SSTC

High Performance and Productivity

スケーラブルシステムズ株式会社

ハイパフォーマンス|パーソナルクラスタ

スケーラブルシステムズ株式会社

ハイパフォーマンス|パーソナルクラスタ

はじめに 2

マイクロプロセッサのマルチコア化がもたらすパラダイムシフト 3

パーソナルクラスタ 4

パーソナルクラスタの特徴と利点 6

パーソナルクラスタの利用方法 7

パーソナルクラスタの製品紹介 9

おわりに 11

スケーラブルシステム株式会社では、IT 技術とHPC システムに関する様々な調査レポートを発行しています。
ご購入の際は（Tel：03-5875-4718 E-mail：biz@sstc.co.jp ）までお問い合わせ下さい。

2006.8.10
2008.10.10 (改訂)

はじめに

シュミレーションやデータ処理のために利用される目的で構築され、利用されるHPCシステムは、様々なアーキテクチャで構築されてきました。現在では、マイクロプロセッサとそのマイクロプロセッサを複数搭載することで、HPC システムを構築することが一般的となっています。このようなHPC システムは、その利用目的や投資規模に応じて、1）数千プロセッサを同時に利用し、数テラバイトのメモリを利用するような高度なシュミレーションを行う、大規模なデータセンターに設置されるスーパーコンピュータのようなシステムから2）部門や個人で利用される小・中規模のサーバやクラスタシステムまで、非常に多岐にわたる利用形態があります。

HPC の二極分化： 基盤技術やコアの IT テクノロジの共通化を図りながら、この極端に分極化したHPCシステムへの対応を図ることが必要となります。HPC で要求されるシステムの仕様が、大きく分極化し、この双方を一つのシステム・アーキテクチャで実現するのは、技術的に可能だとしても、経済性や生産性の点で問題があるためです。

実際にHPCシステムは、大規模なスーパーコンピュータも部門や個人で利用される小・中規模のサーバやクラスタシステムでも、その基盤とするテクノロジーには共通のものが多々あります。これらの基盤技術をベースとして、それぞれの用途向けに最適化や特別の技術を追加しています。現在の HPC システムは、大規模なシステムは、より大規模な構成になる傾向があり、これらの大規模なシステム構築のための技術開発などが続いています。一方、個人や部門で利用されるHPC システムのニーズも広がりを見せています。その意味では、個人や部門で利用される HPC システムはより広範囲に利用が広がっており、大規模システムは、特定のユーザ向けにシステムの規模が拡大しています。

マイクロプロセッサのマルチコア化がもたらすパラダイムシフト

コンピュータシステムの導入、維持・管理などにかかる費用の総額としてのTCOを考え、それを如何に低減するかが、多くのユーザの間で注目されています。コンピュータシステムのコストは製品価格(導入費用)で評価されることが多かったが、近年のコンピュータシステムの複雑化や製品価格の下落などにより、コンピュータシステムの維持・管理やアップグレード、ユーザの教育、システムダウンによる損失など、導入後にかかる費用(ランニングコスト)が相対的に大きな存在となっています。このうち、電力とシステムの冷却などの空調設備などのランニングコストは、非常に大きな比重を TCO の中で占めています。この TCO の削減のための一つの解決策として、マイクロプロセッサのマルチコア化とよりエネルギー効率に優れたマイクロアーキテクチャが提案され、製品化が進んでいます。マイクロプロセッサのマルチコア化によって、プロセッサは従来の動作周波数の向上による性能向上以上に、その性能を向上させる可能性を持つことになります。高いピーク性能をより少ない消費電力と空調設備で実現でき、また、高いエネルギー効率はよりコンパクトな筐体を可能としています。

このようなマルチコアプロセッサの製品化とよりエネルギー効率を実現したシステムアーキテクチャが現実になることによって、従来とはすこし異なったコンセプトでの HPC システムの構築とその利用方法の提案も可能となります。このような新しいコンセプトの一つとして、「パーソナルクラスタ」というシステムが今、注目されています。マイクロプロセッサのマルチコア化と64ビット化は、そのエネルギー効率の改善によって、従来のコンピュータルームに設置されていたクラスタシステムをより身近に、また、パーソナルな用途で利用可能なシステムとして、構築することを可能としています。

ここで、HPC 分野でのクラスタの開発の歴史を見てみます。クラスタシステムは、最初は、パーソナルコンピュータ(PC)をネットワークで接続し、並列アプリケーションを実行することを目的として提案されました。その後、パーソナルコンピュータだけでなく、処理性能をより重視したサーバシステムをより高速のネットワークや専用のインターコネクトで接続することにより、より高い性能と大規模なシステムでのスケーラビリティを実現しています。一方、パーソナルコンピュータ自身も進化し、現在では、サーバシステムと同じようにマルチコアプロセッサを搭載し、非常に強力なコンピュータシステムとなっています。このクラスタと PC の 2 つのシステムの利点を持ち、また、ユーザの PC での利用環境を自然な形で発展させたものがパーソナルクラスタです。

PC はコンピュータに大きな変革をもたらしました。今、新しい PC - **P**ersonal **C**luster が HPC システムを大きく変える可能性があります。

Hrothgar - 1995
- 16 Intel Pentium100 MHz
- PCI
- 1 Gbyte memory
- 6.4 Gbytes of disk
- 100 base-T Fast Ethernet (hub)
- 240 Mflops sustained

Hyglac-1996 (Caltech)
- 16 Pentium Pro 200 MHz
- PCI
- 2 Gbytes memory
- 49.6 Gbytes of disk
- 100 base-T Fast Ethernet (switch)
- 1.25 Gflops sustained

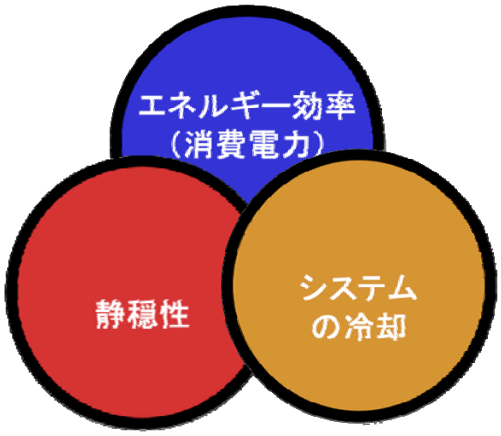
初期の Beowulf クラスタ: PC を‘ラック’に重ねて設置し、ネットワークで接続したものですが、余剰 PC を利用することで、システムのコストを低減し、非常に高い価格・性能比を実現しています。また、MPI などの標準化された API が利用できるようになり、アプリケーションの整備が急速に進むことで、多くのユーザがこのようなシステムの構築を行うようになってきました。

パーソナルクラスタ

クラスタシステムを PC のように利用する、又は、PC の特徴を持つクラスタシステムを考えてみましょう。この考察を行う前に、現在の PC とクラスタに関して、いくつかの点を議論したいと思います。

- PC のプロセッサとサーバのプロセッサの差異が無くなりつつあります。特に、インテル社のマイクロプロセッサの場合、モバイル、デスクトップ、サーバのすべての用途向けのマイクロプロセッサの基本アーキテクチャを統一して、すべてのプラットフォームでのマイクロプロセッサは、インテル Core マイクロアーキテクチャに基づく製品になっています。
- PC でも 64 ビット化が進み、マイクロプロセッサ、OS、アプリケーションの 64 ビット版を利用できるようになっています。サーバは 64 ビット、PC は 32 ビットという区分けは、今後、急速に無くなっていきます。

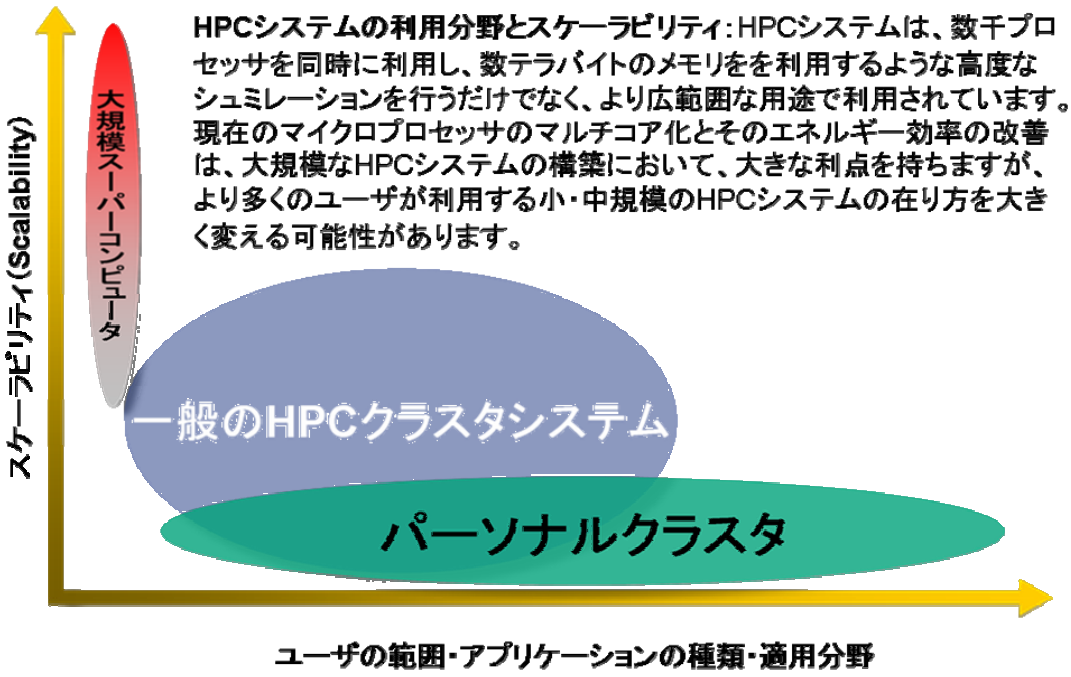
- 現在の IT インフラにおける大きな問題として、その消費エネルギーとシステムの冷却(と設備)の問題があります。クラスタシステムでは、もう一つ、その騒音という問題があり、専用のコンピュータルームなどで稼動させる場合には、特に問題にはなりませんが、一般のオフィス環境で稼動させた場合、利用環境の悪化が懸念されます。



パーソナルクラスタに求められるもの: クラスタを PC のように利用しようとした場合、一般のクラスタシステムよりも更に、静穏性や電源、冷却が問題になります。

- サーバシステムの高密度実装は、Blade 型のマザーボードを利用することで、従来からも可能となっています。ただ、Blade サーバでは、あくまで、コンピュータシステムの構築要素(Building Block)をその設計基盤としているため、Blade サーバ単体をオフィス環境で利用するようなことは、一般的ではありません。

TCO を考えた場合、電力と冷却の問題は、今後、更に深刻化します。その意味では、計算機の利用技術の変更によって、このような問題に対応することも必要になります。同時に、従来のパーソナルコンピュータの延長として、HPC システムを考え、より使い易いシステムの構築を目指すことも必要です。初期の Beowulf クラスタの構築では、大規模な投資と専従の運用管理者を必要とするスーパーコンピュータシステムではなく、自分達で自由にスーパーコンピュータクラスのシステムを構築し、利用することを目的としていました。現在のクラスタシステムは、構成要素としては、一般の商用製品を利用し、コストを下げて、処理能力の向上と同時にそのコスト・パフォーマンスの向上が図られています。HPC システムとして、従来のスーパーコンピュータと同じような位置に置かれています。もちろん、このような大規模なシステムは、今後も重要なコンピュータシステムであり、更に技術革新が求められ、進化していきます。しかし、このようなスーパーコンピュータシステムとは別に、新しいコンセプトで、スーパーコンピュータクラスの性能と PC のような利用環境を提案するのが、パーソナルクラスタです。



HPC システムの利用分野とスケーラビリティ： HPC システムは、数千プロセッサを同時に利用し、数テラバイトのメモリをを利用するような高度なシュミレーションを行うだけでなく、より広範囲な用途で利用されています。現在のマイクロプロセッサのマルチコア化とそのエネルギー効率の改善は、大規模な HPC システムの構築において、大きな利点を持ちますが、より多くのユーザが利用する小・中規模の HPC システムの在り方を大きく変える可能性があります。

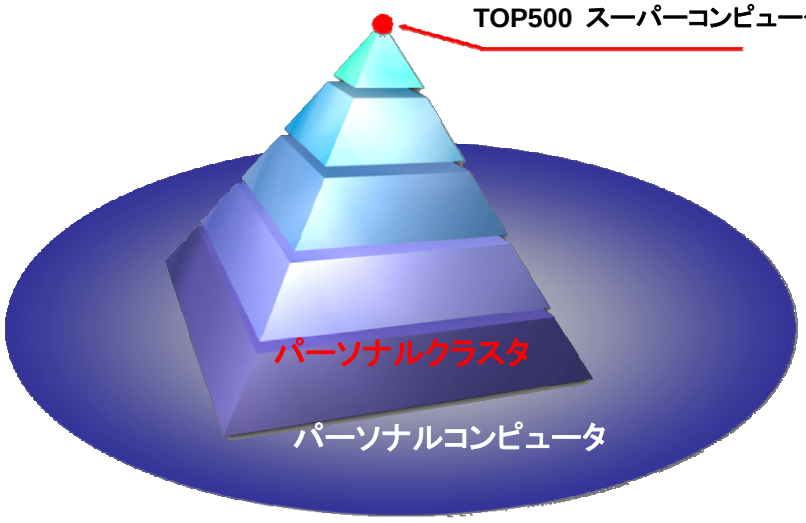
パーソナルクラスタの特徴と利点

パーソナルクラスタをHPCのワークロードに対して利用する場合、多くの利点があります。HPCシステムに要求される高い処理性能とコスト低減の課題に対するソリューションを提供するものです。

- 設置場所の制限が少ない
- 必要とする電源や冷房設備がオフィス環境でも対応可能
- HPC システムは、一般にはシステムの規模が大きくなるとその利用効率に急激に低下します。高いピーク性能に見合う実効性能を示すことが難しくなります。パーソナルクラスタでは、従来のクラスタ以上の高い性能を示すことが可能となります。
- HPC システムは、24 時間稼動させるような運用が一般的です。このような運用は、システムがアイドルしていても、また、特にジョブの実行を予定していない場合にも、システムを稼動さ

せておくことになります。このような運用は、エネルギーの利用効率としては、あまり良いことではありません。パーソナルクラスタでは、PC と同じように必要なときだけ、電源を入れて利用することが可能であり、TCO の削減に貢献します。

- マイクロソフトは、Windows Compute Cluster Server 2003 で、初めて HPC 向けの製品をリリースします。パーソナルクラスタは、この Windows CCS のプラットフォームとして最適です。マイクロソフトは、パーソナルクラスタでの Windows CCS の活用に関する様々な資料を提供しています。Windows CCS が目指すマーケット、利用形態、対象マーケットは、全てパーソナルクラスタのコンセプトに合致するものであり、双方の利点を最大限に発揮することを可能とします。
- パーソナルクラスタは、用途と構成を明確にしたシステムであり、クラスタシステムで問題となる‘システム構築’の問題がありません。この点は、システムの拡張性やシステム構成の選択肢の少なさとして、問題と思われるかもしれませんが、逆に構成を制限することで、更にコストの削減を図ることを可能とし、また、運用やシステム構築の負担を大幅に低減します。クラスタシステムでは、実際には、このような運用コストと管理の面倒さが大きな導入の障害になっています。
- 実際には、パーソナルクラスタ自身を‘ノード’として、より大きなシステムを構成することも可能です。パーソナルクラスタのような小規模で運用が容易なクラスタを初期の Beowulf クラスタが PC を計算ノードにしたように構成し、より大規模なクラスタを構築することは、クラスタの構成を容易にし、TCO の低減に大きく寄与します。



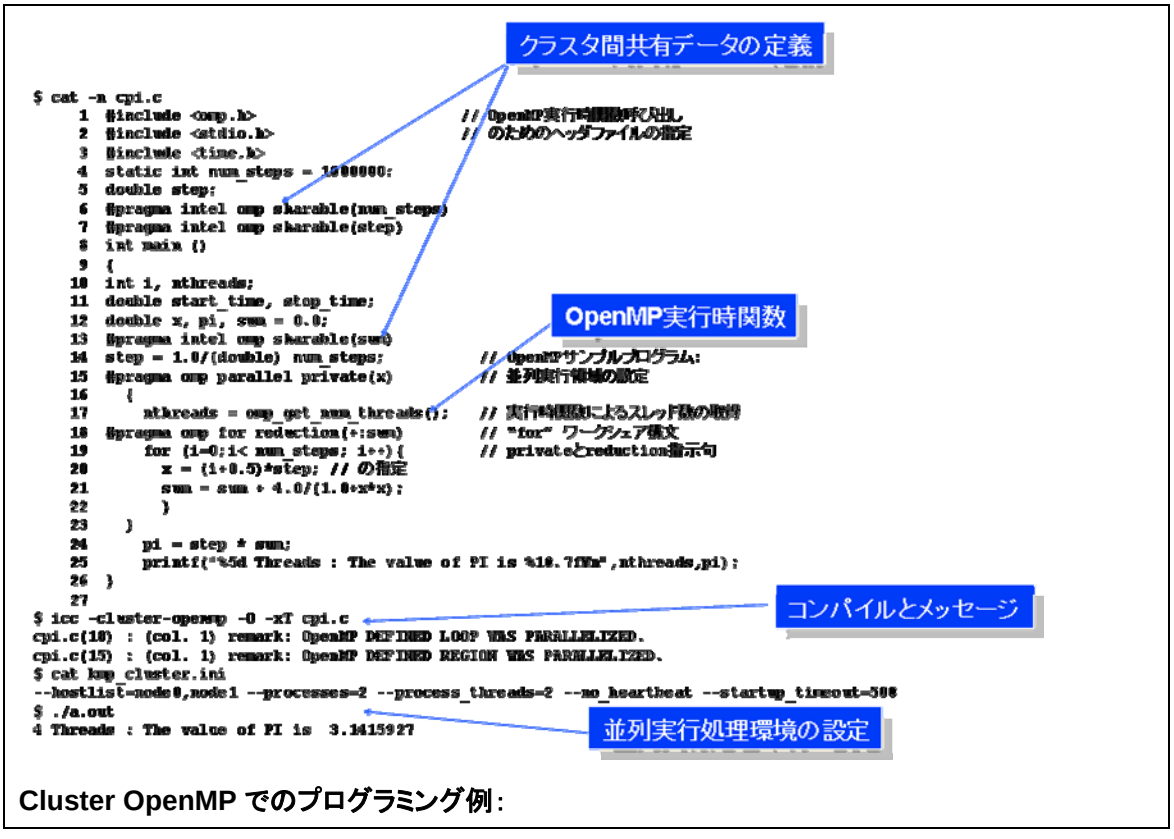
パーソナルクラスタの位置づけ： パーソナルコンピュータの発展として、クラスタシステムをより身近なものとし、HPC ワークロードのより効率的な処理を目指します。

パーソナルクラスタの利用方法

パーソナルクラスタは、PC と同じように、デスクサイドで利用することが可能です。高性能の PC を端末として、利用することも可能ですし、このような構成の PC をヘッドノードとして利用することも出来ます。また、パーソナルクラスタの一つのノードをヘッドノードとして利用し、端末として利用することも可能です。このように利用方法の自由度が高いこともパーソナルクラスタのメリットです。もちろん、パーソナルクラスタを従来のクラスタと同じように、コンピュータルームに設置して、ネットワークで利用することも可能です。

パーソナルクラスタは、Windows Compute Cluster Server 2003 でも様々な Linux のディストリビューションで利用可能です。先に示したように、Windows Compute Cluster Server 2003 とパーソナルクラスタは、ベストな組み合わせとなることでしょう。もちろん、Linux では、スケーラビリティや最新のマルチコアプロセッサのサポート、コンパイラを含む豊富な開発環境とオープンソースの最新テクノロジーが利用可能です。

インテルの最新コンパイラ（バージョン 9.1）では、Cluster OpenMP が利用可能です。Cluster OpenMP は、従来の共有メモリ上でのマルチスレッドプログラミングのための API である OpenMP をクラスタ環境に拡張しています。パーソナルクラスタでは、この Cluster OpenMP をプログラミングのオプションとして利用可能です。




```
$ cat -n cpi.c
     1  #include <omp.h>                          // OpenMP実行時関数呼び出し
     2  #include <stdio.h>                        // のためのヘッダファイルの指定
     3  #include <time.h>
     4  static int num_steps = 1000000;
     5  double step;
     6  #pragma intel omp sharable(num_steps)
     7  #pragma intel omp sharable(step)
     8  int main ()
     9  {
    10  int i, nthreads;
    11  double start_time, stop_time;
    12  double x, pi, sum = 0.0;
    13  #pragma intel omp sharable(sum)
    14  step = 1.0/(double) num_steps;            // OpenMPサンプルプログラム:
    15  #pragma omp parallel private(x)           // 並列実行領域の設定
    16    {
    17      nthreads = omp_get_num_threads();     // 実行時関数によるスレッド数の取得
    18  #pragma omp for reduction(+:sum)          // "for" ワークシェア構文
    19      for (i=0;i< num_steps; i++){          // privateとreduction指示句
    20        x = (i+0.5)*step; // の指定
    21        sum = sum + 4.0/(1.0+x*x);
    22        }
    23    }
    24      pi = step * sum;
    25      printf("%5d Threads : The value of PI is %10.7f\n",nthreads,pi);
    26  }
    27
$ icc -cluster-openmp -O -xT cpi.c
cpi.c(18) : (col. 1) remark: OpenMP DEFINED LOOP WAS PARALLELIZED.
cpi.c(15) : (col. 1) remark: OpenMP DEFINED REGION WAS PARALLELIZED.
$ cat mp_cluster.ini
--hostlist=node0,node1 --processes=2 --process_threads=2 --no_heartbeat --startup_timeout=500
$ ./a.out
4 Threads : The value of PI is  3.1415927
```

Cluster OpenMP でのプログラミング例:

パーソナルクラスタの製品紹介

既に、パーソナルクラスタを製品化している会社も数社あります。その中でも Ciara Technologies[1] が開発したNEXXUS 4000は、パーソナルクラスタとして、非常にユニークな特徴を持ちます。この NEXXUS 4000 は、2005 年の「Supercomputing 2005」のマイクロソフトのビルゲイツ会長の基調講演のデモでも紹介されています。[2]

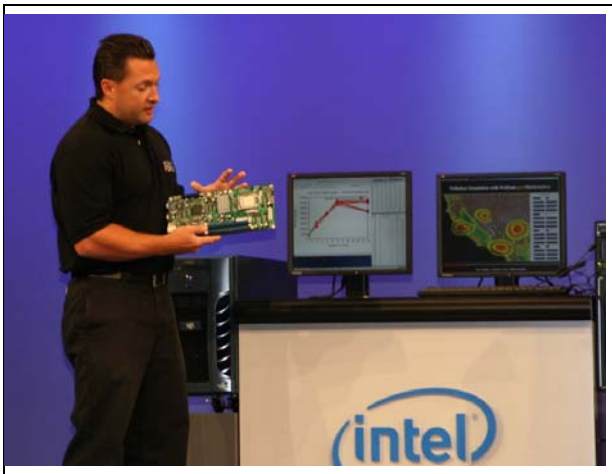
NEXXUS 4000: デスクサイドにも設置できる HPC システム

NEXXUS 4000 は、デスクサイドに設置できる大きさの筐体に最新のインテルの Xeon.プロセッサとCore 2 Duoプロセッサを最大 8 プロセッサ（16 コア）搭載であり、メモリも最大 128GB まで、ディスクは最大 4TB まで拡張可能です。

通常のサーバをベースとするクラスタでは、一つのマザーボードに複数のスロットがあり、ノードは、マルチコア・マルチプロセッサ構成になります。NEXXUS 4000では、このようなデュアルソケットをサポートするマザーボードだけでなく、シングルソケットのマザーボード（インテル S3000PT）を利用したシステム構成も可能です。マルチソケット構成のシステムの場合、MCH が処理を行うデータ量が多くなり、それらデータ処理がボトルネックになる可能性が高いことも事実です。シングルソケットでは、プロセッサとメモリをより密に接続することを可能とします。シングルソケットのマザーボードは、そのコンパクトな形状のため、通常の1U サイズのノードであれば、2 枚以上搭載可能なため、ノード上のソケット数は、デュアルソケットを搭載したノードボードと同等以上の実装密度になります。

NEXXUS 4000 では、また、ギガビットイーサネットのスイッチを最初から搭載しています。このスイッチを利用して、ノード間の専用のネットワークを構築することも可能です。また、搭載されるスイッチとしては、InfiniBand のオプションも用意されており、より高速なインターコネクトを必要とする場合には、このオプションの選択も可能です。

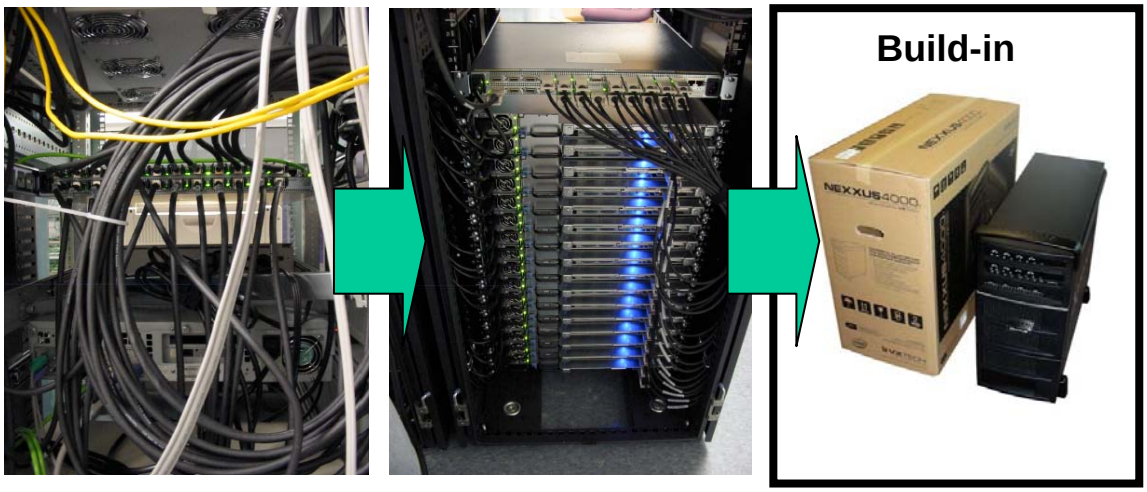
[1] http://www.vxrack.com
[2] http://download.microsoft.com/download/B/F/1/BF149A37-F0E2-4E2C-B7D2-2A93DAFC02DA/KeynoteDatasheet.pdf



S3000PT “Port Townsend”: インテルデベロッパーフォーラム（IDF）で紹介される S3000PT

S3000PT を利用すれば、4U のサイズに 40 コアを搭載したサーバの構築も可能となります。IDF では、Tyan のパーソナルクラスタ Typhoon PSC の紹介などもあり、HPC におけるデスクサイトスーパーコンピュータを紹介しています。

‘PC’をラックに積み重ねて構築した Beowulf クラスタの構築に始まり、サーバを利用しより高性能な HPC システムとして、利用可能となったクラスタシステムは常に進化しています。初期のクラスタでは、ネットワークや電源、インターコネクト用の高速ネットワークなどをシステム構築のために選択し、システムを組み立てるために配線を行っています。そのため、しばしば、この配線は非常に複雑に絡み合い、また、構築作業も容易ではありませんでした。その後、InfiniBand などが一般化し、電源、ラックなどをクラスタ向けに設計し、クラスタシステムとして、製品化された HPC システムが登場し、システムの構築作業は大幅に改善されています。パーソナルクラスタは、このクラスタの進化の一つの結果として、更に容易な導入とその利用を可能とするものです。



クラスタ構築の変遷:クラスタは初期の ‘構築する’ システムから、‘構築された’ システムへの変遷し、パーソナルクラスタによって、‘配送される’ システムとなっています。

おわりに

マイクロプロセッサのマルチコア化によって、HPC システムの構築も変化してきます。 また、プロセッサの省電力化が進み、より高密度な実装が可能となっています。 このような状況で、HPC システムもよりコンパクトで、より使いやすいものが求められています。また、HPC システムとして要求されるシステムの仕様が、 大きく二極分化し、その双方を一つのシステム・アーキテクチャで実現するのは、技術的には可能でも、その経済性や生産性の点でも問題があります。基盤技術や IT インフラに関して、その共通化を図りながら、その二極分化した HPC システムへの対応を図る必要があります。 このような状況に最適なシステムとして、「パーソナルクラスタ」というコンセプトのクラスタシステムを今回ご紹介いたしました。