、従来のように大規模な計算・ストレッジ・ネットワーク能力を生かしたハイエンドなシミュレーションの機能を実現しながら、同時にあり余るそれらの資源を生かして現代のIT環境にシームレスに接合し、様々なサービスを提供する源となって、学内や組織内の多数のユーザのIT ニーズを吸収し、大学における教育・研究アーカイブなどの教育研究業務、さらにはメイルや事務処理などの日常業務などの、統合的なホスティングを行うこととした。これにより、それぞれの研究者ユーザに統合的なIT 環境を提供することが可能となり、学部生から大学院生、教員系から事務系組織まで全員が日常で利用する**「みんなのスパコン」**環境が実現されつつある。のみならず、学生が日常的にスパコン環境に触れ、利用のハードルを非常に低くすることで、次

世代のシミュレーションユーザを多数生み出し、ひいては科学技術の発展に大幅に寄与することが期待されている。

2. TSUBAME の設計要件

TSUBAME の設計にあたり、いくつかキーとなる技術要素が実質的には 2002 年から検討されていた。前回の調達で 2000 年 1 月に GSIC センターの前身である総合情報処理センターに導入されていたスパコン群は、ベクトル型である NEC SX-5 (16CPU、128GFlops), 並列共有メモリ型である SGI Origin 2000 (256CPU、202GFlops), 並びに Compaq GS320 (64CPU、128GFlops)で、合算ピーク性能は約 448GFlops、ストレッジはテープを合わせても 50TByte 程度であった。また、OS も独自色の強い、しかもマシンごとに異なる Unix 系の OS であり、ユーザ環境とのファイル共有は言うに及ばず、マシン間でもあまりファイルや環境の共有ができていなかった(NFS サーバは存在したが、低性能であった)。

そこで、東工大のセンターが GSIC に衣替えした 2001 年の 1 年後の 2002 年 4 月より、PC クラスタを中心としたクラスタ・グリッドの運用実験である「東工大キャンパスグリッド」プロジェクト[3]が開始され、2004 年には総計で 1000CPU を超える大規模な PC クラスタ群＋従来のスパコン環境となり、センターとして 4 年あまり様々な運用経験を積み、またユーザに並列環境への段階的な移行を促した。同時期に、2006 年 4 月に導入される新スパコンの技術目標と、それを満たすデザインに着手した。

新スパコン TSUBAME の大きな技術目標は、「みんなのスパコン」としての幅広いユーザ層獲得、学内研究利便性向上&大規模シミュレーション計算ユーザの要求を同時に満たす新世代計算科学インフラ構築であるが、以下の要件に細分化された：

(a) 2006 年 3 月の構築時点で我国 No.1 となる、(1) 40TFlops 超の総合演算性能、(2) 1 PByte 超オンラインストレッジ、(3)それらを数 Tbit 級で接続する相互結合網結合、をコモディティ技術を用いて実現
(b) これらが全キャンパス・(他大学や企業など外部組織を含む)外部の仮想的研究組織からアクセスがシームレスに可能になるようなグリッドを構築、運営
(c) ヘビーユーザ(Capability)向けの情報サービスのみならず、全学の大幅な研究利便性を高めるライトユーザ(Capacity)向けサービスとの両立
(d) 高いコストパフォーマンス・低消費電力・安定性を実現、それにより東工大 GSIC センターが大規模計算基盤の世界的リーディングセンターとして認知

(a)(1) の 40GFlops 超は地球シミュレータを超えるもので、いわゆる「トロフィー・ハイエンド」つまり最先端の計算能力の提供を目指したものであるが、それと(b)の「みんなのスパコン」としての汎用な環境との両立、さらには(d)のコスト・設置面積・時間、さらに困難なセンター全体の電力の制約を満たすのは、設計次点のテクノロジでは無論困難であった。そこで設計の各種パラメタの決定は 2006 年初頭に登場するであろう各種 CPU や他のハードウェアの制約を予測して行われ、性能や消費電力の絶対値やそのトランジエントな特性などの基礎データとは研究室の Presto III を用いて取得し、安定性や運用性の基礎データは、キャンパスグリッドの各クラスタの数年にわたる運用状況を参考とした。その結果、CPU ノードとインタコネクトは以下のような要求が洗い出された：

- **CPU は高性能・64bit で、かつ x86 互換性の高いプロセッサ。**評価の観点は、性能・ソフトウェアの汎用性、特に Linux やその他汎用ソフトウェアの作動・低消費電力・低コストなどである。特に、
  - 高い FP 性能を提供するマルチコア CPU による高コストパフォーマンスと高信頼の確保
  - 汎用高性能プロセッサによる最先端の設計とプロセス技術による高性能かつ低消費電力
  - x86 互換 CPU による OS や Web・グリッドミドルを含むソフトウェアのデスクトップとの互換性、並列環境を含む各種コンパイラ・ライブラリ等のプログラミング環境の存在、各種 ISV アプリの最新版の提供

  が重要なポイントとなった。
- また、通常のスパコンに習い、**ノードの物理的な構成は、なるべく多くの CPU を物理的なノードのパッケージに含むいわゆる FAT Node のアーキテクチャ**を高く評価した。多くの MPP 型のスパコンは、近年はほとんど 8-32CPU 以上、メモリ数十 GB 以上(1-4GByte/CPU FLOP)以上の FAT Node 型である。それは、以下の理由によるところが大きい：
  - 大規模 SMP での並列プログラミングの容易性
  - 大規模ハッシュやデータベースなどのインコア化による検索系アプリケーションの高速化
  - ノード数の削減による管理の省力化・信頼性の向上・相互結合網の単純化・消費電力の削減

  問題は、x86 バイナリを実行できるアーキテクチャにおいて、HPC に適した高密度かつ高性能の 64bit SMP が存在するか、であった。設置予定であった 2006 年 3 月の段階では、Intel Itanium 次世代デュアルコアの Montecito ベースのものと、AMD Opteron 800 シリーズのデュアルコアベースのものが選択肢として残った
- **相互結合網においても、スケーラビリティ・高バンド幅・低レーテンシ・高コストパフォーマンス・高信頼、並びに２フロアに跨る設置に耐える長距離配線**などの要求項目が洗い出された。これらは大規模な並列マシンに要求される当然の項目であるが、さらにそれらに加えて、以下の要件を重要とした：
  - **将来の性能向上・アップグレード可能性：** 今後計算やストレッジノードの性能を向上させて行くにつれ、ネットワーク自身の性能向上が安価かつ比較的容易に図れることとした。
  - **標準規格採用による汎用接続性：**TSUBAME 全体のアップグレードなどを行う際に、他のネットワークインフラ、特に WAN との高速な接続性を確保できることとした。
  - **TCP/IP ベースなどのマルチプロトコルのサポート：**TSUBAME では、ノード間の MPI 通信のみならず、ストレッジに対するアクセスなど、全てを単一の高速ネットワークで行う設計とした。すでに多くの MPP ではとられている手法だが、PC クラスタではストレッジへの接続は専用の SAN だったり、あるいは MPI 用のネットワークとは別途 IP ネットワークであることが多かった。今回 TSUBAME では、TCP/IP を含むマルチプロトコルをサポートし、冗長性のある超高速ネットワークを用いることで、MPI やストレッジのネットワークを完全統合することとした。これは、ストレッジネットワークを含む全体のコストパフォーマンスや信頼性を上げるとともに、コストや複雑性を下げる。実際、東工大キャンパスグリッドや過去のスパコンの経験から、MPI 等の計

算トラフィックに対してストレッジのトラフィックはバースト的でかつ圧倒的に少ない。

- ケーブリングの少ない間接網・Fat Switch 構成：Fat Node を採用することにより、マルチレーン構成による Fast and Wide な結合網の構築が可能となった。このため、幅広く多くのアプリケーションに対応可能な低レーテンシ・高バンド幅・均質な間接網の採用が可能となった。しかしながら、ノード数が CPU 数と比較して少ないとはいえ、数十ポート程度のスイッチ群の FAT Tree や CLOS 構成では、最低限でも数十台規模のスイッチ群が必要となることが予想された。TSUBAME の設置面積は 300m2 以上と試算され、本センター内のフロアスペースを考えると階を跨って設置される予定だったため、このような結合網ではケーブリングが困難となることが予想された。また小規模のスイッチは、しばしば電源などの冗長性が不足しており、かつ間接網では一つのスイッチの故障が(ノードの故障と異なり)多数のノード群に多大な影響を及ぼすので、耐故障性の面からも好ましくない、と判断した。そこで、数百ポート規模の大規模スイッチのみを少数台接続する構成をとることにより、この問題を解決することとした。これにより、スイッチを比較的ノードの近くに置くことができ、通常の短距離用のケーブルを用いることができる。一方、スイッチ間はフロアを跨ぐなど、距離が長くなるので、並列の光ファイババンドルが用いれることとした。

- 東工大キャンパスグリッドや旧スパコンからの経験のみならず、今回多数のユーザを従来より遥かに広がった用途でサポートするため、ストレッジの設計は当初から最重点項目とされた。特に、容量は、地球シミュレータの総合データ容量や、San Diego Supercomputer Center のアーカイブシステムの前例から、最低１ペタバイトは必要とした。一方、TSUBAME の用途が多岐に渡ること、前述のネットワーク構成により数十 GByte/秒の I/O 速度が可能となったなど、また最近の技術傾向として、MAID 技術など、HDD ベースのストレッジ技術が進歩したことで、「4 年間データを高信頼保存可能で、かつランダムアクセス性の高い様々な用途に耐え、さらには数十 TFlops 級の大規模並列処理に必要な数十 GB/s の I/O 速度の確保を、コンパクトかつ省電力に実現する HDD ベースのストレッジの構築」が技術目標となった。

  これを満たすには、従来のエンタプライズ系 SAN スイッチを中心としたストレッジは、耐故障性は高いものの、性能面・コスト面で不適であると判断した。特に、先述の相互結合網が存在するので、それと SAN ネットワークを共存させることは二重投資となってしまう。むしろ、相互結合網が直接接続される NAS ストレッジを中心とした高密度システムが目的によく合致する。

  しかしながら、NAS の問題は、(1) いわゆる single point of failure がストレッジノードおよびネットワークに発生しやすいこと、および (2) NAS 自身の Disk I/O やネットワークバンド幅と HDD/CPU 比のバランスや消費電力などの問題である。これらは、以下のように解決するのが最善と判断した：

  - (1) に関しては、まずは一部 SAN で二重化したストレッジを設けて、それをユーザの/home やマシン全体の動作に重要なストレッジを設けることとし、一方 NAS は主に/work エリアとして、かつ一時的には unavailable となっても、修復の時間が十分短い(数時間内)であれば、データ損失さえなければ当初は OK とした。これは主にネットワークやコントローラ基盤などの、部品スワップで容易に回復できる修理を想定した。NAS 内の HDD 自身は当初 RAID5 を、将来は RAID6 を実現し、1PByte でも 4 年間でのデータ損失確率を 0.25%として設計を行った。無論、通常のテープバックアップをエミュレーションする差分バックアップも disk-to-disk で当然行う。

    さらに、NAS のノード間で将来的には striping RAID を検討した。すでに当初から Lustre[4]などの並列ファイルシステムが莫大な I/O 速度を生かすためには必須であることは明白であり、NAS 間 RAID はそれの自然な拡張である。このように、NAS 内の RAID と NAS 間の RAID、さらには他の高信頼ファイルシステムなどを組み合わせることによって、高速性・堅牢性と低コストを両立させることとした。
  - (2)に関しては、なるべく電力効率が高く、高性能で、かつ I/O 性能が高い(少数の)CPU に大量の HDD が接続される NAS が好ましいとした。特に、Infiniband や 10GbE の大域を埋めるためには、NAS 全体で 1GB/s 以上の内部 RAID アレイの合算速度が要求される。通常の 4+1 の RAID5 アレイの I/O 速度は 100MB/s~150MB/s 程度なので、7 台から 10 台のアレイ、つまり HDD が最低でも 35 台から 50 台程度 NAS ノードに接続可能でなくてはならない。NAS 全体としては SATA チャンネルが最大で 50 本で、10Gbps (HW RAID) (SW RAID の場合は数十% ほど高い)), 相互結合網の 10Gbps の I/O 速度を加えると、2GByte/s(SW RAID の場合は 3GByte/s)以上の I/O スループットが維持される必要がある。

    また、NAS はプロトコル処理などで高い処理能力が必要とされることが多い。今回のように 50 台近い HDD が接続される場合はなおさらである。逆に HDD の数が多ければ、より消費電力が高い高速 CPU を用いても、平準化されるので、低電力 RISC プロセッサを用いた NAS より好ましい。

  以上をまとめると、50 台近い HDD を搭載可能で、かつそれに見合った処理能力と上記の I/O スループットを実現する NAS を必要としたが、通常の 3U の NAS では HDD の数がせいぜい 14 台程度であり、到底この要求を満たさなく、問題となった。

- 低消費電力・高密度アクセラレータによる計算の加速:
  現状の 64bit の x86 系の CPU の性能は、SpecFP2000 などを鑑みても、数倍から十倍以上もコストがかかるハイエンド RISC プロセッサとほぼ遜色が無くなっている。それどころか、SpecINT2000 においては、x86 の優位性は圧倒的である。また、近年 HyperTransport, PCI-Express などによって、I/O 速度も劇的に向上しており、結果としてインタコネクト性能が上がっている。これらを組み合わせることにより、複雑度の高い並列アルゴリズムでも、従来のスパコン以上に高効率に実行させることが可能となっている。

  しかしながら、近年 CPU の省電力化が最重要課題の一つとなり、確かに FLOPS/W などの指標は向上し続けているものの、100TeraFlops 級のスパコンインフラは、2006 年 3 月の段階で 2-3MW 級となることが予想された。実際、調達後の TSUBAME の実測値では、最大負荷想定の Linpack 実行で、1 ノードあたり 1300W 程度、Opteron 部分だけの 50TeraFlops の合算では 850KW 程度である。これにストレッジや空調を加えると、100TFlops 級では約 2MW となってしまう。

  そこで、一部のワークロードでは FLOPS/W 値が著しく

高い SIMD ベクトル型のアクセラレータを用いることにより、マシンサイズを著しく大きくせずに、性能加速を可能とすることを検討した。近年、グラフィックス系や組込みマルチメディア形は言うにおよばず、GRAPE[5]など、アプリケーション用途をかなり限定すれば、素晴らしい性能を示すものが実際に出ている。問題は、「みんなのスパコン」という、広いユーザベースのコンテクストでは、あまりにも用途や使い方が限定されるものは(少なくとも当初は)、一般ユーザにとってのメリットが少なくなってしまうことにあった。具体的な要件としては：

1. BLAS、FFTW など、一般のユーザが多用する IEEE754 準拠の倍精度の数値ライブラリがユーザから透過に加速されること
2. Matlab, Amber など、幅広ユーザベースを持つ商用や public domain のアプリケーションが加速されうること
3. それら以外の用途でも、先進的なユーザが自分のプログラムのコア演算部分を SIMD ベクトル化することによって、汎用的に加速できうること。特に、そのような汎用のプログラミング環境や SDK が提供されること

とした。得に、1, 2 の用途は即効的な効果を期待するのはもちろんであるが、実際はアクセラレータの適合範囲は汎用 CPU と比較して狭いので、複雑な問題でスケーリングさせるのは逆に難しい、という事態がある。逆に、アクセラレータが得意な分野で、通常の CPU を用いるのは、様々な効率面で無駄が多い。従って、アクセラレータで容易にスケーリングされるような状況は、小さなスケールの問題などでむしろそれを積極的に活用してもらい、汎用 CPU のパワーが必要な複雑な問題のスケーリング時に、いわば「どいて」もらうことが第一義的な価値とした。

無論これは、アクセラレータを用いて大規模な問題にスケーリングさせることを否定していない。将来プログラミングモデルや SDK,アルゴリズムが進歩すれば、そのようなスケーリングは可能であるし、また、実際に後述の Linpack などにおいて、そのような研究を進めている[]。ただ、当座は用途 1,2 において、小さいスケールだが加速がよく効く問題領域への適合の方が、TUBAME 全体としてのメリットが大きい、と判断しているのである。

2. TSUBAME の誕生

以上のような設計要件の元に、TSUBAME は NEC が Sun Microsystems, AMD 社などと協業してその詳細設計や製造・設置の任にあたり、2006 年 3 月末日に誕生し、翌月の 4 月 3 日から第一次の試験運用を開始した。その概要は図 1 に示している。技術的なポイントとしては、以下の項目が挙げられる：

- 5120 ソケットに、Dual Core CPU 採用による 10480 個の AMD Opteron コア、2.4Ghz (一部 2.6Ghz)。
- 8 ソケットを Coherent HyperTransport で接続し、16 core のほぼ SMP に近い共有メモリ・Fat Node を 4U のコンパクトなサイズで実現したサーバマシン Sun X4600 が 655 ノード、約 50.4TeraFllops
- メモリは 1 ノード 32GB (一部 64GB)で合計 21.4TByte (地球シミュレータの約 2 倍)である。これにより、チューニングやライブラリ・コマンドの細かい差異は別として、地球シミュレータで動作するプログラムはほとんど TSUBAME で動作する（そのままのアルゴリズムでもメモリ不足に至らない)。
- 相互結合網は、Infiniband 4x の Dual Lane 構成で、1 ノードあたり 20Gbps である（ストレッジの Thumper は single lane)。スイッチは 288 ポートの Voltaire ISR 9288 で、エッジスイッチとして 6 台、コアスイッチが 2 台である。エッチスイッチとコアスイッチは 24 本ずつのアップリンク(合計 48 本)で結ばれているので、バイセクションバンド幅は(1.44Tbps + 1.44Tbps)である。
- ストレッジは、二つのシステムにより構成されている。一つは Sun の Thumper システム群で、一台あたり 4U で 48 台の 500GB SATA HDD を備え、合算容量 24TByte を備える。ストレッジコントローラは Dual Core / Dual Socket Opteron であるが、ストレッジ用に HyperTransport 経由で多数の SATA ブリッジが接続されており、48 台分に十分な I/O 速度を確保している。TSUBAME では、Single Lane の Infiniband で接続して、RAID5 構成で全体で 1GB/s 以上のシーケンシャルアクセスの外部バンド幅を達成している。TSUBAME では全部で 42 台で、全容量 1PByte, 全ストレッジバンド幅 40GByte/s となるが、その高速性を生かすため、/work ディレクトリでは Lustre 並列ファイルシステムを動作させ、並列化による高速 I/O を実現している。今後、それ以外にもバックアップや、WebDAV など、様々な一般ストレッジサービスにも活躍する予定である

  二つ目は NEC の iStorarge S1800AT システムであり、SATA を用いながら、RAID6+冗長 SAN 構成により、高い信頼性を確保している。全容量は 100TByte で、主に /home に用いられている。
- アクセラレータは、ClearSpeed 社の SIMD ベクトル型のチップ CSX600 二個づつ搭載した PCI-X のボードを用いた(図 1)。詳細は[6]に譲るが、一枚あたりのピーク性能は 96GFlops、ピーク消費電力は約 25W で、現状では BLAS の最高性能が 50GFlops であり、将来さらに改善の予定がある。全体のボード数は現状では 360 枚で、合算ピーク性の葉 34.6TFlops であり、カードとメイン CPU 間の転送速度を鑑みて、1 ノード(一部)ごとに 1 枚づつ装着されている。全体の消費電力は 9KW 以下（TSUBAME の 1 ラック以下)。

  ClearSpeed の利用法は前述の利用モデルを想定しており、大規模な並列計算への適用はこれからだが、今後の HPC の方向性として、高い期待を抱いている。
- TSUBAME の設置面積はサービスエリアを含めて約 350m2 (GSIC センターの計算機室全体で 600m2 程度)であり、ストレッジ(10 ラック)やネットワークを含め 76 ラックで構成される。計算ノードはラックの高さによって 10 台または 11 台格納され、スイッチを含めて 65 ラックを占め、残りがストレッジや制御・管理ノード、外部接続ネットワークなどのラックである。これに、32 台の冷却ユニットが加わる。全重量は冷却器を含めると約 60 トンであり、2F と 1F の合計 3 部屋に跨って設置されている。

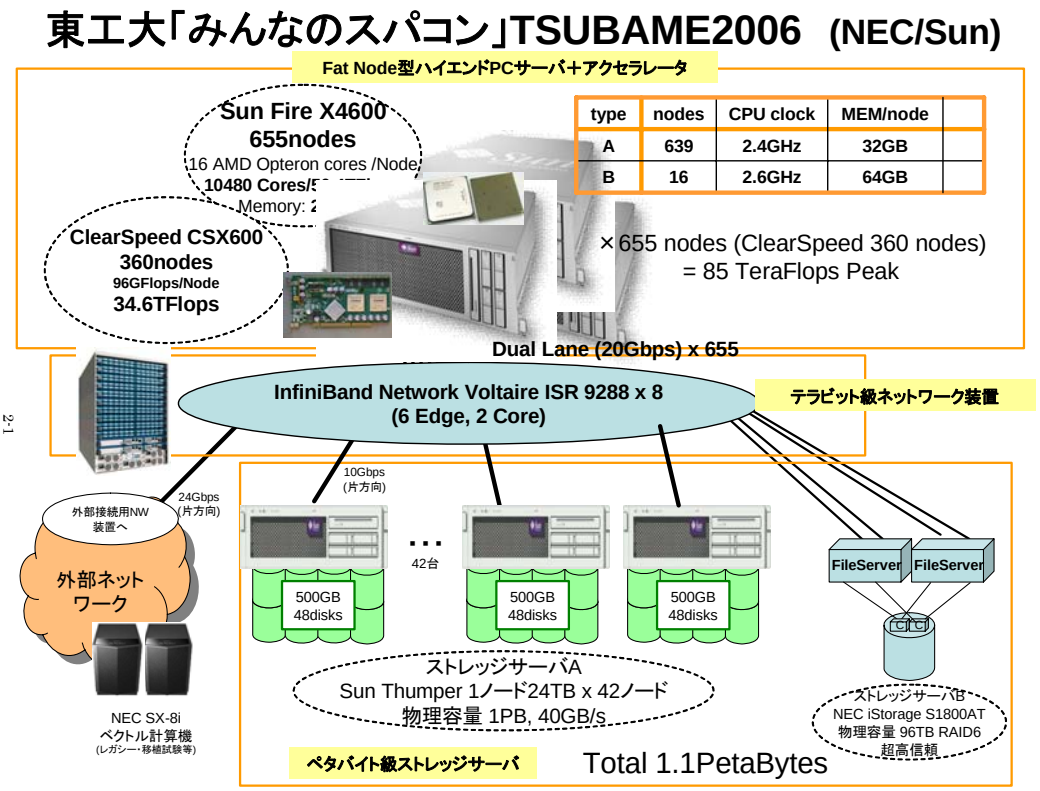

図 1: TSUBAME の概要構成

- TSUBAME の消費電力は、定格では 1153kVA、冷却装置を含むと 1630kVA であるが、実際の観測地では High Performance Linpack 動作時、かつ冷房器を最大パワーで動作させて約 1.2MW 程度である(5 月ごろ)。実際、25 度吸入における単体ノードの HPL 実行における省電力は 1330W と計測されており、よって TSUBAME の計算部分の実際的な最大消費電力は 900KW 以下と推定される。現在、Opteron の DVS による省電力機能 PowerNow! をオンにすべく、追加の電力計測や、冷房器の出力のチューニング・ダクト配置の向上・ラック間隔壁の追加、などを行っている。
- 詳細は別項に譲るが、TSUBAME が約 480 ノード置いてある 2F のメイン計算機室は、最新のデータセンターで用いられる、吸気側の冷気と排気側の暖気とを完全分離(Cold Row vs. Hot Row Separation)して、かつ暖気を滞留させずにすぐに冷却装置で吸気して冷やし、高圧力で吸入側に冷気を排出することによってが高い冷却効率が得られる新しい手法をとっている[7]。こにに対し、旧 GSIC を含むわが国の多くのスパコンセンターが、旧来スパコンのレガシーである冷暖気の攪拌による部屋全体の平均温度を下げて手法を用いているが、この手法は特に高い熱源がある場合は大変非効率であることが知られている。TSUBAME では、新冷却方式実現のために、冷却装置を部屋の真中の暖気列のすぐ後ろにも設置し、天井のダクトを通して暖気と混合せずに吸入列に冷気を供給するという、専用の配置設計・実装を行っている。
- TUBAME の OS は Linux Suse 9 を現状では全ノードで用いており、各ノードには Boot & Swap 用の 2.5inch HDD を装備している。管理&ジョブスケジュリングは、Sun N1 Grid Engine を用いているが、将来これは NAREGI のグリッドミドルと接続され、かつ Condor [www.cs.wisc.edu/condor/] などの他のジョブスケジュラなどと混合される予定である。

  無論 TSUBAME の心臓部は x86-64bit ベースのアーキテクチャなので、将来は Suse 10 へのバージョンアップは言うに及ばず、状況に応じて Sun Solaris や Windows Cluster Computing Server を部分的に採用することも検討している。それらの運用の際には、OS イメージベースの動的なプロビジョニング・ディプロイメントツールや、仮想機械技術による OS ホスティングなどによる実現を果たす予定である。

以上のように、TSUBAME は初期の設計思想をほぼ反映したマシンとなっている。

3. TSUBAME の飛翔 – ベンチマークと初期運用

平成 18 年 7 月 3 日現在、TSUBAME は The 27th Top500 の Linpack 性能で、わが国およびアジアー、世界では 7 位である 38.18TFlops（を達成し、その絶対性能や容量もあいまてって、アジアの代表的なスパコンとなった[6]。また、Rmax/Rpeak 効率も専用設計のスーパコンピュータに比肩する 76. 56%を達成した。わが国のスパコンが Top10 に入るのは珍しいことではなく、むしろ Top500 の開始以来何らかのマシンが必ず Top10 入りしていたが、昨今の大幅にレベルが上がった状況で今回の成績が達成できたのは一つの成果と言えよう(10 位が地球シミュレータ)。

現状では TSUBAME の運用範囲は段階的に拡大し、現状では約 300 ノード、4800CPU 近くが一般運用に供され、稼働率は 60-80%程度である。残りの約半分強のノードは、地球シミュレータの大規模並列コードなど、一部の先進的なユーザのコードを 1000CPU 級にポーティング・チューニングしたり、HPL を含む ClearSpeed の利用実験に供されたり、一般のベンチマークを行うなどの、様々なレベルでのテストに日々用いられている。今後テストやベンチマークを繰り返し、電力や冷却を含む運用体制のファインチ

図 2: TSUBAME の一部

ューニングを行ったうえで、10 月を目処に、一部のテストノードを除いてほぼ全ノードを学内全体、および学外との共同利用に公開する予定である。

また、TSUBAME は今後様々なベンチマークを行い、その結果を一般に広く公開していく予定である(http://www.gsic.titech.ac.jp/~ccwww/tgc/bm/)。現状では HPL は勿論だが、MPI の各種 microbenchmark, HPC Challenge Benchmark (約半数のノード)(http://icl.cs.utk.edu/hpcc/)、NPB 2.3 (Class B/C), Gaussian、Amber, などであり、今後も増加する。ここでは、いくつか特徴的なデータを紹介する。

- MPI の通信性能は、ノード間でのレーテンシは 4.7~5.7μ秒程度、バンド幅は 1.10~1.15GB/s、B1/2(最高バンド幅の 1/2 の性能を示すデータ長)が 16384~32768 バイト程度で、ペアとなるノードの配置にかかわらずほぼ安定した性能を示す。これは、SGI Columbia, Cray Redstorm、BG/L などの同規模のシステムと同程度の数値である。
- NPB Class C では、ほぼ全てのベンチマークにわたって、ノードを分散させて 128CPU 程度まで計測を行ったが、CPU 増加に伴う Mops/s/CPU はほぼ一定であった。つまり、これは TSUBAME のネットワークが高速でスケーラブルであることを意味している。一方、16CPU までノード内でスケールさせると、低下が見られた。この原因はまだ定かではないが、メモリ周りの不適切なアロケーションに伴うバンド幅制約によるオーバーヘッドが原因として考えられる。
- Gaussian ベンチマークは、test397 で Linda 並列化で最大 1024CPU まで試みた。これは筆者の知る限りでは今までで最大の CPU 数である。結果として、16CPU と比較して、48CPU までは比較的良くスケールするが 2.02 倍)、それ以降は加速がなまり、256CPU までスケール(2.78 倍)したあと、それ以上の CPU 数では速度低下した。これより、(より原子数の多い)高分子においては、従来予測されていたより多い数十 CPU 以上でのスケーリングが強く期待される。
- 38.18TFlops Linpack における実行時間は約 11 時間半である。これは、Rmax, Nmax から推定される Top500 上の上位マシンのどの実行時間よりも長い(Top10 では Columbia の約 7 時間 40 分が次点)。この理由は様々に考えられるが、少なくとも TSUBAME は Top500 の中で歴史上もっとも安定に Linpack を実行したマシンであると言えよう。

4. おわりに– TSUBAME は「みんなのスパコン」へ

TSUBAME は単純に従来型のスパコンではなく、次世代の計算科学者を培う教育から日常の研究、さらには事務系のサービスホスティングまで、教育研究期間としてのあらゆる IT ニーズの中心となるべく、「みんなのスパコン」体制を目指す新たな利用法・活用法を実施しようとしている。現状では、同じく 2006 年 4 月より導入された「キャンパス共通認証・認可システム」の認証ポータルへの接続が試験的に行われており、全学へのアカウント付与と SSO が実現される。実はこの認証システム自身、一部 TSUBAME の資源を用いたものである。さらに、OCW (Open Courseware)、全学ストレッジサービス、Windows のリモートデスクトップなどが、TSUBAME 上で仮想マシンなどの技術を用いてホスティングされようとしている。スパコンという観点から鑑みれば、これらが利用する資源は全体では微々たるものだが、全学の IT 資源の TCO の削減、さらには情報サービスの集約化や均質化による利便性の向上の効果は大変大きいと期待している。

最後に、TSUBAME の今後の発展の計画を進めている。技術的には 2008 年ごろに 300TFlops 級や、数ペタバイトのマシンへ、現在の設置面積・消費電力・運用体制を維持しながらアップグレードするのは容易であり、あとは投資次第である。さらに、TSUBAME の設計・契約上の「寿命」は 2010 年 3 月までであるが、その後継としては、1PFlops 級のマシンを考察中である。

謝辞

TSUBAME に「関わった」人数は本学トップから GSIC や事務方までの各関係者・ユーザは言うに及ばず、各ベンダーの多数のの方々、文部科学省、さらには常日頃 HPC に関して議論・共同研究している方々まで、その数は軽く見積もっても 100 人単位となる。紙面が足りないので個々人への御礼ができないのが残念であるが、ひとまずこの場を借りて深く御礼を申し上げ、今後も TSUBAME の発展へのご協力をお願いさせていただければ大変幸いである。

参考文献

[1] Thomas Sterling et. al. *How to Build a Beowulf: A Guide to the Implementation and Application of PC Clusters,* The MIT Press, May, 1999.

[2] IDC Worldwide Quarterly Server Tracker Press release 1Q 2006, IDC, http://www.idc.com/getdoc.jsp?containerId=prUS20180706, May, 2006.

[3] 東工大キャンパスグリッドプロジェクト、http://www.gsic.titech.ac.jp/TITechGrid/

[4] Cluster Filesystems Inc. Whitepaper, "*Lustre: A Scalable, High-Performance File System*", Nov. 2002., http://www.lustre.org/docs/whitepaper.pdf

[5] Junichiro Makino, et. al. "A 1.349 Tflops Simulation of Black Holes in a Galaxy Center on GRAPE-6", Proc. IEEE Supercomputing 2000, IEEE Press. Nov. 2000.

[6] 遠藤敏夫, 長坂真路, 後藤和茂, 松岡聡: 「ヘテロ型スーパーコンピュータ TSUBAME の Linpack による性能評価」情報処理学会ハイパフォーマンスコンピューティング研究会予稿集 (SWoPP2006), 2006 年 8 月(発表予定).

[7] Chandrakant D. Patel et. al. "Thermal Considerations in Cooling Large Scale High Compute Density Data Centers", Itherm2002---8th Intersociety Conf. on Thermal and Thermomechanical Phenomena in Electronic Systems, San Diego, 2002.