

ビッグデータ時代における IT アーキテクチャ

丸山 宏

情報・システム研究機構 統計数理研究所

1. はじめに

IT を利用した産業構造のイノベーションを促進するため、経済産業省は Cyber-Physical Systems(CPS) [1]の高度利用を目指した「IT 融合政策」を発表している[2]。その議論の中で出てきたのが、「CPS に適したアーキテクチャは何か?」という問いであった。CPS ではセンサーから大量のデータが生成される。従ってこの問いは「ビッグデータ時代の IT アーキテクチャは何か」という問いでもある。

IT システムはその設計・運用に多くのステークホルダが関わる複雑なシステムであり、ここでは「システムの構造・働きに関する共通理解」としてのアーキテクチャの概念が極めて重要である。また、よいアーキテクチャを持つことは産業競争力の上でも重要である。

にも関わらず、IT 業界で主流になったアーキテクチャのほとんどは、米国発である。今までの我が国における研究開発の主眼は差別的な要素技術の開発にあり、アーキテクチャの研究開発投資が軽視されていたきらいがある。

特に、CPS やビッグデータなどの流れから、新たな IT アーキテクチャが生まれてくることは必至であり、情報・通信分野の研究開発もそのような流れを無視してはならない。

2. なぜアーキテクチャが変わるのか

そもそも、アーキテクチャはなぜ時代と共に変わってくるのだろうか。その主な理由として「技術の進歩」と「ワークロードの多様化」がある。半導体技術の進歩は高密度実装を可能にし、高並列の計算機を生み出した。一方、Web 検索などの新しいタイプのワークロードは、Hadoop に代表されるような新しいアーキテクチャを必要としている。

では、CPS あるいはビッグデータに特徴的なワークロードとは何だろうか。リアルタイムで発生するセンサーデータを処理するストリーム・コンピューティング（あるいは Complex Event Processing - CEP）や、MATLAB/Simulink に見られるような、連続的な物理世界を微分方程式で記述するプログラミングモデルは、CPS のワークロードに適したアーキテクチャと言うべきであろう。

3. ビッグデータとは

ビッグデータが喧伝されているが、ビッグデータは、「感や経験に基づく専門技能がデータ分析に置き換わりつつある」という流れの一部であるにすぎない。この「データに基づく意思決定」こそが大きな流れであることを認識すべきである[3]。このため、経営におけるデータ分析のリテラシーが求められている。単に IT に投資するだけではだめで、データの分量の大小に関わらず、データ分析を経営におけるプロセスとして定義し、それらを利用する仕組みを構築しなければならない。その中で、本当に必要ならばいわゆる「ビッグデータ」の処理を行うことを、投資対効果を考えて行うべきである。

4 Edge-Heavy Data

CPS が普及すると、センサで発生する大量のデータを扱う必要がある。その際に必要とされる、新たなアーキテクチャはどんなものだろうか。Preferred Infrastructure の岡野原大輔氏は、データの大部分がネットワークのエッジ部分で格納・処理される「エッジ・ヘビー・データ」の時代が来ると予測している[5]。

1970 年に Codd は、データこそ価値のあるものであるから、データはアプリケーションから独立させなければならないと唱えた[6]。この結果、今ではデータはデータセンタで集中的に管理されるのが普通である。しかし CPS の時代では、データの価値はますます増大するものの、全データをセンタに送ることは現実的ではない。理由は 2 つある。

一つは、センサ、特に画像センサからの生データはあまりにも容量が大きく、それらを中央のデータセンタにそのまま送るには通信路やストレージのコストが高すぎるということである。送ることが必要であったとしても、何らかの処理をして情報を縮退させざるを得ない。

もう一つの理由は、センサからのデータには、必ずしも利用目的が予め決められていないものがあるということだ。McKinsey のビッグデータに関するレポート[7]によれば、ビッグデータのほとんどは“exhaust data”（排気データ）、つまり、意識して集めたのではなく、他の事業活動の副産物として生成されたデータであるということである。将来使うかわからないデータを、高価な通信路やストレージに負荷をかけてセンターに送るのは無駄である。

上記 2 つの理由により、我々は、今後はデータの多くがセンサあるいはセンサに近い場所（ネットワークのエッジ）で管理されるようになることを予想する。

既にエッジにデータを溜めている例として、監視カメラがある。現在、監視カメラのデータはサーバには送られず、ローカルに保存されている。昨年の監視カメラの市場規模は、世界で 750 万台といわれており、100GB/カメラとすると、750PB のデータをローカルに保存していることになる。また、現在データを溜めているとは言い切れないが、溜めることができるものとしてスマートフォンが挙げられる。昨年、国内で 2000 万台売れたということは、少なめに見積もって 10GB/台としても 200PB のデータをローカルに保存できる状態にある。McKinsey のレポートでは、日本は年間 400PB のデータを生成していると記載されている。つまり、このうちの半分はスマートフォンにあってもおかしくはない。

Edge-Heavy Data は既に起こりはじめているのだ。

では、Edge-Heavy Data に特化したアーキテクチャとは、どのようなものだろうか。これから多くのアイデアが現れてくるだろうが、ここでは下記の 5 点を挙げておく。

1. SSD・消費電力・ストリーミングデータなどを考慮したハードウェア
2. 画像・映像処理を組み込んだハードウェアとミドルウェア
3. データは動かさず、計算を動かすプログラミングモデル（GFS + MapReduce はその萌芽ではあるが、さらに進化するだろう）
4. 欠測値・外れ値・分布表現を取り込んだプログラミングモデル
5. 分散データマイニングを可能にするミドルウェア。例として、Preferred Infrastructure と NTT が共同開発しているオープンソース・ソフトウェア Jubatus がある[6]

これらは、ハードウェアとソフトウェア（プログラミングモデル）を統合したアーキテクチャになるだろう。

5 おわりに

ICT におけるアーキテクチャの流れは、ほとんど常に米国を発信源としていた。どのアーキテクチャが商業的に主流になるかについては、技術的な観点のみならず、経済や産業の動向、政治的な動き、特定の企業の思惑など多くの要素があり、正確に見極めることは難しいだろう。しかし、次に何が現れてもすぐに対応できるように準備を整えておくことが、ICT 産業に携わるものにとって、重要であると考え。我々は、今後現れてくるワークロードを考え、そこで必要なアーキテクチャを考えて、1 歩先を読んで準備しておくべきであろう。

参考文献

1. 丸山 宏, 佐々木 康裕, IT 融合社会 -情報技術の新たな地平線, 情報処理, Vol.54 No.2, 2013.
2. 経済産業省における IT 融合政策,
<http://www.meti.go.jp/press/2012/05/20120501002/20120501002-4.pdf>.
3. Ian Ayres, *Super Crunchers: Why Thinking-by-Numbers Is the New Way to Be Smart*, ISBN-13: 978-0553805406, 邦訳その数学が戦略を決める, 文春文庫, ISBN-13: 978-4167651701, 2010.
4. 樋口知之, ビッグ・データを操る者が勝つ, DIAMOND ハーバード・ビジネス・レビュー, 2012 年 2 月号, <http://www.ism.ac.jp/news/2012/120116.html>.
5. 丸山宏, 岡野原大輔, “Edge-Heavy Data: CPS・ビッグデータ・クラウド・スマホがもたらす次世代アーキテクチャ,” GICTF 総会 特別講演, <http://www.gictf.jp/doc/20120709GICTF.pdf>, 2012.
6. 18. E. F. Codd, "A Relational Model of Data for Large Shared Data Banks," Comm. ACM, Vol. 13, No. 6, 1970.
7. McKinsey Global Institute, Big data: The next frontier for innovation, competition, and productivity,
http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation, 2011.
8. Jubatus : Distributed Online Machine Learning Framework, <http://jubat.us/>

Contact

質問、ご意見、反論など、何でも結構です。気軽にメールください。時間の許す限り、対応します。メール：maruyama@acm.org, Twitter: <http://twitter.com/maruyama>