SC17 参加報告

星野 哲也、有間 英志、塙 敏博、下川辺隆史
東京大学情報基盤センター

本稿は、2017年11月12日から17日までアメリカ合衆国コロラド州デンバーにて開催された、SC17 (The International Conference for High Performance Computing, Networking, Storage and Analysis) の参加報告である。SC17は、高性能計算(HPC)分野に於いて最大級の国際会議であるとともに、様々な情報技術関連企業や研究所・大学等の技術展示会としても知られている。本稿では、SC17に参加した東京大学情報基盤センター教職員が現地で見聞きした中から気になったことを記す。

1　はじめに

本会議は以前はSupercomputing-XY(XY：開催年)という名称が用いられており、1997年にSC-XYという名称に変更された。1988年フロリダ州オーランドで第1回が開催されてから、毎年11月にアメリカ各地を転々としながら開催されており、今回はSupercomputing-88から数えて29回目である。SC17の開催地は、SC13以来4年ぶりのコロラド州デンバーであった。会場も4年前と同様、巨大なクマの像が人気のコロラドコンベンションセンターであった。本センターはクマ以外にも会場がコンパクトにまとまっていることが評判であり、実際、展示場から各セッション部屋までの移動が非常にスムーズであった。

図1　SC17 会場 (Colorado Convention Center)

本会議は、基調講演、研究発表、パネル討論、BoF(Birds of a Feather: 特定のトピックを定めた小規模集会)、主要技術の理解を助けるチュートリアル、併設される多数のワークショップなどで構成されている。研究発表については61件行われ、また12件のパネル、99件のポスター、41件のチュートリアル、37件のワークショップ、13件の招待講演、79件のBoFなど盛りだくさんの内容であった。また、企業や各種研究機関による最新の製品や技術の展示発表も注目すべき内容であった。主催者発表によると、SC17の参加登録者数は12,000人を越える見込みである。そのうちアメリカ以外からの参加者は2,800人程とのことで、日本からも多くの研究者が参加した。展示には334団体が参加し、大変な賑わいであった。

2　東京大学情報基盤センターによる展示

東京大学情報基盤センターは昨年同様、「ITC/JCAHPC, The University of Tokyo」の名義によるブース展示を行った。筑波大学計算科学研究センターと共同で設立した最先端共同HPC基盤施設をJCAHPCと呼び、Oakforest-PACSスーパーコンピュータの運用を行なっている。今回のブース展示においても、筑波大学計算科学研究センターによる「CCS/JCAHPC, University of Tsukuba」と隣合わせのブースにて、一体的な展示を実施した。また例年同様に、情報基盤センターの提供する計算資源や研究事例を紹介するポスターの展示を行い、パンフレット・チラシ・グッズの配布を行った。恒例となっているブースでのプレゼンテーションも両センターで協力して実施した。ブースプレゼンテーションでは、Oakforest-PACSなどを使った最新の研究など、以下の7件の講演が行われた。Oakforest-PACSは稼働開始から1年が経過し様々な成果が創出されてきている。システム全体を利用した電子動力学シミュレータや地震工学アプリケーションなどの実アプリケーションでの研究成果や、著名なMPIライブラリの一つであるMVAPICH2の実行性能やスケーリングについての発表が行われた。

- “Optical Material Simulation on Full System Scale of Oakforest-PACS,” Taisuke Boku (CCS/U.Tsukuba)
- “Optimizing MPI Startup on Oakforest-PACS,” Kenneth Raffenetti (Argonne National Laboratory)
- “An Overview of the IHK/McKernel Lightweight Multi-kernel based Operating System for Extreme Scale HPC,” Balazs Gerofi (AICS/RIKEN)
- “Accelerating Dynamic Implicit Finite-element Method with Adaptivemultistep Predictor on Oakforest-PACS,” Kohei Fujita (Earthquake Research Institute/UTokyo)
- “Scalability and Performance of MVAPICH2 on OakForest-PACS,” Dhabaleswar K. Panda (The Ohio State University)
- “Improving the Scaling of Conjugate Gradient Eigensolvers,” Osni Marques (Lawrence Berkeley National Laboratory)
- “Optimizations of H-matrices Library for Many-core Processors,” Tetsuya Hoshino (ITC/UTokyo)

3　各種のランキングについて

毎年のSCではスーパーコンピュータの性能に関する様々なランキングが更新される。最も有名なスパコンランキングであるTop500、スパコンの電力性能を競うGreen500、実際のアプリケーションに近いと言われるHPCGや、今回より始まったスパコンのIO性能を競うIO-500などがある。本稿ではTop500から見える全体的な傾向に加え、JCAHPCが運用するOakforest-PACSが1位となったIO-500について詳しく取り上げる。

Top500 (http://www.top500.org/)は世界のスーパーコンピュータの性能をLINPACKという係数行列が密の連立一次元方程式を解くベンチマークの処理速度によって競うものである。1993年の開始以来、6月にヨーロッパで行われる会議であるISCと、本会議SCにて年2回の更新を続けている。近年は特に上位のシステムの入れ替わりが少なくなっていたが、今回の更新で

図 2　ブース展示の様子 (ゲストによるブースプレゼンテーション、展示物を紹介する様子)

表 1　Top500 の上位 10 のスーパーコンピュータ (http://www.top500.org/ より抜粋、一部編集)

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Rmax / Rpeak	Power (kW)
1	Sunway TaihuLight	10,649,600	93,014.6	125,435.9	0.74	15,371
2	Tianhe-2 (MilkyWay-2)	3,120,000	33,862.7	54,902.4	0.62	17,808
3	Piz Daint	361,760	19,590.0	25,326.3	0.77	2,272
4	Gyoukou	19,860,000	19,135.8	28,192.0	0.68	1,350
5	Titan	560,640	17,590.0	27,112.5	0.65	8,209
6	Sequoia	1,572,864	17,173.2	20,132.7	0.85	7,890
7	Trinity	979,968	14,137.3	43,902.6	0.32	3,844
8	Cori	622,336	14,014.7	27,880.7	0.50	3,939
9	Oakforest-PACS	556,104	13,554.6	24,913.5	0.54	2,719
10	K computer	705,024	10,510.0	11,280.4	0.93	12,660

は海洋研究開発機構(JAMSTEC)に設置された暁光が4位（国内1位）にランクインし、6月のランキングでは10位であったTrinity (DOE/NNSA/LANL/NSL) が7位にランクインした。これにより、前回7位であったOakforest-PACSは2つ順位を下げ9位（国内2位)、京コンピュータが10位（国内3位）となった。表1はTop500の10位までのスーパーコンピュータのCores（総コア数)、Rmax（Linpackにより計測された性能)、Rpeak（理論性能)、Rmax/Rpeak（理論性能比の効率)、Power（Linpack測定時の消費電力）を示したものである。

まず今回4位にランクインした暁光に着目すると、コア数が桁違いに多いことがわかる。

Rmaxが1位のSunway TaihuLightの20%強であるのに対し、コア数は2倍近い。経験的に、扱うべきコア数が増すほど、それを扱うためのプログラミングコストは増大する傾向にあり、また効率を出すのも難しくなってくるが、暁光の理論性能に対するLinpackの性能は68%と標準的であり、開発チームのプログラム最適化技術の高さが伺える。特筆すべきは消費電力であり、Piz Daintと比較して60%程度の消費電力で同程度の性能が得られている。この低消費電力技術により、暁光を開発したPezy Computing社のシステムが今回のGreen500ランキングのTop3を独占した。ひとまずLinpackベンチマークにおいて十分な存在感を示した暁光であるが、真価が見極められるのは今後、実際のアプリケーション実行時の性能が出てきてからであろう。ベンチマークプログラムと違い、実際に利用されているアプリケーションはソースコードの規模が大きくなりがちであり、システム向けの最適化には膨大なプログラミングコストを伴うためである。暁光同様、膨大な数の演算コアを持つSunway TaihuLightを設置した中国では、アプリケーションのSunway TaihuLight向け最適化に、中国ならではの人海戦術で臨んだと聞く。暁光はその膨大な演算コアに向けたアプリケーション最適化にどのような方法で挑むのか。注目である。

次に、7位にランクインしたTrinityに目を向けてみると、驚くほどに理論性能に対する効率が悪い。Linpackベンチマークは、最も効率を出しやすいベンチマークの一つであり、30%台というのは驚きである。理由の一つとして考えられるのは、TrinityのプロセッサであるIntel Xeon Phi Knights Landing (KNL)である。8位のCori、我々が運用する9位のOakforest-PACSも同プロセッサを利用しているが、いずれも効率は50%代でありかなり低い。実はKNLの演算コアは周波数1.4GHzを謳っているが、演算密度の高いLinpackのようなベンチマークを実行する際には、1.2GHzまで周波数が下がる。チップの温度によってはさらに周波数が下がる仕組みとなっているため、非常に効率が悪く見えてしまっている。それにしてもTrinityは、同じKNLを用いるCori・Oakforest-PACSよりもさらに低い効率となってしまっている。TrinityはHaswell搭載ノードとKNL搭載ノードの混成であるとのことで、この点が悪影響を及ぼす要因となってしまったのかもしれない。あるいは異なるプロセッサが1システムに搭載されているため、プログラムの最適化がうまくいかなかったのかもしれない。いずれにせよ，なぜそのような性能値になったのか理由を調査したい。

本センターが運用するスパコンでは、Oakforest-PACSがTop500の9位となった他、Oakleaf-FXが157位、Reedbush-Lが291位、Reedbush-Hが295位となった。また、Green500においてはReedbush-Lが11位、Reedbush-Hが16位、Oakforest-PACSが22位となっている。

今回SC17において、最初のIO500 List (http://io500.org/)が公開された。スパコンシステムにおけるIO性能について、カタログ値だけではなく実際の性能値を測定し比較することで、バランスのとれたスパコンシステムの構築を促し、調達の際のベンチマークとして利用されることも考慮されている。

測定の内容は、

- バンド幅性能
  - IOR easy read/write (ユーザが決めて良い)
  - IOR hard read/write (single-shared file, small unaligned, POSIX)
- IO処理性能

表 2　IO500 のスーパーコンピュータ (http://www.io500.org/ より)

Rank	System	ファイルシステム	クライアントノード数	スコア	バンド幅 (GiB/s)	IO 処理性能 (kIOP/s)
1	JCAHPC Oakforest-PACS	IME	2048	101.48	471.25	19.04
2	Kaust Shaheen	DataWarp	300	70.90	151.53	33.17
3	Kaust Shaheen	Lustre	1000	41.00	54.17	31.03
4	JSC JURON	BeeGFS	8	35.77	14.24	89.81
5	DKRZ Mistral	Lustre	100	32.15	22.77	46.64
6	IBM Sonasad	Spectrum Scale	10	21.63	4.57	102.43
7	Fraunhofer Seislab	BeeGFS	24	18.75	5.13	68.55
8	PNNL EMSL Cascade	Lustre	126	11.17	4.88	25.59
9	SNL Serrano	Spectrum Scale	16	4.25	0.65	27.98

– mdtest easy create/stat/delete (ユーザが決めて良い)
– mdtest hard create/stat/read/delete (single-shared directory, 3901 byte fles, POSIX)
– “find” の操作

である。これらの結果、バンド幅性能として上記4種の測定値の幾何平均、IO処理性能として上記8種の測定値の幾何平均を求め、最終的に $\sqrt{(\text{バンド幅性能} \times \text{IO処理性能})}$ の値を IO500 のスコアとする。

表2は、各システムにおけるファイルシステム、測定に用いたクライアントノード数と、IO500のスコアの内訳を示したものである。この表からもわかる通り、本センター (JCAHPC) の Oakforest-PACSがトップになった。バーストバッファであるIMEを用いたことにより、特に高いバンド幅が得られていることがわかる。なお、現時点ではIO500と言いつつ9つのエントリしかないため、今後のリストの充実が望まれる。

4　メイントラック論文について

4.1　Leveraging Near Data Processing for High-Performance Checkpoint/Restart by Abhinav Agrawal, Gabriel H. Loh, and James Tucki

North Carolina State UniversityとAMDの著者による “Leveraging Near Data Processing for High-Performance Checkpoint/Restar” という論文である。特に共著者の一人であるGabriel H. Lohはメモリアーキテクチャの世界で有名であり、本論文もNear Data Processingと呼ばれる近年注目を集めつつある計算パラダイムを、大規模HPCシステムにおけるチェックポイント・リスタートの高効率化に応用するというものである。

エクサスケール級の大規模HPCシステムにおいては、CPU、メモリ等コンポーネント数の劇的な増加により、障害発生頻度の大幅な増大が危惧されている。従って、その様なシステム上で長時間ジョブを実行する際には、アプリケーションの途中経過(チェックポイント)の定期的保存、及び、障害発生後の最新チェックポイントからの復帰(リスタート)が必須の技術であると考えられている。しかし、システム規模が大きくなればなるほど、保存すべきチェックポイントの量も膨大になり、結果としてアプリケーションの実行にリソースを割ける時間が限られてしまうという問題がある。

本研究では、ノードごとのローカルなストレージの利用、特に、ストレージデバイス内のロ

ジックを用いたNear Data Processing(NDP)を行うことによって、並列ファイルシステムへのアクセスオーバーヘッドを削減し、この問題に対処している。NDPとは、一般的に、メモリ・ストレージデバイスのそばにあるロジック(例えば、三次元積層DRAMの最下層のロジックやSSD内部のコントローラロジック)に軽量の計算をオフロードすることで、CPUとこれらデバイス間との通信オーバーヘッドを削減し、高性能・低電力化を図るという計算パラダイムである。本研究では、チェックポイントをローカルなストレージに保存するというところまでをCPUが担当し、(1)「チェックポイントデータの圧縮」、(2)「並列ファイルシステムへの圧縮データの転送」をストレージデバイス側のロジックが担当しており、それによってチェックポインティングのオーバーヘッドの削減を行っている。(1)の具体的な狙いは、並列ファイルシステムへのアクセストラフィックの削減による、チェックポイントの高頻度化とリスタートの高速化を可能にする点にあり、(2)の具体的な狙いは、アプリケーションの実行とチェックポイントの保存とをオーバーラップさせることによる、チェックポインティングオーバーヘッドを隠蔽するところにある。

今現在において製品化されているストレージデバイス内のロジックは、圧縮や転送機能を持たないため、実システムを用いた有効性の評価は不可能である。そこで本論文では、性能モデルを用いた単純な評価行っている。評価の結果、並列ファイルシステムに直接チェックポイントを書き込む場合に比べ、大幅にオーバーヘッドを削減できることを論文では示している。

エクサスケール以降の大規模HPCシステムにおいては、チェックポイント・リスタートのオーバーヘッドを削減することは重要な課題である。その際、本提案にある圧縮やオーバーラッピングは有力な方針であると言える。しかし、NDPロジックに関して要求性能を満たすためのハードウェア要件、および、そのオーバーヘッド(コスト・電力)に関する議論がないため、実用性に疑問が残る。特に、HPCは市場が小さいため、これらオーバーヘッドが小さくなければ、デバイスベンダー側が実装するとは考え難い。この様に、アイディアとしては面白いものの、実際のシステムに実用するという点については、まだまだ考慮することがあると言える。

4.2 sPIN: High-performance streaming Processing in the Network by Torsten Hoefler, Salvatore Di Girolamo, Konstantin Taranov, Ryan E. Grant, Ron Brightwell

ETH ZurichとSandia National Laboratoriesの著者らによる“sPIN: High-performance streaming Processing in the Network”と題された論文である。筆頭著者であるETH ZurichのTorsten Hoeflerは、HPCのネットワーク関連分野において大変著名であり、本研究では前述のNear Data Processingに似た概念であるProcessing in the Networkを駆使し、大規模HPCシステムにおけるネットワークの高性能化を図っている。

エクサスケール級の大規模HPCシステム上で並列アプリケーションを動かす際には、ノード間通信の高速化は必要不可欠である。特に昨今のHPCシステムでは、GPUに代表されるアクセラレータの利用等によってノードの演算性能は大幅に向上しており、結果としてこの通信オーバーヘッドが特に問題になりつつある。通信の高速化のため、近年では、Remote Direct Memory Access (RDMA) と呼ばれる、ネットワークで接続されたリモートノード上のメモリに、オペレーティングシステム(OS)なしに直接データ転送を行う仕組みがInfiniBand等にて実装されている。本研究では、このRDMAをベースとしてさらなる通信の高効率化を図っている。

本研究の核となる提案はstreaming Processing in the Network (sPIN)と呼ばれる通信の高効率化の仕組みとそれを実現するハードウェアアーキテクチャである。具体的には、上述のRDMAに基づく高性能ネットワークにおいては、送受信データの処理をCPUを介して行う部分にこそボトルネックが存在しており、sPINではこの処理をCPUからNIC内のハンドラ処理ユニット(HPU)等へオフロードすることでこれを解決している。例えば、ノード間のPing-Pongでは、現状ではすべてのメッセージがメインメモリとCPUを介して行われるものの、sPINでは、NIC内のローカルメモリ及びハードウェア内で処理を完結させることで、レイテンシの短縮を行っている。例えば、ノード間で行列のアキュムレートを行う場合でも、受信側ノードにてNIC内のHPUがCPUの代わりにローカルメモリを使って計算を行う。その際、ローカルメモリとメインメモリの間のデータ転送はNIC内のDMAユニットが行う。論文ではPing-Pong、ブロードキャスト、RAIDのパリティアップデート等でsPINによる高性能化が可能であることを定性的・定量的に示している。

本研究の特筆すべき点は、異なる2つのシミュレータ(LogGOPSim、Gem5)を駆使した緻密な評価である。ここで、LogGOPSimはネットワークや分散メモリのシミュレーションを担当し、Gem5はノード内のCPUやNICのシミュレーションを担当している。本研究の様にハードウェアの変更を伴う提案の場合、実機での検証が困難なため、シミュレータを用いた評価を行うことが多い。その場合、より広範囲なコンポーネントをより緻密に評価を行うためには、一般的に実装や実験により長い時間を費やす必要がある。例えば本研究の場合、Gem5上にsPIN用ハードウェアを実装した上でLogGOPSimと協調動作させる必要がある。さらには、Gem5は高機能で拡張性が高い反面、コードが複雑かつ膨大なため、使いこなすのは困難である。従って、本評価のためには非常に多くの労力を費やしたことが予想される。

その反面、実用上では幾つか疑問が残る。例えば、現実のHPCアプリケーションに対してsPINがどの程度効果的かについての検証はなく、Ping-Pong、ブロードキャスト、RAIDのパリティアップデート等の単純なタスクのみ検証がなされている(これらの検証だけでも多大な労力を必要とするが)。また、結局のところNIC内のハードウェアはどの様な機能をカバーすれば十分なのかについても、ここでは明らかにはされていない。さらに、各ノードにおいて、NIC内のローカルメモリとメインメモリの間でどの様に一貫性を保つかについても、議論の余地がある。従って、本研究もアイディアとしては面白いものの、実用化までにはまだ長い道のりがありそうである。

5　おわりに

次回、SC18は2018年11月11日から16日の日程でテキサス州ダラスにて開催される予定である。