IBM Data Engine for NoSQL - Power Systems Edition

ホワイト・ペーパー

Brad Brech – Juan Rubio – Michael Hollinger
IBM Systems Group

2015 年 4 月

目次

1　エグゼクティブ・サマリー

近年、モバイル・ユーザーの期待に応えるために、新しい顧客対応アプリケーションで、かつてないほど優れた応答時間と規模が必要となっているため、NoSQL の採用が爆発的に進んでいます。標準的な NoSQL 実装は、インメモリーでの稼働か、キャッシュとしてのメモリーに大きく依存するため、コストが高くなり、拡張が困難です。また、従来型のストレージ IO の遅延はアプリケーション要件に適合しません。NoSQL アプリケーションの拡張性をサポートするには、新しいソリューションが必要です。

IBM Data Engine for NoSQL - Power Systems Edition は、従来型の IO ストレージの遅延の問題を発生させることなく、最大 57 テラバイトのフラッシュ・メモリーをプロセッサーに接続することにより、新しいメモリー層を作成します。DRAM ほど高速ではありませんが、特にネットワーク経由でデータにアクセスする場合、遅延は大半のアプリケーションの許容範囲内です。フラッシュは、DRAM より大幅に低コストでもあり、カスタマー・ソリューションを提供するための導入コストと運用コストの削減に役立ちます。お客様、MSP、ISP はすべて、NoSQL アプリケーションエリアおけるこの新しいテクノロジーの応用のメリットを享受します。IBM の主力製品である POWER8 のオープン・アーキテクチャーに組み込まれたハードウェアとソフトウェアを活用すると、お客様はソリューションに関して「規模」と「速度」のどちらかを選ばずにすむことを意味します。

2　ビジネス上の問題

SQL データベースでは常にアプリケーション応答時間が重要な要素であったため、業界では、その最適化に相当な労力が費やされてきました。しかし、クラウド・ソリューション・アーキテクチャーを中心に構築されたモバイル・アプリケーションが急激に普及していることから、規模、レジリエンシー、シンプルさの点で NoSQL データベースの採用が進んでいます。

これまでは、比較的高いコストがかかるメモリー[1]と DRAM の保持に必要な、スケールアウト型ノードの数が原因で、NoSQL データベースの総所有コストは非常に高いものでした。このように導入コストが高額であることから、Redis のような NoSQL 実装の採用は、データ量が比較的小容量のアプリケーションや、アプリケーションの中でも超高速なパフォーマンスを絶対的に必要とする部分だけに限られていました。

IBM と Redis Labs は、CAPI (Coherent Accelerator Processor Interface) インターフェースを介して IBM のオープンな POWER8 プロセッサーに接続されたフラッシュを活用するという独自の方法を認識し、この問題に対応すべく協業しました。IO バス上でソリッド・ステート・ドライブ (SSD) としてフラッシュを使用する NoSQL ソリューションは、回転式ディスクよりも優れたパフォーマンスを発揮しますが、RAM と比較すると入出力オーバーヘッドが高いために、求められている遅延時間に対応することはできません。フラッシュ DIMM も同様に、システムサイズ、レジリエンシー、その他のパフォーマンスの面で制限があります。IBM と Redis Labs は、DRAM と CAPI 接続フラッシュの組み合わせを使用することによって、導入コストと運用コストの両方を大幅に削減すると同時に、より高速、大規模、スケーラブルなアプリケーションを稼動できるソリューションを提供できます。

12 TB のデータベースの場合、導入コスト (TCA) は従来の導入コストの 3 分の 1 になり、ソリューションに必要なノード数を 24 分の 1 まで減らすことで、ネットワーク、設置スペース、電力、冷却、運用のオーバーヘッドの運用コスト (TCO) が大幅に削減されます。

[1] 標準的な DRAM のコストは、ギガバイト単位で比較すると、ディスクのコストの 250 倍、フラッシュのコストの 10 倍です。

3　IBM ソリューション

IBM Data Engine for NoSQL は、プロセッサーにデバイスを接続するための高性能なソリューションを提供する POWER8 システム上の新しい Coherent Accelerator Processor Interface (CAPI) を土台に構築されています。本書のこのセクションでは、CAPI、フラッシュ、Redis ソフトウェアについて説明するとともに、標準的な NoSQL 実装時の拡張性の問題への対応に役立つ独自のソリューションがどのようにして共同で作成されたかを説明します。

3.1　Coherent Accelerator Processor Interface (CAPI) の概要

POWER8 のオープン・アーキテクチャーにおける重要なイノベーションは、Coherent Accelerator Processor Interface (CAPI) です。CAPI は、パートナー企業のデバイス、POWER8 コア、オープン・メモリー・アーキテクチャーの間に高帯域幅で低遅延のパスを提供します。CAPI アダプターは、通常の PCIe x16 スロットに取り付けられ、基本的に送信メカニズムとして PCIe Gen 3 を使用します。ただし、その他の入出力カードやアクセラレーターとの類似点は、この点のみです。CAPI 対応デバイスは、コア上で稼働するアプリケーション・プログラムに取って代わることも、カスタム・アクセラレーション実装を提供することもできます。CAPI は、入出力サブシステムのオーバーヘッドと複雑さを取り除き、アクセラレーターがアプリケーションの一部として稼働できるようにします。そのため、アプリケーションはカーネルを介することなく直接的にアクセラレーターと対話でき、コード・パスが減少します。IBM のソリューションは、プログラミングへの投資をはるかに低く抑えながらシステム・パフォーマンスを高めることができ、さらに幅広いアプリケーションでハイブリッド・コンピューティングの稼動を可能にします。

CAPI のパラダイムでは、アクセラレーションに固有のアルゴリズムは、FPGA の中の「Accelerator Function Unit (AFU またはアクセラレーター)」という 1 つのユニットに収容されています。AFU の目的は、アプリケーションのパフォーマンスを向上させてホスト・プロセッサーの負荷を軽減するために、アプリケーションのカスタマイズされた機能用の計算単位密度を高めることです。アプリケーションを加速するために AFU を使用すると、幅広いアプリケーションでコスト効率の良い処理が実現します。CAPI における重要なイノベーションは、クライアントの AFU を POWER8 プロセッサーと一貫性のある対等機能として扱うためのインフラストラクチャーを提供するカスタム・シリコンが、POWER8 システムに収容されていることです。

各 CAPI アクセラレーター・プロセッサーは、システム内で「ファースト・クラス」の存在であり、サーバー内の POWER8 プロセッサーが使用するのと同じメモリーやアドレス・スペースを処理します。IBM は、アクセラレーター・デバイスの管理を簡素化する耐久性の高いサービス層と抽象化層を各アクセラレーターに提供して、ソリューション設計者がアプリケーション固有の課題への取り組みにさらに集中できるようにします。例えば、アクセラレーターは、その他の POWER8 スレッドと同様にロックに参加でき、デバイスへの通信のオーバーヘッドを大幅に軽減します。さらに、簡素化されたアドレッシングにより、アクセラレーターの使用とプログラミングが簡単になります。CAPI に適したアプリケーションとしては、モンテカルロ・アルゴリズム、キーバリューストア、財務や医療のアルゴリズムが挙げられます。

IBM Solution for Flash Optimized NoSQL でも見られるように、CAPI は、フラッシュ・メモリー拡張の基盤としても使用できます。この革新的な製品は、CAPI 接続ソリューションによって 40 TB のデータへのシステム・アクセスを提供します。

CAPI の全般的なバリュー・プロポジションは、プロセッサーをハードウェア・アクセラレーターに接続して、同じ言語で通信できるようにする (IO ドライバーなどの仲介をなくす) ことで、新しいアルゴリズム実装の開発時間を大幅に短縮してアプリケーションのパフォーマンスを向上させることです。CAPI について詳しくは、CAPI に関するホワイト・ペーパー (http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=SP&infotype=PM&appname=STGE_PO_PO_JPJA&htmlfid=POS03140JPJA&attachment=POS03140JPJA.PDF#loaded) を参照してください。

3.2 Redis の概要

Redis は、BSD のライセンス提供を受けている、高度なオープン・ソースの**キーバリューのキャッシュ**と**ストア**です。Redis は、*単純な*キーバリューストアではなく、*データ構造サーバー*であり、さまざまな種類のアプリケーションをサポートしています。従来型のキーバリューストアでは、ユーザーは、ストリング・キーをストリングバリューに関連付けます。Redis では、単純なストリングに制限されず、さらに複雑なデータ構造を保持することもできます。次に、現在 Redis によってサポートされているすべてのデータ構造をリストします。

データ構造	説明
バイナリー・セーフのストリング	
リスト	挿入の順序に従ってソートされたストリング・エレメントの集合
セット	ソートされていない固有のストリング・エレメントの集合
ソートされたセット	すべてのストリング・エレメントが*スコア*という浮動小数点数値に関連付けられる
ハッシュ	値に関連付けられたフィールドで構成されるマップ。フィールドと値はストリング (Ruby または Python のハッシュに非常に似ている)。
ビット配列 (または単にビットマップ)	特殊なコマンドによって操作されるビットの配列として表現されるストリング値。ユーザーは、個々のビットを設定またはクリアしたり、1 に設定されたすべてのビットをカウントしたり、最初に設定されたビットや未設定のビットを見つけたりすることができる。
HyperLogLogs	セットの基数を見積もるために使用される確率的なデータ構造

表 1: Redis によってサポートされているデータ構造

3.3 Flash Optimized NoSQL ハードウェア

IBM Data Engine for NoSQL は、フラッシュ・メモリー・アレイへの高スループットで低遅延の接続を提供する能力を活用して、NoSQL 実装のスケーリングの問題に対応する独自のメモリー層を作成します。この設計では、メインメモリーを使用して「ホット」データをキャッシュに入れたり保持したりすることで、プロセッサーのメインメモリーは、アプリケーションで必要となる高速な応答時間を提供できます。ただし、このソリューションは、IBM Flash Systems ストレージ・ソリューションを使用して、最大 40 テラバイトのフラッシュ・メモリーへのアプリケーション・アクセスを提供します。フラッシュ・アレイは、CAPI アダプター・カードを使用して取り付けられ、プロセッサーとフラッシュの間に高帯域幅で低遅延のパスを提供します。そのために、アダプターは、Field-Programmable Gate Array (FPGA) チップとファイバー・チャネル入出力ポートを使用します。FPGA デバイスは、接続されている Flash Systems ストレージ・アレイの管理とアクセス制御のための専用ロジックを収容しています。このロジックは、FPGA に内蔵しているため、必要に応じて更新できます。POWER8 プロセッサーに DRAM とフラッシュの両方への直接アクセスを提供することで、アプリケーション・ソフトウェアは、メモリーとフラッシュの使用量の比率を調整して、特定のサービス・レベル・アグリーメントに基づいてパフォーマンスとコストを最適化することができます。

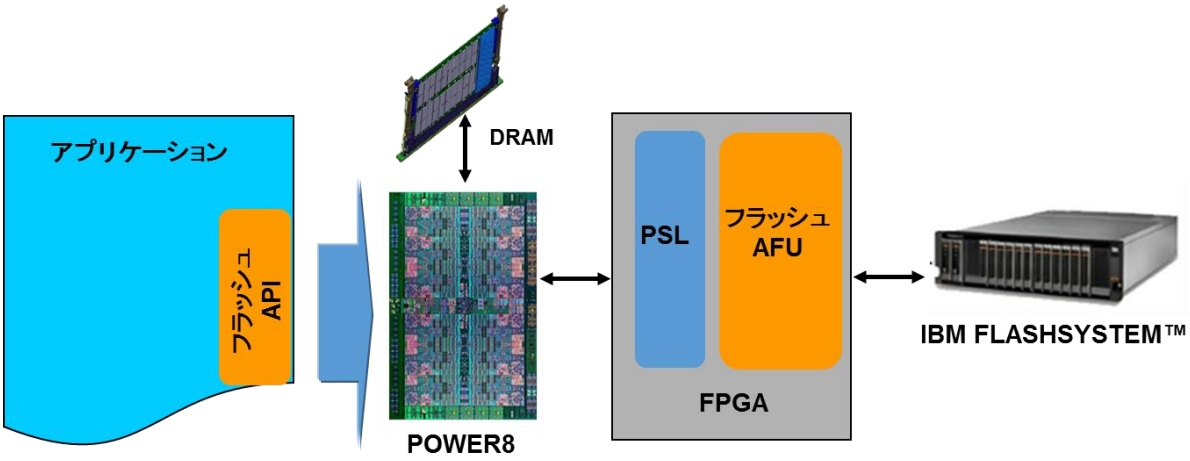

図 1: IBM Data Engine for NoSQL の概念図

3.3.1 IBM Data Engine for NoSQL ソフトウェア

IBM Data Engine for NoSQL ソリューションは、フラッシュ内のデータの管理とアクセスのためのキーバリューとロー・ブロック入出力のインターフェースを提供する一連の開発 API を通じて、フラッシュへの直接アクセスをアプリケーションに提供します。ソフトウェア・パッケージ全体は、次の 4 つのコンポーネントで構成されています。

マスター・コンテキスト (MC) - クライアント・アプリケーション・ソフトウェアの代わりに、アダプターの初期化、LUN のディスカバリーとマッピング、エラー・リカバリーとヘルス・チェックを実行して、訂正不能エラーに対応し、リンク・イベントを管理します。

ブロック入出力 API – 特定のブロックの読み取り/書き込み要求を処理して、フラッシュ内の論理アドレスにあるデータの読み取り/書き込みのために AFU に対してコマンドを直接発行します。さらに、それらの要求への応答も処理します。

キーバリューストレージ (KV) API - Redis と上記のブロック入出力 API の間のブリッジを形成する汎用キー値データベースを提供します。

次の図に、このソリューションのソフトウェア・コンポーネントを示します。

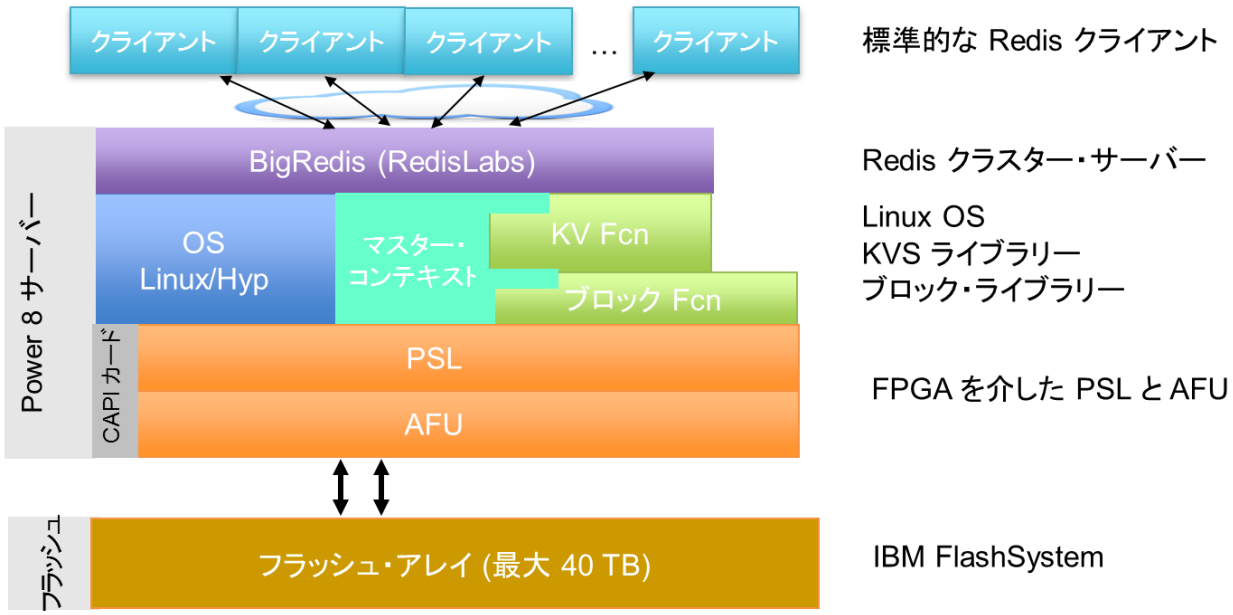

図 2: ソフトウェア・コンポーネントの概要

3.4 BigRedis の概要

BigRedis は、Redis Labs がこのソリューションに付けた名称です。BigRedis は、CAPI インターフェースが提供する大規模なフラッシュ・アレイを活用するために拡張された、Redis Labs Enterprise Cluster の 1 バージョンです。大容量フラッシュだけでなく、Power S822L システムの機能も活用して、最大 192 の実行スレッドを実現するため、事実上、このソリューションは「ボックスに収まったクラスター」の実装となっています。そのため、Redis サーバーは、単一のノードで使用できる大容量フラッシュ・アレイと多数の実行スレッドの両方を使用して拡張できます。BigRedis では、ユーザーは、必要なストアのサイズを選べるだけでなく、個々のニーズに合っていると判断する価格対性能比やサービス・レベル・アグリーメントも選ぶことができます。ソリューション・コストの予測を踏まえて、次の図に、ユーザーによるメモリーとフラッシュの比率の変更に伴う代表的なパフォーマンスとコストの相対的な関係を示します。

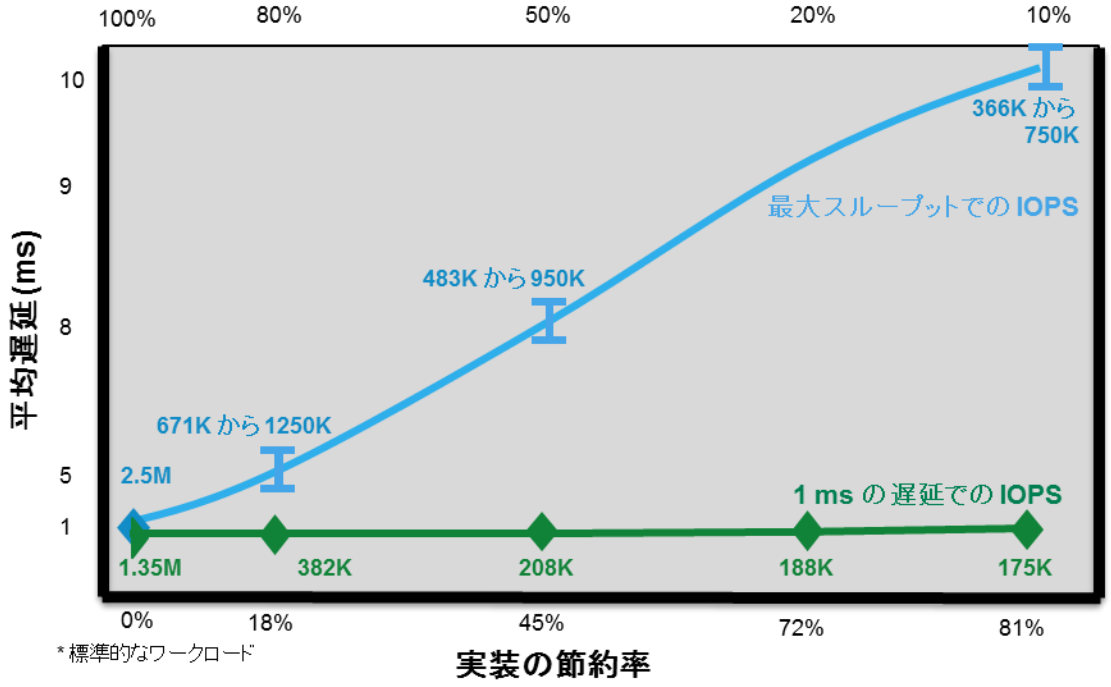

図3: ユーザーは、IOPS 率でも遅延の要件でも、ソリューションのニーズに合った費用対性能比を選択できます。Redis インスタンスのフラッシュの相対量が増えると、RAM に対するフラッシュの相対コストにより、システム実装にかかるコストの節約につながります。CAPI によって加速されるフラッシュを使用すると、ユーザーは、特定のアプリケーションに望ましい遅延とコストを選択することができます。この柔軟性は、すべて RAM で構成された Redis では得られません。

CAPI フラッシュを備えた BigRedis の概要のビデオは、https://www.youtube.com/watch?v=wXO9_Hp3p60 で見ることができます。

4　対象となる市場セグメント

IBM Data Engine for NoSQL で構築されたソリューションのターゲットは、現在、NoSQL ソリューションを使用しているか、NoSQL ソリューションに投資している多様な業界と研究分野に属しています。現在、Redis Labs が提供する BigRedis は、入手可能な唯一の NoSQL MSP ソリューションですが、将来、その他の MSP ソリューションが発表されるでしょう。最適化された NoSQL データベースの市場は広大です。次の図 4 に、NoSQL ベースのソリューションを使用する業界と対象市場をいくつか示します。これらの市場は、CAPI Flash Optimized NoSQL が適用される可能性がある市場全体のほんの一部です。

図 4 POWER8 と CAPI アクセラレーションを活用するソリューションの市場

5　成長と採用の戦略

IBM Data Engine for NoSQL は、最初は NoSQL 実装のスケーリングの課題を解決することを目標としていますが、IBM は、この新しいメモリー層とその特性を活用できる可能性があるその他のアプリケーション・タイプを評価しています。多様な NoSQL ソリューションが市場で提供されているなか、 Redis オファリングは、幅広い NoSQL オファリングの費用効率特性を改善するスターティングポイントとして考えられます。また、標準的な IO 接続よりも大幅に短い遅延での大容量ストレージへのアクセスを必要とするその他のアプリケーションに独自のソリューションも提供します。インメモリー・データベースとその他のビッグデータ・アナリティクスのソリューションも、このソリューションの潜在的な対象アプリケーションです。

5.1　OpenPOWER Foundation の活用

POWER8 の CAPI インターフェースは、OpenPOWER アーキテクチャーの重要部分であり、パートナーがサーバー・ソリューション・アーキテクチャー全体でイノベーションを起こせるようにしています。今日、OpenPOWER Foundation の複数メンバーが新しい CAPI フラッシュ・ソリューションを開発しています。このようなアーリーアダプターは、前述した多くの対象市場においてソリューションを提供するでしょう。OpenPOWER Foundation のメンバーが将来提供するその他の NoSQL ソリューションに期待していてください。

5.2　製品オファリングと将来の方向性

IBM Data Engine for NoSQL は、開発中であり、開発の初期実動段階にあります。IBM の初期製品オファリングは、Redis Labs の製品である BigRedis で使用されています。IBM は、今後もその他の NoSQL プロバイダーと協力して、将来、新しい API をさらに幅広い開発で使用できるようにする予定です。Redis オファリングについて詳しくは、次の Web サイトを参照してください。 https://**redislabs**.com

6　まとめ

IBM Data Engine for NoSQL - Power Systems Edition は、ミドルウェアとアプリケーションの開発者が新しいストレージ層にアクセスできるようにするイノベーションです。このストレージ層は、POWER8 プロセッサーのオープンな CAPI バスを介して接続されるフラッシュ・テクノロジーに基づいており、事実上、フラッシュをプロセッサーに「より近い場所」に配置します。このハードウェアとソフトウェアのソリューションは、スケールアウト型の DRAM ベース・ソリューションに代わる、コストとパフォーマンスを重視したソリューションとなります。

IBM®

Printed in JapanSeptember 2015

本書の情報は、予告なしに変更される場合があります。本書に記載された製品は、移植、生命維持などのアプリケーションや、その他当該製品の不具合が生命の危険や身体傷害または大きな物的損害を招く可能性のある危険な用途での使用を想定していません。本書に記載される情報が、IBM 製品仕様または保証に影響を与える、またはこれらを変更することはありません。 本書の内容は、IBM またはサード・パーティーの知的所有権のもとで明示または黙示のライセンスまたは損害補償として機能するものではありません。 本書に記載されている情報はすべて特定の環境で得られたものであり、例として提示されるものです。他の操作環境で得られた結果は、異なる可能性があります。

本書に記載の内容は正確を期していますが、これらの情報は準備段階のものであり、その正確性や完全性について表明も保証もするものではありません。

注: 本書には、開発時の設計、サンプリング、または初期製造段階における製品の情報が含まれています。この情報は予告なしに変更される場合があります。最終設計を確定するにあたっては、IBM のフィールド・アプリケーション・エンジニアにご確認の上、本書の最新バージョンをご利用ください。

本書は、Power Architecture® と互換性のあるテクノロジー製品の開発のみを目的として使用できます。本書を変更または配布してはなりません。本書により、明示、黙示、禁反言その他によっても、いかなる知的所有権についてのライセンスも付与されることはありません。

本書に記載された情報は、現状のまま提供されます。IBM は、いかなる場合も、本書に記載された情報の使用に直接または間接的に起因して発生した損害について責任を負いません。

日本アイ・ビー・エム株式会社
〒103-8510 東京都中央区
日本橋箱崎町 19-21

IBM ホーム・ページ: **ibm.com®**

2015 年 4 月 8 日

POW03130-JPJA-01