一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
VLD2018-81,CPSY2018-91,RECONF2018-55(2019-01)

[招待講演] 大規模 PC クラスタ技術
－これまでの 20 年と今後の展望－

中島　耕太†

†（株）富士通研究所 〒211-8588 神奈川県川崎市中原区上小田中 4-1-1
E-mail: †nakashima.kouta@fujitsu.com

あらまし　PC クラスタは，現在主流の HPC システムの構成方法であり，約 20 年にわたって開発されてきた．PC クラスタは, x86 アーキテクチャの CPU, DRAM, GPGPU といったコモディティハードの組み合わせで構成される．このため，コストパフォーマンスに優れる．また，コモディティハードの性能向上により，PC クラスタシステム全体の性能も向上してきた．一方で，主要な部品がコモディティであるため，それを組み合わせたシステムの性能問題の根本原因を分析することが困難であるため，性能分析技術が必要となる．また，部分的に故障が含まれるネットワークでも安定運用させる必要があるため，ネットワークの管理技術が必要となる．本発表では，大規模 PC クラスタ技術の 20 年間の発展について説明し，その中で，性能分析技術やネットワーク管理技術について説明する．
キーワード　PC クラスタ，性能分析技術，ネットワーク管理技術

Large Scale PC Cluster Technologies
－20 years and future perspectives－

Kohta NAKASHIMA†

† Fujitsu Laboratories Ltd. 4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki, 211-8588 Japan
E-mail: †nakashima.kouta@fujitsu.com

Abstract　PC cluster systems, which are mainstream as major supercomputer implementation, have been developed for 20 years. Major components of PC cluster system are commodity based hardware such as x86 CPU, memory modules, GPUs and so on. Commodity-based hardware not only provides cost-effective systems but also improves total system performance with commodity component performance improvement. However, it is hard to find root-cause of system performance issue because the systems are combination of various kind of commodity hardware. To resolve these problems, system performance analysis technologies are required. Moreover, in order to stabilize large scale network with some failure components, network management technologies for large scale PC cluster are required. This paper describes development of large scale PC cluster technologies for 20 years including performance analysis technologies and network management technologies.
Keywords　PC Cluster，Performance analysis，Network management

1. はじめに

PC クラスタは，現在主流の HPC システムの構成方法であり，Intel アーキテクチャ CPU を搭載した PC サーバを多数接続することにより構成される．HPC システムの性能ランキングである Top 500[1]によると，2018 年 11 月時点で 500 システム中 480 システムが PC クラスタである．

PC クラスタは CPU や DRAM は，x86 アーキテクチャ CPU に代表される汎用的に使われるコモディティ部品により構成される．これにより，コストパフォーマンスに優れたシステムを構成することができる．コストパフォーマンスに優れるサーバ群を InfiniBand[2]や OmniPath といった高性能なインタコネクトで接続することにより，システム全体での並列性能を高めている．さらに，GPGPU のようなアクセラレータを活用することで，高い演算性能を達成している．

PC クラスタは約 20 年にわたって開発されてきた．特に 2000 年代以降, サーバ間を高速に接続する技術が広がったことから，HPC システムの主流のシステム構成となった．本格的に PC クラスタが HPC システムとして使われるようになると，性能や挙動における品質を高めることが重要になってきた．

PC クラスタは, 主要部品がコモディティ部材であるため，ハードウェアの構造の中身を完全に知ることはできない．またその製造ばらつきから初期段階においては動作や性能面においてある程度の割合で不良部材が混入する．このようなハードウェアを多数組み合わせてシステムを構成すると，動作や性能障害を引き起

こすが，その根本原因を見つけ出すのは困難であり，これを分析する技術が必要となる．また，サーバ間を接続するネットワークは大規模化するため，一部に故障が含まれるようなネットワークにおいても安定的に全体動作させることが重要である．

本発表では，大規模 PC クラスタ技術の約 20 年間の発展について述べ，その中で，実際に HPC システムとして PC クラスタを構築するにあたって必要となった性能分析技術やネットワーク管理技術について述べる．また，PC クラスタ技術の今後の展望について議論する．

2. PC クラスタの登場

PC クラスタ登場前は，HPC システムとして，まずベクトルプロセッサによるシステムの登場した後，複数のスカラ型プロセッサを共有メモリで結合した共有メモリ型プロセッサシステムや複数のプロセッサとメモリの組を高速インタコネクトで接続した分散メモリ型の超並列システムが登場した．このような HPC システムの変遷の理由の一つに，広く用いられているスカラ型プロセッサを結合することによりコストパフォーマンスを高めたいという要求があった．一方で，高い性能を達成するためには，高速に CPU 間を接続する必要があり，このための開発コストが高価である課題があった．

1990 年代に入ると，Beowulf project[3]や Network of Workstations[4]のように比較的安価な PC やワークステーションといったコンピュータを接続することによって構成された PC クラスタや WS クラスタの構築が試みられるようになった．当時の課題は，接続するネットワークの性能であった．文献[3]では，10Mbps の 10Base-T を 4 チャネル使用して接続するといった工夫により 16 ノードまでのスケーラビリティを実現しているが，高速化には限界があった．

Myrinet[5]が登場すると，Myrinet を利用した PC クラスタシステムの研究が盛んになった．Myrinet はネットワークカード上にプロセッサを搭載したインタコネクトであり，1.28Gbps の高バンド幅とプロトコルオフロード機能を実装することを可能とした．特に RDMA の実装を可能としたことが性能向上に大きな影響を与えた．著名な実装事例として SCore[6]プロジェクトにおける PM ライブラリがある．

このような PC クラスタ研究が進むにつれ，2000 年代にはいると，各社が PC クラスタを HPC システムとしてリリースするようになった．理化学研究所が導入した RIKEN Super Combined Cluster はその一例であり，1,024 台の Intel アーキテクチャサーバを接続して構成された．2004 年稼働当時，日本最大の PC クラスタとして Top500 において第 7 位にランクインするなど，本格的な HPC システムとしての地位を確立するようになった．

3. 性能分析技術

1,000 ノードを超えるような大規模な PC クラスタでは，システム全体の挙動を分析する技術が重要となる．PC クラスタは多数のコモディティ部品から構成されるため，実際に組み合わせると，予想通りの性能が達成できず，その性能問題の原因がどこにあるのか分析するのが困難になる．この問題を解決するための性能分析技術が開発された[6]．本技術では，各サーバの性能情報をクラスタリングすることにより，どこに問題があるかを見つける機能と，時系列に性能の変化を分析することで性能の間欠障害を見つけ出す機能が実現されている．これらの機能により，大規模な PC クラスタにおいてもその性能分析を容易にし，問題個所の特定を効率化している．

4. PC クラスタの複雑化

PC クラスタが HPC システムの主流になってくると，ファイルシステムを共有するために複数の PC クラスタを接続して構成した複雑なネットワーク運用が求められるようになった．実際のシステム運用においては，スイッチやケーブルといったネットワーク部品が一部故障するケースもあり，故障個所を回避しつつ，複雑なネットワークを安定運用する技術が求められた．この問題を解決するために，運用中のネットワークであっても安全に故障個所を回避する経路に切り替える技術[7]や，複数の Fat Tree トポロジーを接続したようなネットワークでもデッドロックを効率的に回避するルーティング技術[8]を開発した．

5. GPGPU とメニーコア

2008 年以降，GPGPU を活用した PC クラスタシステムが構成されるようになった．これは GPGPU がコストパフォーマンスだけでなく電力あたり性能においても優れているためである．東京工業大学が導入した Tsubame 1.2 は最初に Top500 に登録された GPGPU クラスタであり，以降，GPGPU を活用したシステムが多数利用導入されている．最近では 2016 年に稼働した産業技術総合研究所が導入した AI Bridging Cloud Infrastructure も大規模な GPGPU ベースの PC クラスタであり，導入当時 Top500 において第 5 位となる性能を記録した．

GPGPU はコストパフォーマンスに優れる一方で，そのプログラミングスタイルが一般の CPU と異なることから，プログラミングにおける敷居が高いという課題があった．電力効率を高めつつ高性能を実現し，さらにプログラミングの課題を解決する技術として，メ

ニーコアアーキテクチャが提案されており，その実装形態の一つとして Intel 社の Xeon Phi[10]がある．Xeon Phi は，Intel アーキテクチャの命令セットを処理できる小型のコアを 60 個以上搭載した CPU で，CPU と同じプログラミングスタイルのままプログラミングが可能となり，GPGPU の課題を解決する技術として注目された．この Xeon Phi を搭載したサーバを高速インタコネクトの一種である OmniPath によって 8,208 台接続して構成し JCAHPC(Joint Center for Advanced High Performance Computing)が導入した Oakforest-PACS は 2016 年稼働当時，Top500 において第 6 位になる等，大規模システムとしてメニーコアアーキテクチャが活用できることを実証した．

6. 今後の展望

2018 年において，PC クラスタは Top500 に登録されているシステムのうち 480 システムを占めており，今後も HPC システムの主流であり続けると考えられる．一方で克服すべき課題が 2 点ある．

一つ目の課題は，相対的なネットワーク性能の低下である．GPGPU やメニーコアアーキテクチャの登場により，演算性能は飛躍的に向上した．一方で，サーバ間を接続するネットワーク性能の向上は比較的緩やかであり，ネットワーク性能と計算性能のバランスが悪くなっている．ネットワーク性能が向上しない原因の一つとして，PCI Express 性能の向上の鈍化がある．PCI Express は InfiniBand や OmniPath といった高速インタコネクトとサーバを接続する I/O バスであるが，この性能の向上が停滞している．この課題の解決のアプローチとして，NVIDIA 社が開発する NVLink があり，IBM Power プロセッサとの直接接続を実現しているものの，Intel アーキテクチャのような広く普及している CPU との接続には用いられておらず，この課題の解決が重要である．

二つ目の課題は，プロセッサ性能の向上の鈍化である．Top 500 に登録されている 500 システムの合計値や 500 番目のシステム性能の向上を見ると，2013 年頃より性能の伸びが鈍化している．この鈍化の主要な要員の一つにムーアの法則の鈍化がある．これまでのように半導体のプロセスルールの進化が進まなくなってきており，この課題を解決する必要がある．解決の一つの手法としてある特定の用途に特化するドメイン特化コンピューティングがある．GPU も一種のドメインコンピューティングとも言えるが，Google 社が開発する TPU[11]のように，ニューラルネットワーク計算に特化することで高速化するものも登場している．今後はドメイン特化プロセッサを PC クラスタに適用することでシステム性能を向上させていく必要もある．

7. おわりに

本発表では，大規模 PC クラスタ技術の約 20 年間の発展について述べ，その中で，実際に HPC システムとして PC クラスタを構築するにあたって必要となった性能分析技術やネットワーク管理技術について述べる．また，PC クラスタ技術の今後の展望について議論した．

PC クラスタは，今後も HPC システムの主流であり続けると考えられるが，克服すべき課題として，ネットワーク性能の向上とドメイン特化プロセッサの活用がある．これらの課題を解決することで，今後も PC クラスタシステムの性能を向上させ，増え続ける計算需要に応えていくことが重要であるといえる．

文　献

[1] Top 500: http://www.top500.org.

[2] InfiniBand Architecture Specification Release 1.3: https://www.infinibandta.org.

[3] T. Sterling, D. J. Becker, J. E. Dorband, D. Savarese: Beowulf: A parallel workstation for scientific computation, International Conference on Parallel Processing, 1996.

[4] Thomas Anderson, David Culler, David Patterson, and the NOW Team. The Case for NOW (Networks of Workstations). IEEE Micro vol. 15, no. 1, February 1995, pages 54 - 64.

[5] N. J. Boden, D. Cohen, R. E. Felderman, A. E. Kulawik, C. L. Seitz, J. N. Seizovic and Wen-King Su. “Myrinet - A Gigabit-per-Second Local-Area Network”. IEEE MICRO, 15(1):29–36, 1995.

[6] Ishikawa Y., Tezuka H., Hori A., Sumimoto S., Takahashi T, O’Carroll, and Harada H.: RW PC Cluster II and Score Cluster System Software – High Performance Linux Cluster, Proceedings of the 5th Annual Linux Expo, 1999.

[7] 山村周史，平井聡，小野美由紀，松本和宏，住元真司，久門耕一，“時系列データの統計解析手法による PC クラスタシステム解析手法の提案,” 情報処理学会論文誌コンピューティングシステム(ACS), 47(SIG12(ACS15)), 250-262, 2006

[8] 中島耕太，久門耕一，成瀬彰，住元真司，“大規模 InfiniBand システムにおける経路更新手法の提案,” 電子情報通信学会技術研究報告．CPSY, コンピュータシステム 109(168), 67-72, 2009.

[9] 中島耕太，成瀬彰，住元真司，久門耕一，“通信量バランスの良いデッドロック回避ルーティング手法の提案とクラスタネットワークにおける評価,” 情報処理学会論文誌コンピューティングシステム (ACS), 4(4),191-202, 2011.

[10] Avinash Sodani, Roger Gramunt, Jesus Corbal, Ho-Seop Kim, Krishna Vinod, Sundaram Chinthamani, Steven Hutsell, Rajat Agarwal, Yen-Chen Lui, “Knights Landing: Second-Generation Intel Xeon Phi Product,” IEEE Micro vol: 36 Issue: 2, 34-46, 2016

[11] Norman P. Jouppi, et. al, “In-Datacenter Performance Analysis of a Tensor Processing Unit,” Proceedings of the 44th International Symposium on Computer Architecture (ISCA), 2017.