複雑化・巨大化するAI処理を高速化するContent-Aware Computing技術とは

白幡　晃一　　原　靖　　坂井　靖文　　三輪　真弘　　田渕　晶大　　高　虹

あらまし

複雑化・巨大化する計算ニーズに対応するコンピューティング技術として，富士通はCAC（Content-Aware Computing）技術の研究開発を進めている。これは，処理内容とデータの分析に基づいたソフトウェアの最適化により高速化を行う技術である。計算量の動的削減，並列学習高速化，I/O高速化からなるCAC技術を組み合わせることにより，2020年11月には大規模機械学習処理のベンチマークであるMLPerf HPCで処理速度世界第1位を達成した。

本稿では，CAC技術とその適用事例について述べる。

1. まえがき

IoTによるデータ量の爆発的な増加と，AIによるデータ分析技術の高度化により，データを新たな資源として活用し社会や産業を変革するデータ駆動型社会が到来しつつある。膨大なデータの高度な分析には，複雑化・巨大化を続けるAIモデル学習のためのコンピューティング能力が重要となる。例えば，文章生成言語モデルGPT-3は1,750億個のパラメーターを持ち，1台のGPUで学習した場合，355年かかるとも言われている [1]。

このような複雑化・巨大化する計算ニーズに対応するために，富士通はCAC（Content-Aware Computing）技術の研究開発を進めている [2]。CAC技術とは，処理内容とデータの分析に基づいてソフトウェアの高速化を行う技術である。

本稿では複雑化・巨大化するAI処理を高速化するCAC技術とその適用事例について述べる。

2. AI処理高速化への課題

複雑化・巨大化を続けるAIモデルを高速に学習させるためには，コンピューターのハードウェア性能と，その性能を最大限に活かすソフトウェア技術が重要となる。今回は，AI処理の中でも特に膨大な計算量を必要とするディープラーニングを主な対象とする。

ディープラーニングでは，画像・音声・文章などのデータを多層ニューラルネットワーク（DNN）に与え，分類や回帰といったタスクを学習させることで，画像・音声の認識や文章の翻訳などを行うモデルを獲得する。

ディープラーニングでは，計算量が膨大である一方，高い演算精度は必要とされないため，演算精度を最適に制御して計算量を削減することが重要となる。また，複数のコンピューターを同時に用いた並列計算では，学習精度を維持しながら高効率で計算することが重要となる。更に，膨大なデータを並列ファイルシステムからローカルディスクにコピーする時間や，ローカルディスクからメモリへの読み出しによる遅延時間を短縮することも重要となる。

3. AI処理を高速化するContent-Aware Computing技術とは

CAC技術 [3] は，処理内容を分析・軽量化し，コンピューターへの処理の割当を最適化することで高速化を行うソフトウェア技術である。CACの概要を図-1に示す。様々なアプリケーションに対して，処理内容を分析しながら，使用するハードウェアに応じて計算量・並列処理・I/Oを自動的に最適制御することで，最大10倍の学習の高速化が可能となる。

本章では，CAC技術について，計算量の動的削減技術，並列学習高速化技術，I/O高速化技術の順に述べる。

3.1 計算量の動的削減技術

ディープラーニングの計算量は膨大である一方，高い演算精度は不要なため，演算精度を最適に制御して計算量を削減することが学習時間の短縮に有効である。計算量の動的削減技術の概要を図-2に示す。学習精度を維持しながら計算量を削減する技術として，ビット幅削減技術，Gradient Skip技術，Pruning，および計算科学シミュレーション高速化を紹介する。

(1) ビット幅削減技術

数値計算を高速化する手法の一つに，ビット幅削減技術がある｛図-2 (a)｝。一般的に32ビットで表

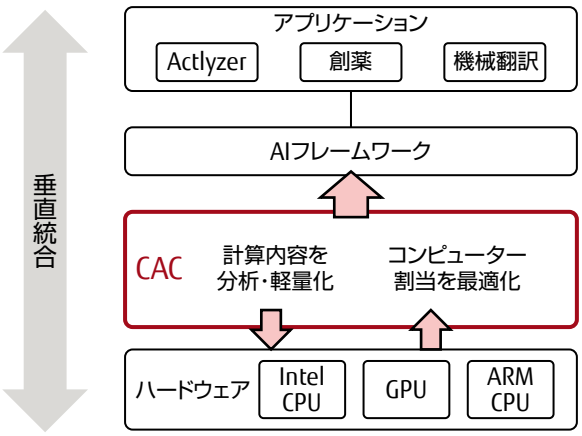


図-1　Content-Aware Computingの概要

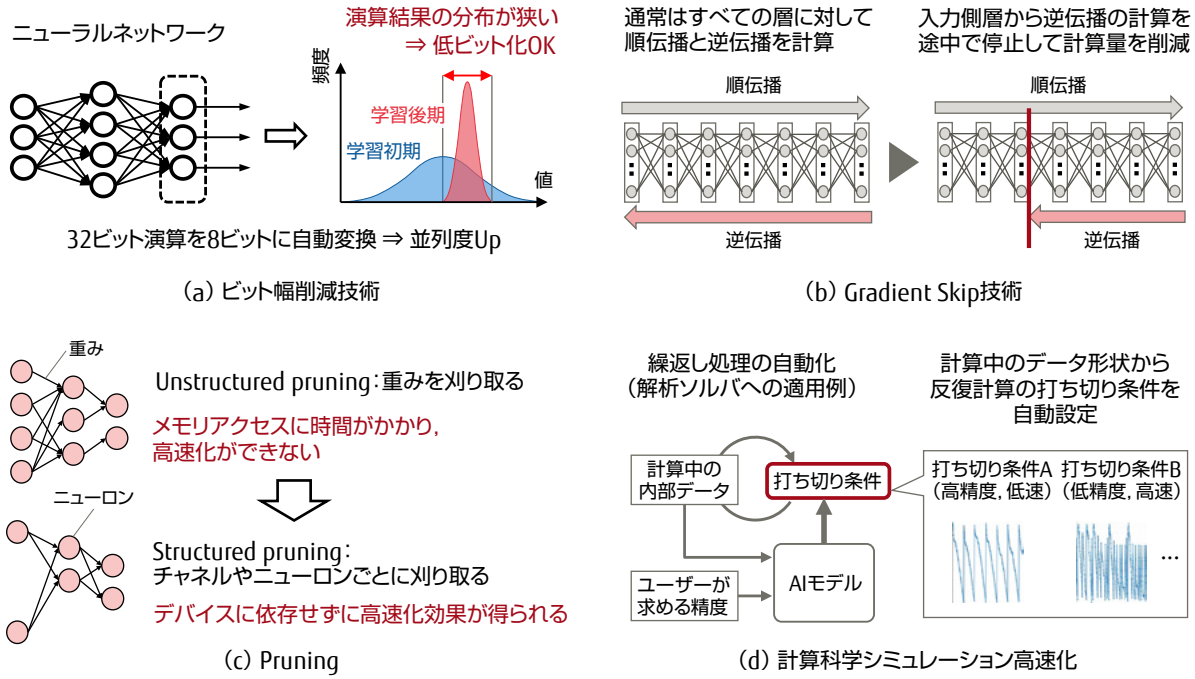


図-2　計算量の動的削減技術

現される数値データを16ビットや8ビットで表現することでデータサイズを減らし，演算速度や通信速度を高速化する技術である。

ディープラーニングにおいて安易にビット幅を削減すると，DNNの精度が著しく劣化する場合がある。従来，精度劣化を防ぐために専門家によるビット幅調整が必要であり，ビット幅削減技術の適用のハードルが高かった。そこで，精度劣化しないビット幅を自動的に決定する技術を開発し，誰でもビット幅削減技術を簡単に利用できるようにした。代表的な画像分類タスクであるImageNetを，画像処理DNNであるAlexNet，ResNet-18およびResNet-50の学習に適用したところ，分類精度を大きく劣化させることなくビット幅を自動的に決定できた。また，AlexNetおよびResNet-18で3.5倍，ResNet-50で2.5倍の学習の高速化を実現した。

(2) Gradient Skip技術

DNNは幾つかの層を組み合わせたネットワークである。学習時の処理には，推論確率を出力する順伝播と，学習におけるパラメーター（重み）に対する修正量（誤差勾配）を算出する逆伝播がある。学習が十分進んだ層では重みの修正が不要となる。また，画像処理DNNでは，入力側の層の学習が早く進むことも判明した。このため，Gradient Skip技術では，学習が十分進んだ入力側層から逆伝播の計算を徐々に停止して計算量を削減することで，高速化を実現する｛図-2 (b)｝。

また，重みの修正を急に停止すると学習最終到達精度が若干低くなることも実験から判っており，修正値をなめらかに0に持って行く精度保証技術を開発した。これにより，Gradient Skip技術の適用による精度劣化は無視できる程度まで抑えられた。実際に，MLPerf HPCのベンチマークの一つである気象データから異常気象を識別するDeepCAMに適用し，最大で1.8倍の高速化を実現した。

(3) Pruning

認識精度の高精度化と複雑なパターン認識を実現するため，DNNの規模は増大し続けている。しかし，学習によって得られたニューラルネットワークには，ニューロン同士のつながりの強さを示す重みが0と見なせる接続がある。このような接続に対する演算（乗算）は結果を0と見なせることが判っているため，計算量を削減できる。Pruning（プルーニング，枝刈り）技術 [4] は，DNNにおいて重みの値が小さな接続を自動に判別し，削除する技術である｛図-2 (c)｝。

Pruningの実装には課題がある。単純な実装例として，メモリから重みを読み出し，0と見なせる場合は演算をスキップするという方法が考えられる。しかし，メモリアクセスに時間がかかり，Pruningによる高速化ができない。演算を飛ばすべきデータがバラバラに存在するためである。今回は，重みの最小のまとまり（チャネル）の絶対値の総和を求め，総和の小さいチャネルを削除した。これにより，不要なメモリアクセスが発生せず，並列処理も活かせ，高速化可能となった。

(4) 計算科学シミュレーション高速化

構造解析，流体解析，分子動力学計算などの計算科学シミュレーションにおいて，詳細な結果を得るための大規模計算や，設計の自動化などを行うための大量のパターンに対するシミュレーションを素早く行う高速処理のニーズが高まっている。

このような計算では，反復計算においてユーザーが求める精度に効率的に達する打ち切り条件の設定が重要となる。従来は，打ち切り条件の設定はユーザーに任せられていた。今回，AIモデルを利用することで，計算中のデータに対しわずかな計算量で精度達成を判定する技術を開発した｛図-2 (d)｝[5]。

また，シミュレーションの過程で得られるデータから入力と出力の関係を学習させることで得られる代理モデル（サロゲートモデル）の活用も進みつつある。シミュレーションの計算を学習済みの代理モデルで置き換えることで，結果を瞬時に得られる。

3. 2 並列学習高速化技術

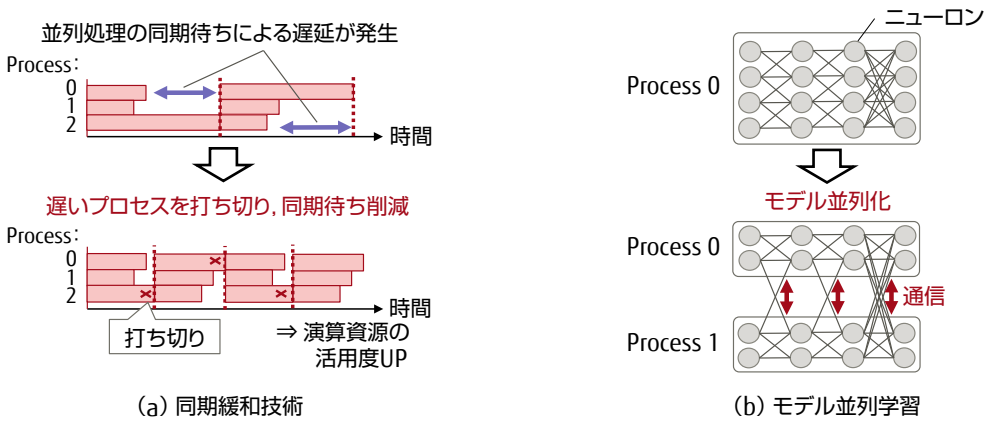
多数のコンピューターを同時に用いた大規模並列学習では，高い効率を保ったまま並列度を高めることが重要となる。並列学習高速化技術の概要を図-3に示す。処理に遅れが生じたプロセスに対する同期待ち時間を削減することで効率を維持しながら計算する同期緩和技術と，データ並列の限界を超えた並列度を実現するモデル並列学習について紹介する。

(1) 同期緩和技術

図-3 (a) に同期緩和技術の概要を示す。ディープラーニングの分散学習において，学習中に処理速度が遅いプロセスが発生すると，学習の反復ごとに必要な同期の待ち時間により全体の処理が遅延する。そこで，処理速度が常に最大化されるように遅いプロセスを動的に切り離す技術を開発し，速度低下を抑えながら学習できるようにした [6]。ResNet-50を用いたImageNetの画像分類において，25％のプロセスを削減してもほとんど精度を保ったまま学習できることを確認した。

(2) モデル並列学習

図-3 (b) にモデル並列学習の概要を示す。モデル並列は1つのDNNモデルを複数に分割し，コンピューターに分散して学習する手法である。学習速度向上の他，1台のコンピューターに載らない巨大なDNNモデルの学習を実現する。複数のコンピューターで異なる学習データを同時に処理するデータ並列では一度に処理するデータ量（バッチサイズ）が増加する。バッチサイズが増加すると，学習によって得られるモデルの精度が低くなる。モデ



(a) 同期緩和技術　　(b) モデル並列学習

図-3　並列学習高速化技術

ル並列ではバッチサイズは変わらないため，並列化による精度への影響がないのが特徴である。

モデル並列におけるモデル分割方法は複数ある。例えば，入力データの空間次元（画像であれば縦横）での分割や，入力データのチャネル（画像であればRGB）での分割などがある。入力データやDNNの層，分割方法によって，コンピューター間の通信の種類や最大並列数が変わるため，状況に応じた分割方法を選択する必要がある。モデル並列の適用には専門知識が必要とされるが，データ並列と組み合わせることで学習を更に高速化できる。我々は，4章で述べるMLPerf HPCに適用し，4分割モデルで1.8倍の高速化を実現した。

3.3 I/O高速化技術

ディープラーニングでは膨大なデータを用いるため，I/O高速化技術が必要となる。ここでは図-4に示した2つの技術，膨大なデータを並列ファイルシステムからローカルディスクに移動させるデータステージング高速化と，計算時にローカルディスクからメモリへのデータ移動の時間を隠蔽するI/Oボトルネック解消について紹介する。

(1) データステージング高速化

多数のコンピューターを利用した学習では，リモートストレージからの読み出しが競合し遅延する場合がある。これは，ステージングによりローカルストレージへのアクセスに置き換えることで軽減できる。更にステージングを高速化するには，高性能なストレージやネットワークを用いてスループットを上げるだけでなく，データを圧縮して転送量を減らすことも有効である。特に学習データを予め圧縮して保存可能な場合，高圧縮フォーマットにより転送時間の大幅な短縮が期待できる。データの展開は転送とオーバーラップできる他，データを多数のコンピューター間に分散してステージングすることで展開を並列化し，展開時間を短縮した。

(2) I/Oボトルネック解消

メモリに載り切らない学習データは，ローカルのHDDやSSDなどのストレージに保持しておき，必要に応じてメモリに読み出して学習を行う。しかし，必要になってから学習データを読み出し始めるのでは，読み終わるまで学習が中断してしまう。そこで，次に学習するデータをストレージから先読みすることで，学習の中断無しに効率良く実行できるようにした。

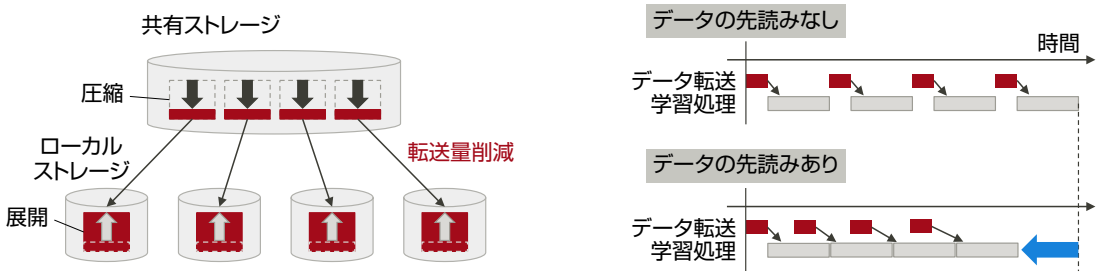
4. 適用事例

本章では，CAC技術の適用事例について述べる。

4.1 MLPerf HPC

MLPerf HPCは機械学習ベンチマークとして広く使われているMLPerfのHPC向けベンチマークである[7]。ダークマターデータから宇宙論的パラメーターを推定するCosmoFlowと，3.1 (2) 節で既出のDeepCAMの二つが含まれている。それぞれステージングを含めた学習時間を計測する。我々は産業技術総合研究所のAI橋渡しクラウド基盤「ABCI」と理化学研究所の「富岳」に対してMLPerf HPCの高速化を行った[8]。

ABCIにおいては，CosmoFlowとDeepCAMの両方でステージングを行い，CosmoFlowではデータ



(a) データステージング高速化　　(b) I/Oボトルネック解消

図-4　I/O高速化技術

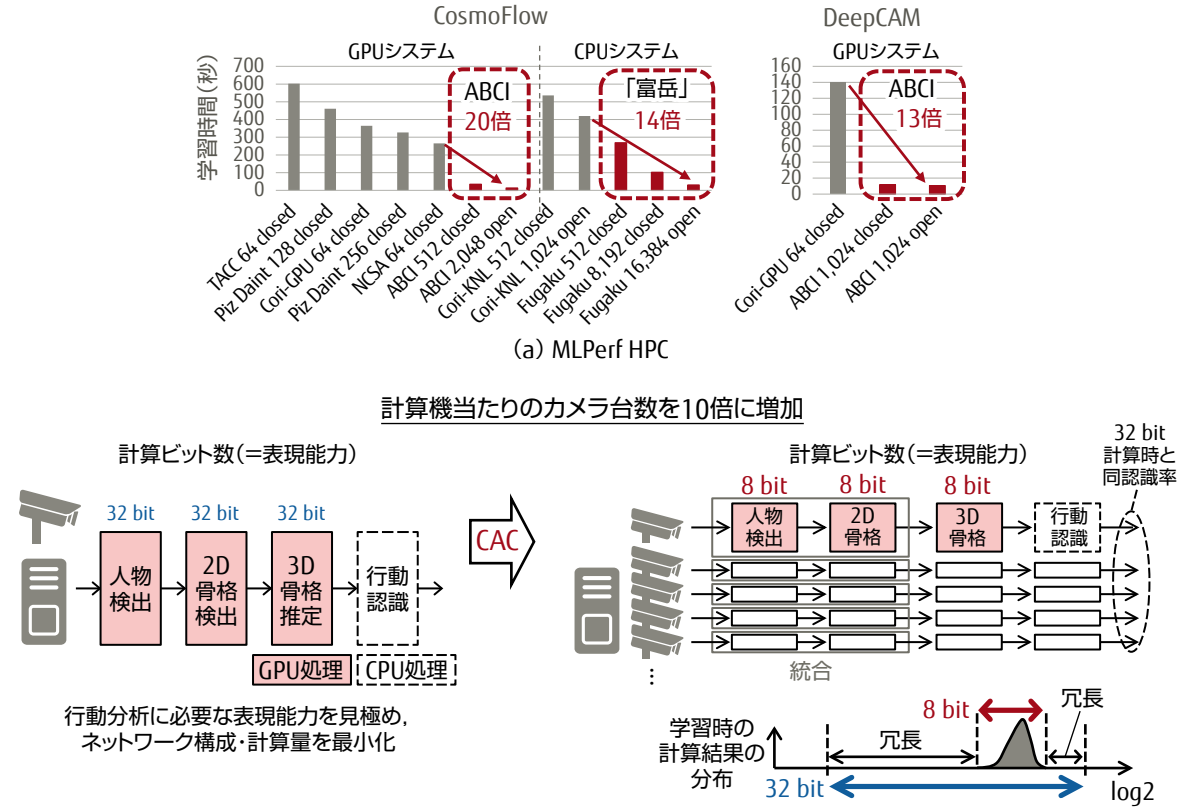


図-5　CAC技術の適用事例

圧縮を適用した。これにより，ステージング時間を最大1/12に短縮した。更にI/Oボトルネックを解消することで，学習効率を最大20％改善できた。またDeepCAMにGradient Skip技術を適用し学習時間を更に10％短縮した。

「富岳」においては，CosmoFlowのみを対象にし，同様にデータステージング高速化・I/Oボトルネック解消を行った。「富岳」のコンピューター1台当たりの性能はABCIのそれより低いが，より多くのコンピューターを利用できるため，データ並列に加えてモデル並列も適用し，使用するコンピューター数を最大16倍増加させた。

CACの適用によってMLPerf HPC v0.7ではABCIはCosmoFlowとDeepCAMでGPUタイプの他システムと比較してそれぞれ20倍と13倍の性能を達成し，「富岳」はCosmoFlowでCPUタイプの他システムと比較して14倍の性能を達成し，世界第1位，第2位を独占した｛図-5 (a)｝。

4. 2　Actlyzer

Actlyzerは富士通と富士通研究開発中心有限公司が開発した行動分析技術である [9]。この技術を活用したソリューションの価値を高めるためには，1台のコンピューターで処理可能なカメラ台数を増やす必要があった。そこで，行動分析の計算の無駄を見つけ，Actlyzerの構成要素である人物検出と骨格検出の統合により計算量を最小化にした。更に，32ビットから8ビットへのビット幅削減技術による計算の効率化と，pruning技術による人物と骨格検出のモデルを軽量化することで，コンピューター当たりの処理可能なカメラ台数を10倍に増加させた｛図-5 (b)｝。

5.　まとめと今後の予定

本稿では，複雑化・巨大化するAI処理を高速化するCAC技術とその適用事例について紹介した。

計算量の動的削減技術，並列学習高速化技術，

I/O高速化技術からなる高速化技術を組み合わせることにより，MLPerf HPCベンチマーク処理速度世界第1位を達成した。また富士通の行動分析技術Actlyzerへの適用によりコンピューター当たりのカメラ台数を10倍に増加させた。

今後は，CAC技術の完成度を高め，更なる高速化や使いやすさの向上を進めていくとともに，医療，創薬，材料，物流など様々な分野に適用領域を拡大し，社会課題の解決に貢献していく。

参考文献・注記

[1] Lambda：OpenAI's GPT-3 Language Model: A Technical Overview.
https://lambdalabs.com/blog/demystifying-gpt-3/
[2] 富士通研究所：計算の厳密性を自動調整しAI処理を最大10倍高速化するコンピューティング技術を開発.
https://pr.fujitsu.com/jp/news/2019/10/25-1.html
[3] 富士通研究所：計算精度を自動調整する「Content-Aware Computing」により、AI処理を10倍高速化.
https://www.fujitsu.com/jp/group/labs/about/resources/article/202002-cac.html
[4] D. Blalock et al.：What is the state of neural network pruning?
https://arxiv.org/abs/2003.03033
[5] A. Haderbache et al.：Acceleration of Structural Analysis Simulations using CNN-based Auto-Tuning of Solver Tolerance.
https://ieeexplore.ieee.org/document/9150415
[6] K. Shirahata et al.：Preliminary Performance Analysis of Distributed DNN Training with Relaxed Synchronization.
https://www.jstage.jst.go.jp/article/transele/advpub/0/advpub_2020LHS0001/_article/-char/ja/
[7] ML Commons：機械学習のイノベーションをすべての人へ.
https://mlcommons.org/ja/
[8] 富士通：機械学習処理ベンチマークMLPerf HPCにて最高レベルの速度を達成.
https://pr.fujitsu.com/jp/news/2020/11/19-1.html
[9] 富士通研究所：映像から人の様々な行動を認識するAI技術「行動分析技術 Actlyzer」を開発.
https://pr.fujitsu.com/jp/news/2019/11/25.html

著者紹介

白幡　晃一（しらはた　こういち）
富士通株式会社
研究本部
Content-Aware Computingの研究に従事。

原　靖（はら　やすし）
富士通株式会社
インフラストラクチャシステム事業本部
Content-Aware Computingの研究に従事。

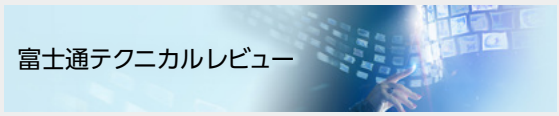
坂井　靖文（さかい　やすふみ）
富士通株式会社
研究本部
Content-Aware Computingの研究に従事。

三輪　真弘（みわ　まさひろ）
富士通株式会社
研究本部
Content-Aware Computingの研究に従事。

田渕　晶大（たぶち　あきひろ）
富士通株式会社
研究本部
Content-Aware Computingの研究に従事。

高　虹（こう　にじ）
富士通株式会社
研究本部
Content-Aware Computingの研究に従事。

この記事は，富士通の技術情報メディア「富士通テクニカルレビュー」に掲載されたものです。他の記事も是非ご覧ください。



https://www.fujitsu.com/jp/technicalreview/