汎用サーバ8千台で、スーパーコンピュータの傑作をつくる！大規模PCクラスタ構築技術（前編）

【未来を創るチカラ Vol.12】先端コンピュータシステムプロジェクト プロジェクトリーダー、中島 耕太

FUJITSU JOURNAL / 2019年2月12日

スーパーコンピュータの主流は今、汎用CPUを搭載したPCサーバを何台も接続する"PCクラスタ"型。その規模が大きくなるほど、PC単体（ノード）も、大量のノード同士を繋ぐ複雑なネットワークも高速化と安定化が課題になります。富士通の研究チームはこの難題に挑戦し、東京大学・筑波大学が共同で導入した"Oakforest-PACS"の開発で世界トップ級の性能を実証。このような大規模PCクラスタ構築技術が認められて2018年4月には文部科学大臣表彰の科学技術賞を受賞しました。チームを率いる中島耕太に、不可能を可能に変える舞台裏や、研究にかける思いを聞きます。

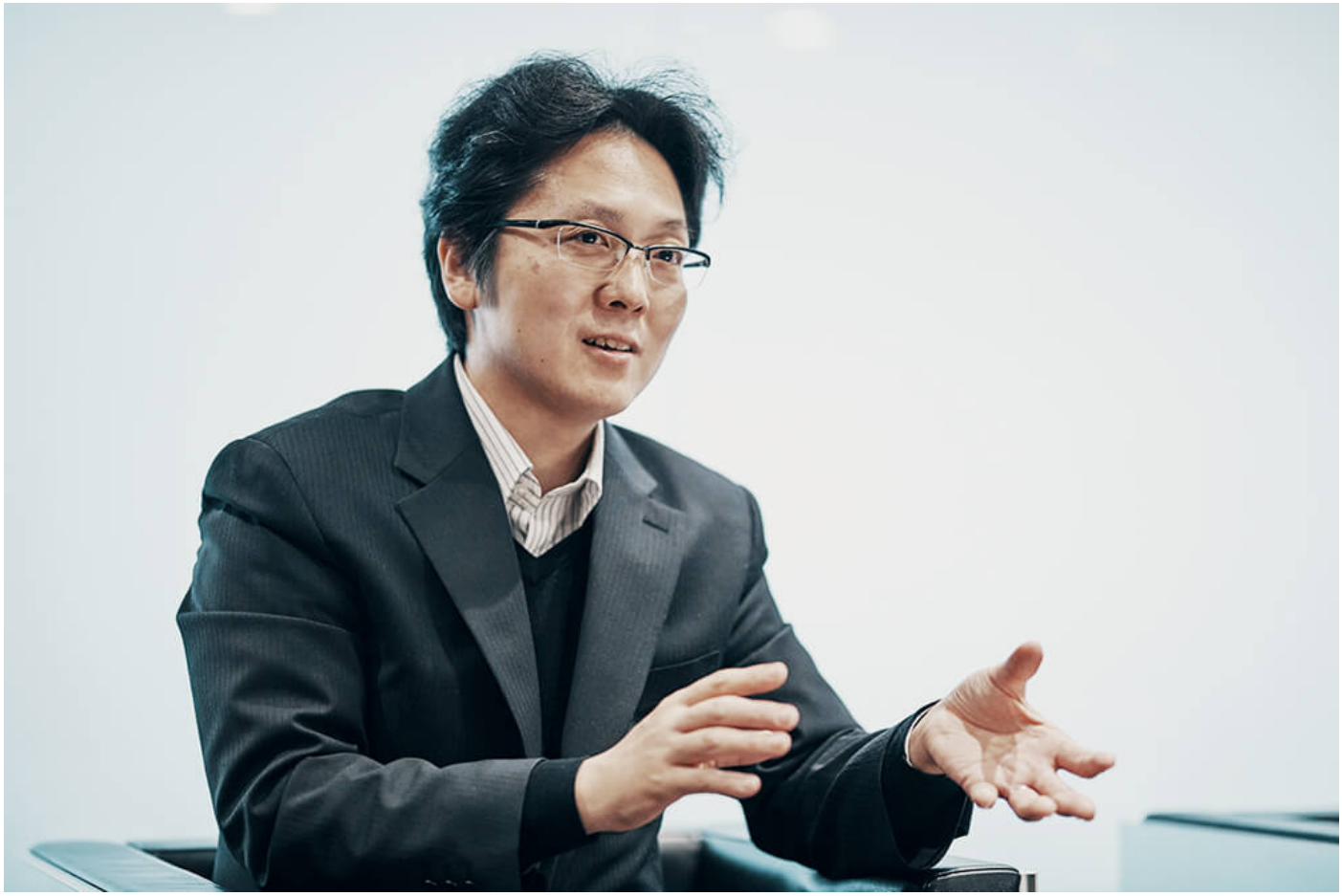
小さなコンピュータを束ねて、大きな性能を引き出す

中島　「PCクラスタというのは、昔はパーソナルコンピュータをたくさん並べて繋ぎ、全体を束ねて一つのスーパーコンピュータにしようというものでした。現在は"PC"という名は付いていますが、基本はコモディティサーバと呼ばれる汎用プロセッサ、つまり世の中でよく使われているコストパフォーマンスのよいプロセッサで作ったサーバをたくさん使います。パソコンよりちょっと豪華な構成のサーバをたくさん並べるという感じです。

サーバ単体は、計算ノードと言いますが、例えばこれ（写真上）はインテルCPUを積んでいるサーバ、計算ノードです。1箱にはノードが8個入ります。東大、筑波大に置かれているOakforest-PACSというPCクラスタは、8,208台ノード。この箱が約千台並んでいます。事業部で作られたこのサーバを束ねて、どうやって性能を引き出すかが私たちの技術のポイントになります。」

研究人生は、PCクラスタの歴史とともに

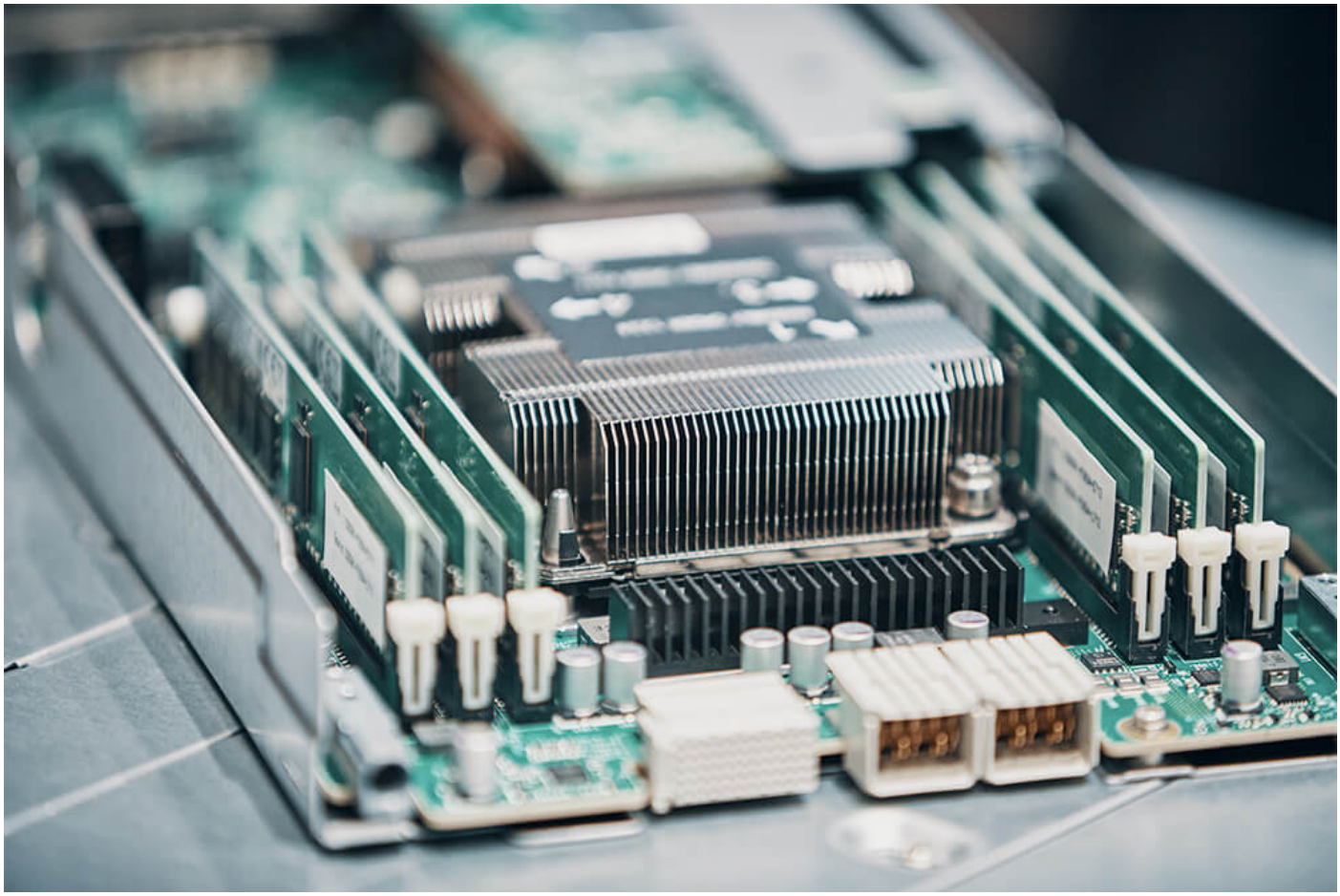
「PCクラスタは、1990年代後半より研究が盛んになった領域です。ちょうどこの時期大学生だった私は、コンピュータの台数を増やすと性能が上がるというPCクラスタの技術にとても関心を持っていました。研究室では、ネットワークで外から入ってきたデータを、できるだけ速くメモリに転送するにはどうすればいいかというソフトウェア技術を研究していました。



富士通研究所に入社した2002年頃は、当時技術的な大きな壁を破った、コンピュータ間を高速に接続する技術（InifniBand/Myrinet）が世の中に広がり始めていました。PCクラスタが一気にスーパーコンピュータの構築手法として主流になっていく中で、富士通が理化学研究所に納入したRSCCというPCクラスタが当時世界7位のスーパーコンピュータとなって、新入社員だった私は、先輩方が進めるプロジェクトを横から見ながら、凄い技術だなと思っていました。」

現場でぶつかった課題から、新しい技術が生まれる

「大規模なPCクラスタを実際に構成してみると、やはり汎用品なので性能品質にばらつきがあります。千台もあれば数台に初期不良もあります。まずそれぞれの部品がどういう挙動や性能を示すか、どこが遅いのか調べなくてはなりません。しかも恒常的ではなく間欠的に悪くなる、どの部品がおかしいのか切り分けることも厄介です。大規模なクラスタを複雑に組み合わせると、これをどう制御するかも課題です。システム構成を組み、最終的に全体で並列計算させた時の性能を上げる方法を考えなくてはなりません。こういった現場での課題が、性能分析技術やネットワーク制御技術の開発に繋がりました。」



Oakforest-PACSの挑戦。理論通りに現実はいかない

「使用する部品が最先端のものであればあるほど、ちゃんと動かした上で性能を引き出すのが難しくなりますし、規模が大きくなればなるほど、些細な問題が引き金で全体の挙動・性能を悪化させます。2016年に稼働させたOakforest-PACSは、最も困難なPCクラスタでした。
当時最先端だったIntel社製のXeon Phiというメニーコアプロセッサを使用したのですが、現在主流のプロセッサがせいぜい20コアくらいなのに対して、68コアも持つ非常にチャレンジングなプロセッサで、また高バンド幅メモリと大容量メモリの2種類のメモリを持った過去にあまりないアーキテクチャでした。
これを使いこなすのが富士通の役割です。プロセッサの特徴を事前に評価し、大規模になった時の課題を想定して構築準備を進めました。ところが実際に8,208台も接続すると、理屈では動くはずなのに、全然動かない（笑）。これまで想定したことのない問題が見えてきます。現場のSEさんやCEさんと相談しながら『ここが悪そうだね』というところを見つけて、分析技術を使って実際に切り分け、サーバを作っている事業部門やプロセッサを作っているIntel社と一緒に構築していく必要があります。いかに状況を分析してわかりやすく整理するかが問題解決のポイントでした。」

未知の難問を解くカギは、積み上げた過去にある

「これまで想定したことのない問題に直面すると、分析の方法から試行錯誤で作っていかなくてはなりません。こういった困難を乗り越えられたのは、これまでの技術の蓄積のおかげです。例えば、Oakforest-PACSでは、その前の大きなプロジェクトで得られた開発経験と実システムでのネットワークの挙動分析の経験によって、構築時のOakforest-PACSで起こっている問題を切り分け、突貫で分析のためのソフトウェアを作っていきました。

やっている間は本当に成功する気がしないくらい、一時は絶望的な状況もあったのですが、最後は何が何でもやってやるぞという気概でしたね。私たちの研究チームはこのプロジェクトに賭けていましたので、一致団結して進められましたが、これを他部門とも共有していく必要がありました。絶望的な状況の時こそ率先して問題解決にあたる姿勢を見せたことで、全員で一致団結して進めていくのに微力ながら貢献できたかなと思います。」

日米のチーム全員が、各地で歓声をあげた

「スーパーコンピュータのランキングはTOP500、つまり世界の500位までのリストが毎年二回更新されて、そこに載せるというのが一つのモチベーションになっています。Oakforest-PACSでは、世界6位の値が取れました（2016年11月）。

プログラムが完走して最初に目標性能を達成した瞬間は、Oakforest-PACSが置かれている柏キャンパスの現場や、今いる川崎の研究所、事業所、それぞれの家、インテル側はアメリカの東海岸と西海岸など、何十名もの全メンバーが電話でつながっている状態でした。結果を知らせた時には、『やったー！！』という感じで、海の向こうからもすごい拍手が起こって、その瞬間はとても良かったですね。」

後編では、イノベーションを起こすチームについて、またAIに関わる現在の中島の研究について聞きます。

株式会社富士通研究所
コンピュータシステム研究所
先端コンピュータシステムプロジェクト
プロジェクトディレクター
中島 耕太

2000年九州大学工学部電気情報工学科、2002年同大学大学院修士課程を修了。
同年富士通研究所に入社。
以来、PCクラスタ高速化技術、高速ネットワーク制御技術の研究開発に従事。
博士（工学）、平成30年度文部科学省科学技術賞開発部門「大規模PCクラスタ構築技術の開発」を他4名と受賞。

FUJITSU JOURNAL / 2019年2月12日