［「京」の後の時代を支えるスパコン］

# ④ メニーコアアーキテクチャに基づくスーパーコンピュータ

朴　泰祐 | 筑波大学計算科学研究センター　中島研吾 | 東京大学情報基盤センター

## メニーコアプロセッサ

### メニーコアプロセッサの概要

現在のスーパーコンピュータ（以下，誌面節約のため「スパコン」と略す）は超並列方式，すなわち大量のプロセッサを何らかの高性能相互結合網で接続し，分散メモリ上で並列プログラミングを行って利用するのが典型的な姿である．特に電力あたり性能を非常に高めたい場合，システムを制御する CPU に加え，演算加速装置を付加する場合も多い．現在主流となっている演算加速装置は GPU（Graphics Processing Unit）であるが，より汎用プログラミングを容易にし，かつ汎用高性能 CPU よりもはるかに高い絶対性能と電力あたり性能を求めるプロセッサがメニーコア（Many-Core）型アーキテクチャである．一般的に，GPU では単一命令で同種の演算を大量に実行し，命令制御系を単純化することで電力あたり性能を向上させている代わりに，そのプログラミングは専用の言語を用いたりプログラム中の並列性について常に意識する必要があり，汎用 CPU に比べ煩雑である．

一般的に入手可能な商用メニーコアプロセッサとして最も利用されているのは Intel 社の Xeon Phi と名付けられたプロセッサファミリーである．初代の Xeon Phi は評価用（非売品）の Knights Ferry（KNF, 2010 年），その後の商用版の Knights Corner（KNC, 2012 年）というコード名で開発・販売された．これらは PCIe（PCI Express）カードとして提供され，GPU のように通常の CPU をホストとして，それに接続される演算加速装置であった．その後，メニーコアプロセッサ本体が通常の CPU と同様に自らブートし，システム管理，つまり OS を実行でき，演算加速も行うような Bootable CPU として提供されたのが Knights Landing（KNL, 2016 年）である．

### Xeon Phi プロセッサの CPU コアの特徴

そもそもメニーコアプロセッサと一般的なマルチコアプロセッサはどう違うのだろうか？　Xeon Phi シリーズは基本的に Intel 社 CPU の標準命令セットである IA-32（いわゆる x86 系）に基づく命令セットアーキテクチャを持つ．このため，各コアでは普通に x86 バイナリが動作し，Linux 等の OS も動作する．一般の x86 系 CPU はマルチコア構成で，数個～ 20 個程度のコアが 1 チップ上に実装されている（コア数は年々増加している）．一方，メニーコアである Xeon Phi は 50 ～ 70 個程度のコア数を持ち，単純比較では数倍のコアが内蔵されている．これらが汎用高性能 CPU コアと同様に動くのであれば，多い方が得，ということになるが話はそう単純ではない．

現在の汎用高性能 CPU は消費電力が非常に大きく，スパコンに用いられるものでは 100W 以上である．コア数が増えると当然電力も大きくなるが，半導体テクノロジー（シリコンの微小化）やアーキテ

クチャの工夫で，コア数と性能が向上しても電力を一定に抑えてきた．したがって，同じ内容のコアを数倍搭載すれば，当然電力が大幅に増えてしまう．そこでメニーコアプロセッサのコアは，命令セットこそ同じだが，演算パイプライン構成，分岐予測機構，キャッシュおよびメモリの制御機構等，CPU性能の本質にかかわるようなアーキテクチャ上の簡略化が行われている．たとえば1つのコアでの整数命令の実行性能を単純比較すると，汎用高性能CPUコアの方がメニーコアのそれより高速である．この点が，メニーコアプロセッサの性能チューニングを難しくする要因となっている．

Intel社は，メニーコアプロセッサのコア単純化を補うため，浮動小数点演算命令を大幅に強化した．Xeon Phiシリーズのプロセッサには512bit SIMD（Single Instruction Multiple Data）命令セットが備えられた．これは1命令で，たとえば64bit倍精度浮動小数点演算であれば4つ同時に，または32bit単精度浮動小数点演算であれば8つ同時に実行する．対象となるレジスタも512bitに拡張されている．四則演算だけでなく，科学技術計算で多用される積和演算（乗算と加算を1命令で実施）も可能である．512bitの積和演算命令を使えば，1命令で16個の単精度浮動小数点演算が同時に実行できるわけである．SIMD命令も多くの電力を消費するのだが，独立な16個の命令を順に実行するよりも演算パイプラインの制御が単純化され，演算あたりの消費電力は大幅に低減される．しかし，メモリ上のデータの並びや，演算順序の制約から，規則的な配列処理だけに向いていて，これが科学技術計算の一般的特性にマッチする．このように，科学技術計算の特性を活かしつつ，コアのアーキテクチャを単純化することで，メニーコアプロセッサをこれらの処理向けに実用化することに成功している．逆に言えば，一般的なOSや事務処理アプリケーションの実行はメニーコアプロセッサではうまく高速化できないともいえる．また，これらの命令をなるべく高頻度で利用できるようにアプリケーションおよびコンパイラが工夫されなければならない．

もう1つの電力削減要因は動作周波数である．一般的な汎用高性能CPUは2.8～3.3GHz程度のクロック周波数を持つが，KNLでは1.4GHz程度と低く抑えられている．CMOSデバイスの消費電力は周波数に大きく影響されるため，これは重要である．命令あたりの浮動小数点演算性能の増加，多数のCPUコア，比較的低い動作周波数の組合せによって科学技術計算において非常に高性能を出すCPUが実現されたといえる（なお，512bitのSIMD命令はその後，高性能向けマルチコアCPUにも導入された）．

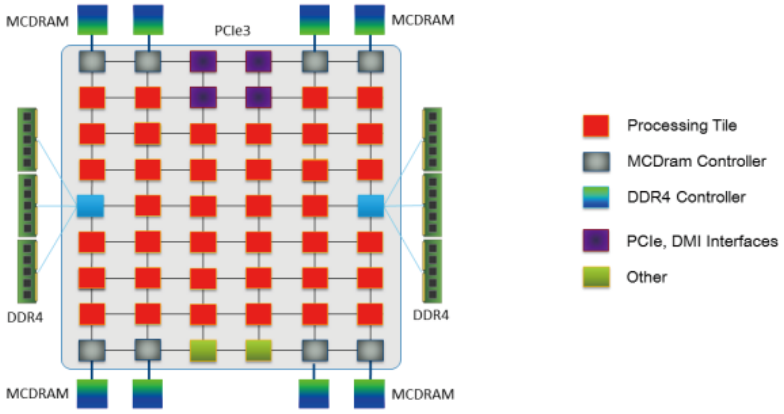


図-1　Intel Xeon Phi 7250 Knights Landing プロセッサの内部構成（http://intel.com より）

## Knights Landing のメモリ構成

図-1に，KNLの内部構成を示す．KNLはそのスペックによりいくつかの型番があるが，ここでは一番用いられているIntel Xeon Phi 7250プロセッサ（1.4GHz, 68 cores）を取り上げる．同プロセッサは合計68個のCPUコアが2個一組でペアになり（タイルと呼ばれる），メモリへのインタフェースと一部のキャッシュを共有している．34個

のタイルはチップ上で2次元メッシュ構造のネットワークで結合されており，この上でチップ上に組み込まれた高バンド幅メモリであるMCDRAMと，チップに外付けされた汎用DRAMメモリ，さらにPCIe等の外部インタフェースに接続される．MCDRAMはメモリモジュールへのアクセスパスを通常のDRAMより広くし，一度のアクセスで大量の連続データを読み書きすることでスループットを向上する特殊なメモリで，チップ上でCPU部と直結配線されている．MCDRAMはチップに組み込まれているため容量に制限がある（7250では16GB）．これを補うため，KNLでは一般的なDDR型のDRAMも利用可能で，両クラスのメモリをうまく使い分けることでメモリバンド幅がボトルネックになりがちな科学技術計算アプリケーションを高速化することが可能であり，この点も性能チューニングの要点である．7250プロセッサの理論ピーク性能は約3TFLOPSであり，同世代のマルチコアプロセッサに比べ7～8倍程度高速である．しかし，アプリケーション性能を最大に引き出すような最適化を施さなければこの性能は発揮できない．

## 国内におけるメニーコア型スパコン

KNCを演算加速装置として用いたスパコンの中で国内最高性能のものは筑波大学・計算科学研究センターのCOMAシステムであったが，これは2019年3月で運用を終了している．一方，KNLの登場により，これをスパコンの本体CPUとして用いることができるようになり，これを用いたスパコンは国内外で多く導入された．KNLはスタンドアローンなCPUであるため，マザーボードをコンパクトに構築でき，特にTOP500リスト（本特集「1. 日本のスーパーコンピュータの現在の状況と今後」の冒頭を参照）で上位にランクするような超並列・高性能システムに用いられている．

KNLを用いたHPCI（革新的ハイパフォーマンスコンピューティングインフラ）資源としては，筑波大学と東京大学が共同運営するJCAHPCが運用するOakforest-PACS，京都大学のCamphor 2，北海道大学のGrand Chariotの一部，大阪大学のOCTOPUSの一部等がある．このうちCamphor 2はCray社が提供する独自の相互結合網を持つ超並列システムであるが，残りはKNLをCPUとするクラスタ型計算機である．本稿の残り部分ではこのOakforest-PACSを例として取り上げる．


図-2　Oakforest-PACSの外観（左）と計算ノードの内部（右）

## Oakforest-PACS

Oakforest-PACS（以下，OFP）は筑波大学と東京大学が共同調達・共同運用しているKNLを用いた超並列クラスタで，2016年11月のTOP500リストに初めて載り，世界第6位，国内第1位にランクされた．つまり，「京」コンピュータを上回る性能を持つスパコンが大学主導で導入された．両大学は，OFPの調達と運用を円滑に行うために，JCAHPC（最先端共同HPC基盤施設：Joint Center for Advanced HPC）を設立し，予算および人的資源，スパコン稼働に必要な光熱水料等をすべて一定の按分で負担している．ただし，スパコン本体は東京大学柏キャンパスの情報基盤センターに設置されている．2018年6月のTOP500リストで，産業技術総合研究所が運用するABCI（AI Bridging Cloud Infrastructure）システムが登場するまで，国内最高性能スパコンであった．

OFPは8208ノードのKNLによる計算ノードからなる．図-2（左）にOFPの外観を，また図-2

（右）に計算ノードを示す．OFP の各計算ノードは Intel Xeon Phi 7250 メニーコアプロセッサを 1 基搭載し，相互結合網インタフェースとして Intel の Omni-Path Architecture による，理論ピークバンド幅 100Gbps が搭載されている．メモリは先述のように 16GB の MCDRAM に加え，DDR メモリが 96GB 接続されている．

超並列スパコンのアーキテクチャ上の大きな特徴は，高性能プロセッサ間での通信を高性能で行うための相互結合網をどのように構築するかにある．OFP ではクラスタ型計算機で一般的に用いられる Fat-Tree と呼ばれるトポロジで結合網を構築している．Fat-Tree とは，Ethernet 等に用いられる単純な Tree 構造と異なり，相互結合網を多階層で構築する際，スイッチの上位側と下位側で同等のバンド幅を確保するようネットワークリンクとスイッチを増強したものである．図 -3 に相互結合網の構成を示す．スイッチは上下 2 つの階層からなり，下位階層では 48 ポートのスイッチ（上位・下位それぞれ 24 ポート）を用い，上位階層では 768 ポートの巨大なスイッチを用いている．この構造により，どの計算ノードからどの計算ノードへの通信も，理論的にはブロックされずに最高の並列通信が可能となる．

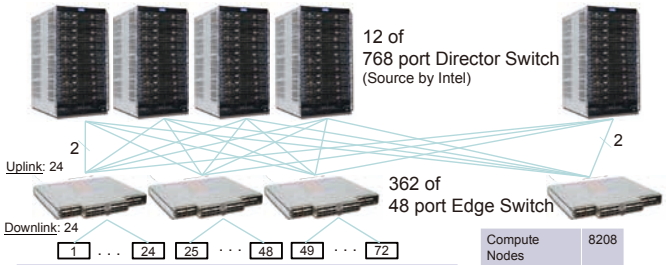


**図 -3　Oakforest-PACS の大規模相互結合網（full bisection bandwidth fat-tree）**

**表 -1　Oakforest-PACS の仕様**

| | | | |
|---|---|---|---|
| Total peak performance<br>Linpack performance | | | 25 PFLOPS<br>13.55 PFLOPS (with 8,178 nodes, 556,104 cores) |
| Total number of compute nodes | | | 8,208 |
| Compute node | Product | | Fujitsu PRIMERGY CX1640 M1 |
| | Processor | | Intel® Xeon Phi™ 7250（Code name: Knights Landing），<br>68 cores, 3 TFLOPS |
| | Memory | High BW | 16 GB, > 400 GB/sec (MCDRAM, effective rate) |
| | | Low BW | 96 GB, 115.2 GB/sec (DDR4-2400 x 6ch, peak rate) |
| Inter-connect | Product | | Intel® Omni-Path Architecture |
| | Link speed | | 100 Gbps |
| | Topology | | Fat-tree with (completely) full-bisection bandwidth (102.6TB/s) |
| Login node | Product | | Fujitsu PRIMERGY RX2530 M2 server |
| | # of servers | | 20 |
| | Processor | | Intel Xeon E5-2690v4 (2.6 GHz 14 core x 2 socket) |
| | Memory | | 256 GB, 153 GB/sec (DDR4-2400 x 4ch x 2 socket) |

OFP のハードウェアおよびソフトウェアの仕様を表 -1 にまとめた．OS は Linux を用いるが，フロントエンド（ユーザがログインしてコンパイルやジョブ投入を行うサーバ）はセキュリティを重視して Red Hat Linux を導入しているのに対し，計算ノード間は一般的なログインができないため，CentOS（無償版）を導入している．

OFP のもう 1 つの特徴として，高性能共有ファイルシステムが挙げられる．すべての計算ノードおよびフロントエンドサーバから均一にアクセス可能な Lustre ファイルシステムによる 25PB の共有ストレージがあり，ユーザはここに初期データや計算結果を保存できる．共有ファイルシステムであるため，次の計算がどの計算ノードで実行されても問題はない．OFP ではさらに，SSD で構成された約 1PB の高速ファイルキャッシュを備えている．非常に高速で高いバンド幅（共有ファイルシステムの約 3 倍）を持つことで，大規模計算中の一時ファイルや結果ファイルを高速に読み書きすることができる．なお，2019 年 6 月時点で，この高速ファイルキャッシュのバンド幅は世界最高性能としてランクされている（IO-500 ベンチマークによる）．

## Oakforest-PACS の運用とアプリケーション

OFP は JCAHPC によって管理されるが，その利用プログラムは若干複雑である．筑波大学と東京大学はそれぞれの調達および運用資金の比率で計算資源を按分しており，各大学の各種利用プロ

グラムはその枠内で自由に設定されている．また，JCAHPC として HPCI 一般利用に共同で資源を提供している．つまり，HPCI を通しての OFP の利用は筑波大学・東京大学が共同で行っている．

HPCI 一般利用のほかに両大学が独自の利用公募を行っているが，全体としてのアプリケーション分野の分布を図 -4 に示す．ここで目立つのは QCD（Quantum Chromo-Dynamics）で，各種プログラムを通じて総資源の 40% 以上が使われている．QCD に次いで多い地球科学・宇宙科学分野の利用者の多くは，前年度まで稼働していた東京大学の Oakleaf-FX（富士通 PRIMEHPC FX10，「京」の商用版）を利用していたが，2018 年度からは OFP を利用するようになっている．全体的には基礎科学分野，ものづくり，地球科学，材料科学等の分野が 70% 程度の資源を利用していることが分かる．

各アプリケーションのジョブサイズは 1 ノードからシステム全体を使うまで多岐に渡るが，数百人のユーザが共有するため巨大なアプリケーションを実行するには資源調整が必要である．JCAHPC では，毎月月末に実施されるシステムメンテナンスに合わせ，OFP のすべてのノードを 24 時間占有利用することができる「大規模 HPC チャレンジ」を実施している．これは計算科学・計算工学分野の大規模シミュレーションを挑戦的に行う研究者への特別措置で，研究の内容やシミュレーションコードの準備状況，予備評価等を添えた申請を行うことで，無償で OFP のフルシステムを 24 時間利用できる．

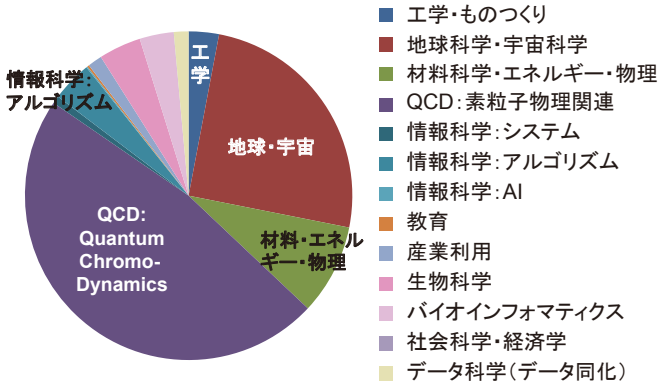


図 -4　2018 年度の Oakforest-PACS の資源利用分布

## ポスト「京」コンピュータ開発への貢献

最後に OFP の重要な役割の 1 つとして，ポスト「京」コンピュータのシステムソフトウェア開発への貢献を紹介しておく．理化学研究所・計算科学研究センターでは，ポスト「京」コンピュータ『富岳』を効率的に運用するための OS である McKernel を開発している．これはメニーコアアーキテクチャに基づくプロセッサという特性に基づき，計算ノード内のスレッド処理を揃え，スレッド間待ち合わせによる性能低下を防ぐ機能や，超並列分散メモリシステム上でのメッセージパッシングを効率的に行う機能等が盛り込まれている．国内でこの開発に貢献できる計算機資源は少なく，OFP はメニーコア型クラスタの代表としてこの研究に大きく貢献している．さらに，同センターで開発中の，PGAS（Partitioned Global Address Space）モデルに基づく大規模並列記述言語である XcalableMP の開発にも用いられている．

（2019 年 8 月 29 日受付）

**朴　泰祐（正会員）**　taisuke@ccs.tsukuba.ac.jp
慶應義塾大学大学院理工学研究科電気工学専攻修了，工学博士（慶應義塾大学）．現筑波大学計算科学研究センター長，同大学院システム情報工学研究科 教授．現 HPCI 連携サービス委員会委員長．超並列アーキテクチャ，相互結合網，クラスタ計算に興味を持つ．2011 年 ACM ゴードンベル賞共同受賞．IEEE，ACM 各会員．

**中島研吾（正会員）**　nakajima@cc.u-tokyo.ac.jp
東京大学工学部航空学科卒業．テキサス大学大学院航空宇宙工学専攻修了（MS）．博士（工学）（東京大学）．三菱総合研究所，高度情報科学技術研究機構，東京大学理学系研究科を経て，2008 年より東京大学情報基盤センター教授，2018 年より理化学研究所計算科学研究センター副センター長，専門は数値流体力学，並列アルゴリズム，日本応用数理学会・日本計算工学会・SIAM・IEEE 各会員．