［「京」の後の時代を支えるスパコン］

⑤ 多数の Xeon プロセッサを用いるスパコン

南里豪志 | 九州大学

現在，スパコンの大半が Xeon プロセッサを搭載しており，しかもその中の多くが，アクセラレータを持たない，Xeon プロセッサを主要な計算資源とするシステムである．本稿では，このようなシステムが普及した背景を説明し，その特徴の 1 つである多様性について言及する．その後，実際のシステムの例として，九州大学のスーパーコンピュータシステム ITO サブシステム A をはじめ，国内外のいくつかのシステムを取り上げ，構成を比較する．また，Xeon プロセッサ搭載システムの今後についても展望する．

Xeon プロセッサを搭載したスパコンの登場と普及

Top500 リストの中の Xeon プロセッサ搭載スパコン

現在，世界のスパコンの大半に Intel 社の Xeon プロセッサが搭載されており，そのうちの多くが Xeon プロセッサを主要な計算資源としている．図 -1 は，スパコンの性能順位リストの 1 つである Top500 リストにおける Xeon プロセッサ搭載システム数の推移である．これは，リスト中のプロセッサの項目に Xeon を含むシステムのうち，Intel 社のメニーコアプロセッサである Xeon Phi を含まないものの数である．図より，2001 年 11 月に最初のシステムが登場して以降，急速に Xeon プロセッサを搭載したシステム数が増加し，ここ数年は常に 9 割を超えていることが分かる．

また，2011 年以降のシステム数については，アクセラレータの項目が None であるものとそうでないものを，それぞれ「アクセラレータなし」と「アクセラレータあり」に分けて示している．それによると，アクセラレータありのシステム数が増加傾向にあるものの，アクセラレータのない Xeon プロセッサ搭載システム数は 7 割前後で推移していることが分かる．これは，Xeon プロセッサの計算資源としての需要が依然として非常に高いことを示している．

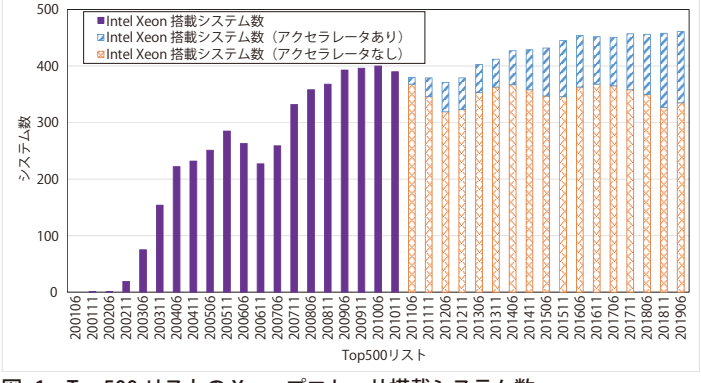

図 -1　Top500 リストの Xeon プロセッサ搭載システム数

Xeon プロセッサ搭載スパコン普及の背景

2001 年以降に Xeon プロセッサ搭載スパコンが普及した 1 つの要因は，Xeon プロセッサの高性能と高信頼性である．Xeon プロセッサは，Intel 社が 1998 年に販売を開始したマイクロプロセッサのブランド名であり，サーバやワークステーションでの要求に応える高性能と高信頼性を特徴としている．

たとえば，本稿執筆時点（2019 年 7 月）での Xeon プロセッサとデスクトップ向けの Core i9 等のプロセッサとの主な違いとして，1 台の計算機に複数のプロセッサを搭載可能であること，およびメモリのエラー訂正機能を利用できること，が挙げられる．複数プロセッサの搭載は，単なる演算性能の増加だけでなく，搭載可能メモリ量や接続可能デバイス数の増加も意味しており，より大規模な問題での利用を可能とする．一方メモリのエラー訂正機能は，メモリエラーによる運用停止の頻度を低減させるため，長期間の安定運用につながる．これらの特性から，スパコンでの利用に耐えるプロセッサとして，Xeon プロセッサの評価が高まった．

しかし，プロセッサだけではスパコンを構成できない．この時期に Xeon プロセッサ搭載スパコンが普及したプロセッサ以外の背景として，汎用高性能ネットワークの登場と，標準的なソフトウェア環境の整備が挙げられる．1995 年以降，Xeon プロセッサを搭載した計算機同士を接続するインターコネクトネットワークとして利用可能な，Fast Ethernet，Myrinet，InfiniBand，Gigabit Ethernet 等の汎用高性能ネットワークが相次いで登場した．さらに，Linux，MPI（Message Passing Interface），OpenMP という，現在のスパコンでは標準となったソフトウェア環境が，これらのネットワークで接続された Xeon プロセッサ搭載計算機群の上で利用可能となった．これらにより，共通の並列計算環境を持つさまざまな規模の Xeon プロセッサ搭載システムが世界中で導入された．それらのシステムで開発されたソフトウェアや得られた知見が，折しも利用が広がりつつあった Web を介して共有された結果，Xeon プロセッサ搭載システム利用者のコミュニティが形成され，スパコンとしての導入事例が増加したと考えられる．

Xeon プロセッサ搭載スパコンの多様性

Xeon プロセッサ搭載スパコンの重要な特徴の 1 つとして多様性が挙げられる．まず Xeon プロセッサ自体が，各世代で多様なコア数やクロック周波数の組合せから選択可能である．そのため，使用するプログラムの並列性や要求されるシングルスレッド性能，および予算や電力の制約に応じて，適切なものを選ぶことができる．さらに，システムの計算ノード数，ノードあたりの搭載メモリの種類や量，ストレージの容量や形態，計算ノード間のインターコネクトネットワークの種類や接続形状，等も，予算や用途に合わせて調整可能である．

例として，インターコネクトネットワークの種類による計算性能への影響を，図 -2 に示す．図の縦軸は，システムの理論ピーク性能（理論的に達成可能な最高性能）に対する LINPACK 性能（Top500 で用いられるベンチマークプログラムの性能）の比で，計算システムの計算効率を示す指標の 1 つである．一方，横軸は Top500 リストの順位であり，2019 年 6 月のリストの各順位のシステムの計算効率を，使用したインターコネクトネットワークの種類を示す点でプロットした．なお，インターコネクトネットワーク以外の影響をできるだけ排除するため，Skylake マイクロアーキテクチャ以降の Xeon プロセッサを搭載し，アクセラレータを持たないシステムのみをプロットしている．

図より，インターコネクトネットワークとして Gigabit Ethernet を選択した場合の性能が，ほかの

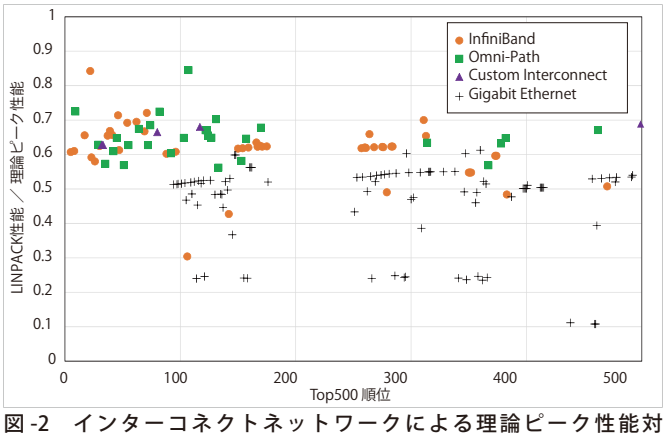

図 -2　インターコネクトネットワークによる理論ピーク性能対 LINPACK 性能比の相違

種類を選択した場合に比べて低くなる傾向が明らかである．この傾向は，アプリケーションの通信量や頻度，およびスパコンの規模によって変動するため，スパコンの設計では，用途に応じたインターコネクトネットワークへの予算配分が重要になる．

Xeon プロセッサ搭載スパコン事例

九州大学スーパーコンピュータシステム ITO サブシステム A

Xeon プロセッサを主要な計算資源としているスパコンの例として，筆者が所属する九州大学情報基盤研究開発センターで 2018 年 1 月に稼働開始したスーパーコンピュータシステム ITO（以降，ITO）のサブシステム A を紹介する．ITO は，主に国内の大学や研究機関に提供される全国共同利用の計算資源である．図 -3 に ITO の構成図を示す．このシステムで大規模計算を担当するバックエンドは，Xeon プロセッサのみを計算資源とするサブシステム A，および Xeon プロセッサと NVIDIA 社の Tesla を計算資源とするサブシステム B で構成される．ほかに，可視化などの対話的な処理を担当するフロントエンドと，バックエンドやフロントエンドから共有されるストレージがあり，これらがインターコネクトネットワークを介して接続されている．

ITO のサブシステム A は，各計算ノードが Intel 社の Skylake 世代の Xeon プロセッサ Xeon Gold 6154 を 2 基ずつ搭載している．このプロセッサは，基本のクロック周波数が 3.0GHz，プロセッサコア数が 18 個で，消費電力を示す TDP（Thermal Design Power, 熱設計電力）が 200W である．これらの数値は，同世代の 2 ～ 4 プロセッサ搭載サーバ向け Xeon プロセッサとしては高めである．このうち高い基本クロック周波数は，並列性が高くないプログラムでも高速に処理できることを意味し，このシステムの汎用性向上に貢献している．

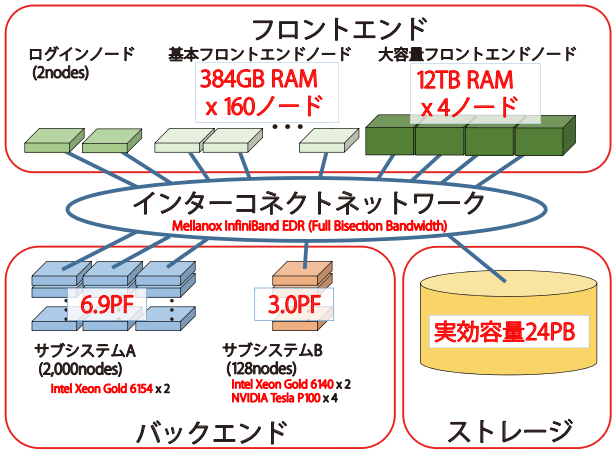

図 -3　スーパーコンピュータシステム ITO の構成図

各計算ノードに搭載されているメモリは 192GB であり，DDR4-2666 のメモリモジュールをプロセッサあたり 6 本ずつのメモリチャネルに装着することで，計算ノードあたりの総メモリバンド幅は約 255.9GB/ 秒となっている．

サブシステム A を含め，ITO の全体を接続するインターコネクトネットワークは，Mellanox 社の InfiniBand EDR である．このネットワークの片方向あたりの理論転送速度は 100Gbps である．また，ITO の任意の計算ノード間の通信遅延時間は 2 μ 秒以下である．

ITO のように多数の計算ノードを接続するシステムでは，接続に用いる形状が性能に大きく影響する．

たとえば 36 ポートのスイッチで単純な木構造を構成した場合，2,000 台の計算ノードを接続するのに必要なスイッチ数は高々 60 台程度である．しかしこの接続では，多数の計算ノードからの通信が 1 本の経路に集中する通信衝突が多発し，通信速度が大幅に低下するため十分な計算性能が得られない．そこで通常，InfiniBand や Omni-Path を用いたインターコネクトネットワークでは，Fat-Tree と呼ばれる形状でノード間を接続する．これは，前述の木構造で根に近づくほど帯域幅を広くするものである．実際のインターコネクトネットワークでは，スイッチ間を接続するための上位のスイッチを多重に用意することで経路を増やし，帯域幅の増加と通信衝突の軽減を図る．

特に ITO では，システム全体で Full Bisection Bandwidth（フルバイセクション帯域幅）となるように Fat-Tree を構成している．Full Bisection Bandwidth とは，全システムの任意の半数の計算ノードが残り半数の計算ノードに対して同時にデータを送信しても通信路の衝突が起こらない接続形態である．そのため，多数のジョブが同時に実行されている状況でも通信路の衝突が発生しにくく，安定した通信性能を維持することができる．

ITO では，407 台の 36 ポートスイッチと 8,044 本のケーブルを使用して，これを実現している．

InfiniBand EDR のもう 1 つの特徴は，SHARP（Scalable Hierarchical Aggregation and Reduction Protocol）である．これは，並列計算で多用される Allreduce という集約通信の中で必要になる，和や論理演算等の計算を，計算ノードのプロセッサではなくインターコネクトネットワークのスイッチ装置で行わせる機能で，これにより集約通信が高速化される上，プロセッサを本来の計算に専念させることで計算効率の向上も図ることができる．

国内外の Xeon プロセッサ搭載スパコン

続いて，国内外の Xeon プロセッサ搭載スパコンの事例として，2019 年 6 月の Top500 リストにおいて 100 位以内に入った国内のスパコンを表 -1 に，10 位以内に入った海外のスパコンを表 -2 に，それぞれ示す．

まず，表 -1 のシステムのうち気象庁の 2 つの Cray XC50 は，2018 年 6 月に運用を開始した同一機種で，片方が数値予報を担当する主系，もう片方が開発業務と主系のバックアップを担当する副系である．次に，東京大学の Oakbridge-CX は，ITO と同じ全国共同利用のシステムで，正式な運用開始は 2019 年 7 月である．また，六ヶ所核融合研究所の JFRS-1 は，同研究所の研究開発を支援する目的で 2018 年 6 月に運用が開始された．

一方，表 -2 のシステムのうちテキサス大学の Frontera は，主に米国の NSF（National Science Foundation）を介して全米に提供されるシステムで，2019 年 6 月に初期運用が開始された．これに対してライプニッツ研究センターの SuperMUC-NG は，ヨーロッパの研究者向けに提供されるシステムで，2018 年 9 月に運用が開始された．

表 -1，表 -2 の各システムに搭載されているプロセッサの名前の，Xeon の後に表記される Platinum，Gold はプロセッサのカテゴリを示しており，Platinum の方がより多くのコアを搭載したプロセッサを選択可能であるものの，Gold より高価である．そのためスパコンの設計における傾向として，1 台の計算ノードあたりの性能や，システム全体の電力対効率を重視する場合に Platinum が選択され，価格対性能を重視する場合に Gold が選択されることが多いと考えられる．また，カテゴリ名の後の数字は，選択可能なコア数とクロック数の組合せの番号を示している．これらの組合せも電力や価格に影響するため，スパコンの導入にあたっては，実際に利

表 -1　Top500（2019 年 6 月）において 100 位以内の国内の Xeon プロセッサ搭載スパコン

名前	組織	Processor	ノード数	Interconnect Network
Cray XC50	気象庁	Xeon Platinum 8160（2.1GHz, 24 コア）	2,816	Aries
Cray XC50	気象庁	Xeon Platinum 8160（2.1GHz, 24 コア）	2,816	Aries
ITO サブシステム A	九州大学	Xeon Gold 6154（3.0GHz, 18 コア）	2,000	InfiniBand
Oakbridge-CX	東京大学	Xeon Platinum 8280（2.7GHz, 28 コア）	1,368	Omni-Path
JFRS-1	六ヶ所核融合研究所	Xeon Gold 6148（2.4GHz, 20 コア）	1,370	Aries

表 -2　Top500（2019 年 6 月）において 10 位以内の Xeon プロセッサ搭載スパコン

名前	組織	Processor	ノード数	Interconnect Network
Frontera	テキサス大学（USA）	Xeon Platinum 8280（2.7GHz, 28 コア）	8,000	InfiniBand
SuperMUC-NG	ライプニッツ研究センター（Germany）	Xeon Platinum 8174（3.1GHz, 24 コア）	6,480	Omni-Path

用を予定しているいくつかのプログラムについて性能を比較し，予定している予算や使用可能な電力の範囲内で最適なものを選択する必要がある．

なお，本節で紹介したシステムを含め，Xeon プロセッサを搭載するスパコンのほとんどが，計算ノードに 2 プロセッサずつ搭載する構成を選択している．計算ノードあたり 1 プロセッサとしない理由としては，計算ノードあたりの性能が低く搭載できるメモリ量も少なくなることが挙げられる．また，計算ノードあたり 3 プロセッサ以上としない理由としては，計算ノードが大規模化して高額になることや，プロセッサとメモリの位置関係による性能変動が大きくチューニングが困難となることが挙げられる．

一方，インターコネクトネットワークについては，表 -1，表 -2 の各システムとも，Ethernet よりも高価で高性能なネットワークが選択されている．Omni-Path は Intel 社の製品で，InfiniBand と同様に 100Gbps の理論転送速度を有している上に，InfiniBand よりもポート数の多いスイッチを使用可能であるため，システム全体のスイッチ数やケーブル数を削減でき，安価にシステムを組むことができる．ただし，前述の SHARP のように集約通信をスイッチ装置で行わせる機能はない．これに対して Aries は Cray 社が自社のスパコンシステム専用に開発しているネットワークで，InfiniBand や Omni-Path と同等かそれ以上の通信速度を有する．

表 -1，表 -2 で InfiniBand や Omni-Path を選択している各システムのネットワークの接続形状は Fat-Tree である．ただし，Frontera と SuperMUC-NG は Full Bisection Bandwidth でない．これは，計算ノード数が 8,000 および 6,480 と多く，Full Bisection Bandwidth のために必要になるスイッチの台数やケーブルの本数が膨大となるためであると考えられる．詳細な形状は不明であるものの，一般にこのようなシステムでは，システム全体をいくつかのブロックに分け，それぞれのブロック内は Full Bisection Bandwidth となるように接続して，ブロック間の接続に用いるスイッチやケーブルを削減した形状とすることが多い．

これに対して Aries を選択した各システムは，Dragonfly と呼ばれる形状でノード間を接続している．これは，計算ノードを大規模，中規模，小規模のように多段階でグループ分けし，それぞれの段階でグループ間を全対全接続するもので，大規模なシステムにおいて Fat-Tree より少ないケーブル数でノード間を接続できる特徴がある．

今後の展望

近年，消費電力が計算機の性能を制約する最も重要な条件となったことから，アクセラレータやメニーコアのような電力対性能に優れた演算装置が登場している．しかしこれらの演算装置は，性能を発揮させることのできるプログラムが限定されるか，もしくは十分な性能を得るためのチューニング作業に時間を要することが多い．そのため計算機センター等で運用する際は，十分な利用者支援が不可欠となる．

これに対して Xeon プロセッサは，過去に開発されたほとんどのソフトウェアが最新の製品でも手を加えずに動作する．さらに，プロセッサを更新するだけで，特にチューニングを施さなくてもある程度の性能向上が得られる場合が多い．このように利用者や計算機センターの負担が少ないことが，Xeon プロセッサを主要な計算資源とするスパコンが依然として Top500 リストの 7 割程度を占めている理由の 1 つだと考えられる．

たとえば，2015 年に発売が開始された，Skylake と呼ばれるアーキテクチャを採用した Xeon プロセッサでは，コア数やメモリバンド幅等の向上に加え，それまでの 2 倍の 32 個のデータに対する SIMD（Single Instruction Multiple Data）演算を行える AVX-512 命令の追加により，大幅な性能向上を図っている．この Skylake アーキテク

チャ以降のプロセッサと，一世代前の Broadwell アーキテクチャ以前のプロセッサで，消費電力に対する LINPACK 性能の比を比較すると（図 -4），Skylake アーキテクチャ以降のプロセッサを使用した場合の方が高くなる傾向が見られる．このことから，少なくとも LINPACK ベンチマークでは，コア数の増加，メモリバンド幅の向上，および AVX-512 命令による効果が得られている，と判断できる．

しかし，コア数と SIMD 演算数の増加による性能向上には限界が見え始めている．たとえば，Skylake アーキテクチャ以降のプロセッサによる計算効率と，Broadwell アーキテクチャ以前のプロセッサによる計算効率を比較すると，図 -5 に示すように Skylake アーキテクチャ以降のプロセッサで計算効率が低下する傾向が見える．この原因として，まず Skylake 以降のクロック周波数の問題が挙げられる．計算効率の分母である理論ピーク性能は，プロセッサの基本クロック周波数と同時実行可能な最大の演算数の積であり，Skylake アーキテクチャでもそのように算出している．しかし Skylake アーキテクチャでは，同時実行可能演算数を最大にする AVX-512 命令を使用する際はクロック周波数が下げるように設計されている．すなわち，AVX-512 命令を基本クロック周波数で実行することはできない．

また，コア数や同時 SIMD 演算数の増加によって得られる性能向上は並列処理による効果である．そのため並列性の低いプログラムでは，ほとんどその効果が得られないか，もしくは性能向上のための大幅なチューニングが必要となる．その結果，Xeon プロセッサを主要な演算資源とするスパコンにおいても，今までのように互換性を維持した性能向上が困難となることが考えられる．これは，今後普及が予想される AMD 社の EPYC プロセッサのような Xeon 互換アーキテクチャも同様である．

これに加え，今後，特に省電力性が求められるスパコンにおいて，Xeon 互換プロセッサのような汎用性の高いアーキテクチャは不利であり，メニーコアやアクセラレータ等，コアの機能や性能を制限して省電力を追求したアーキテクチャが採用される可能性が高い．そのため利用者には，用途等の状況に応じて利用するアーキテクチャを選択し，プログラムを適切にチューニングすることが求められる．また，多くの計算機センターは，複数の種類のアーキテクチャによるスパコンを運用し，それぞれの特徴やチューニング技術を紹介することで，利用者の支援にあたっている．

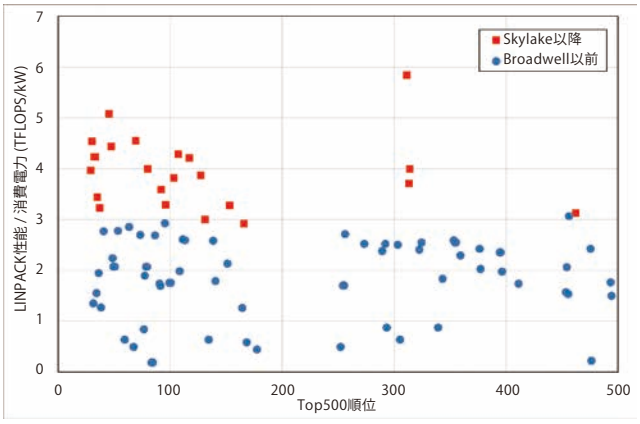

図 -4　Skylake アーキテクチャの消費電力対 LINPACK 性能比への影響

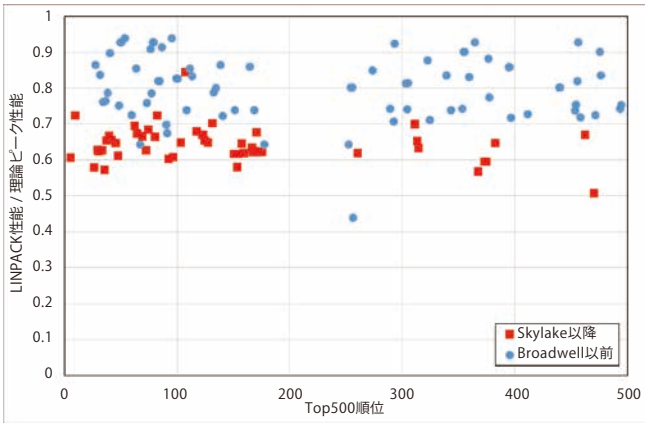

図 -5　Skylake アーキテクチャの理論ピーク性能対 LINPACK 性能比への影響

（2019 年 7 月 31 日受付）

南里豪志（正会員）　nanri.takeshi.995@m.kyushu-u.ac.jp
1995 年九州大学大学院工学研究科情報工学専攻 修士課程修了，1996 年九州大学大型計算機センター 助手，2001 年九州大学情報基盤センター 助教授，現在，九州大学情報基盤研究開発センター 准教授．博士（情報科学）．