

デジタルデータの長期保存における課題と戦略

Issues and Strategy of Long-term Digital Data Preservation

石原 直

Abstract

高度情報化デジタル社会に突入した今、社会の多くの分野において人類の貴重な情報消失の危機が叫ばれている。過去の幾つかの情報消失事例への反省から、デジタル情報の長期保存システムの構築に向けて様々な取組みが行われているものの、いまだ抜本的な解決に至っていない。本稿では、超長期にわたって記憶データを保持できるメモリ媒体の実現を前提として、デジタルデータ長期保存方法として現在のところ唯一の方法であるマイグレーション戦略から、低価格・高信頼なエミュレーション戦略への転換を提案している。

キーワード：デジタル情報，長期保存，半導体技術，メモリ素子

1. はじめに

今日、情報技術あるいはネットワークの基盤が次々と整備され、デジタル技術に支えられた高度情報化社会、高度ユビキタスネットワーク社会が実現されている。ネットワークが発展し、情報流通のコスト、料金が大幅に下がり、スマートホン、携帯電話、パソコンを自由に使って仕事あるいは遊びをする時代になった。ここに至る発展の過程では、通信インフラの整備、情報技術の進歩が大きな役割を果たしてきたが、中でもこの半世紀の間の半導体の技術進歩は強力な原動力であった。

半導体技術の発展の背景を図1に示す。半導体産業は、「半導体集積回路（LSI）の集積度や性能は3年で4倍になる」というムーアの法則⁽¹⁾でけん引され、実際に半導体市場はムーアの法則が提示された1965年から半世紀にわたって常に一本調子で発展してきた。これを技術として支えたのが、CMOSトランジスタのスケールリング則⁽²⁾である。素子の寸法を K 分の1にすると、集積度は K^2 倍になり、トランジスタの性能をスピードと消費電力で評価した性能が K^2 倍になるという法則で、素子の基本構造は変えずに寸法を縮めていくだけで性能が着実に向上することに技術的な裏付けを与えている。実際、トランジスタの基本構造を変えることなく、この

50年間で半導体素子の集積度は100万倍、あるいはインテルのホームページによると、実は1,000万倍に、単体の素子性能は1万倍になったとされる。それにもかかわらず、半導体チップの価格はほぼ一定で推移してきたことから、データ処理のコストは、100億分の1から1,000億分の1に削減されたことになる。このことが、半導体技術が高度情報化社会発展の強力な駆動力であったことを如実に示しているのである。

2. デジタル情報の爆発

このような高度情報化社会の進展に伴って世界で増え続けるデジタル情報の生成状況を図2に示す。米国 International Data Corporation の公開資料⁽³⁾によると、2011年に全世界で生成されたデジタルデータ量は1.8 ZByte、2020年には40 ZByteに到達すると予測され、既に2007年には生成されるデータ量が供給されるメモリ・ストレージの容量を上回ってしまったとされる。身近な例でも、世界中で発表される論文が2000年以降に急激に増えていることはよく知られている。一般に、このような指数関数的な情報量の増大を「情報爆発」と呼んでいるが、これに伴って、「人類が保存すべき情報量」も急速に増大し、比例して「貴重な情報の消失のリスク」が増大しているのである。

3. これまでのメモリ技術開発

高度情報化社会の持続的な発展のためには、基本的に

石原 直 正員 東京大学大学院工学系研究科
E-mail ishihara.sunao@mail.u-tokyo.ac.jp
Sunao ISHIHARA, Member (Graduate School of Engineering, The University of Tokyo, Tokyo, 113-8656 Japan).
電子情報通信学会誌 Vol.98 No.12 pp.1038-1042 2015年12月
©電子情報通信学会 2015



図1 半導体の発展 半導体素子はムーアの法則によって集積度や素子性能を大幅にアップさせながらチップ価格はほぼ一定に抑えられた。これによりデジタルデータ処理においては、 $1/10^{10} \sim 1/10^{11}$ という驚異的なコストダウンが図られた。

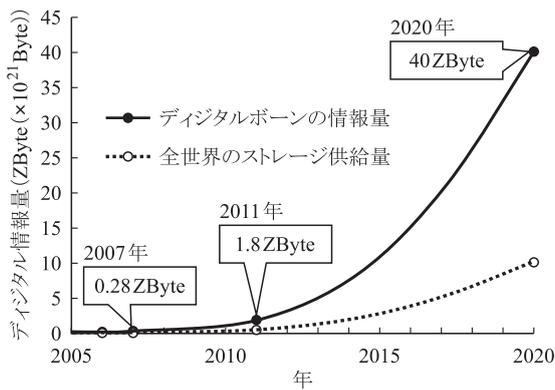


図2 情報爆発 21世紀に入ってからデジタルデータの急速な伸びは、その増大する驚異的なスピードから「情報爆発」と呼ばれる。

2種類のデジタル技術が必要と考えられる。一つはデジタルデータを処理する技術体系で、一般に情報技術 (IT) と呼ばれる。社会活動、経済活動、産業活動を支えるためのデータ処理のテクノロジー、テクニック、及びそのシステムである。社会が必要とするもう一つの技術は、デジタルデータを保存する技術体系である。社会、産業、個々人のアイデンティティ、文化、社会的規範、歴史を継承していくための情報の記憶・保存の技術である。

これまでの処理系デジタル技術におけるメモリ技術開発の状況を図3に示す。情報システムに組み込まれるメモリとして、ワークメモリは年に約10%、ストレージメモリは年に約3倍と急速に発展してきた。しかしながら、ここで使われているメモリの寿命は、当然のことながら情報システムの寿命と同程度で十分あり、100年

あるいは1,000年といった長期記憶寿命は全く考慮されていない。振り返ってみると、デジタル社会の本質的課題であるデジタル情報の長期保存の必要性について社会的に認識されてこなかったことから、保存系のメモリの技術開発は置き去りにされてきたものと考えられる。このデジタルデータ保存に無関心なまま情報爆発が進行すると、人類の貴重な情報の回復不可能な消失につながる危険性は確実に増大していくものと考えられる。

4. これまでの長期保存への取組み

デジタル情報の長期保存の問題は1990年代中頃から米国を中心に関心が高まり、2000年には米国の議会図書館が National Digital Information Infrastructure and Preservation Program (NDIIPP) を設立し、当初から180以上の大学、研究機関が参加した⁽⁴⁾。

2006年にはスウェーデン国立公文書館から電子社会におけるデータ消失の危機に警鐘を鳴らすメッセージが出された⁽⁵⁾。デジタルデータを長期保存するためのプロジェクト推進にとって、初期投資はもちろん、その後のメンテナンスの経費や継続的なりニューアル経費などの資金の手当てを綿密に計画しておかないとデジタル化プロジェクトは破綻するという警告であった。

また、デジタルデータに特有の保存上の課題を明らかにした事例として、2000年にNASAで起きた不測の事態が知られている。1975年にNASAは火星探査機バイキングを火星に送ったが、25年後にバイキングの観測データを読もうとしたところ復元できないという事態に陥った。磁気テープに保存されたビットデータは読み

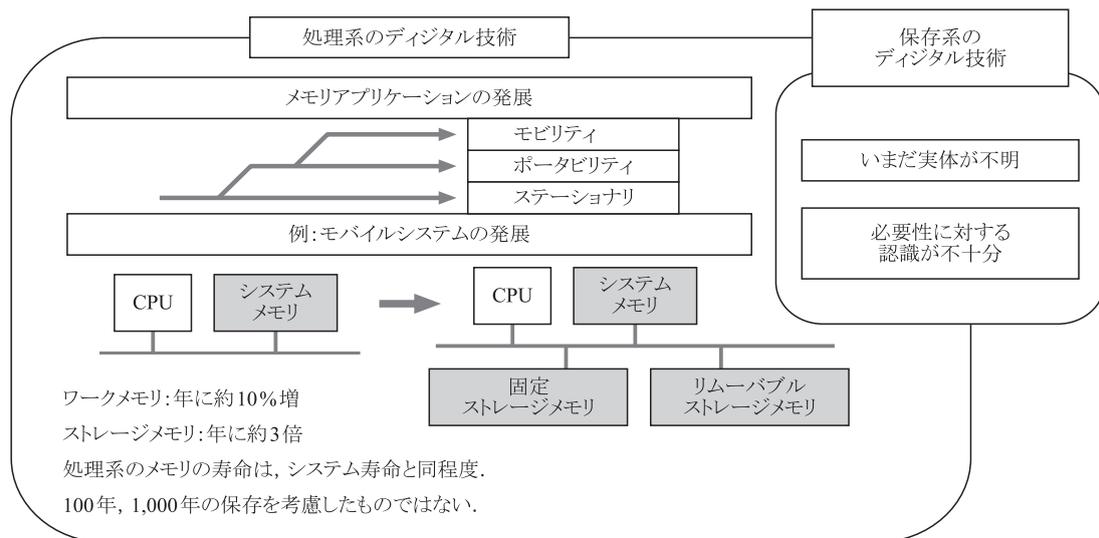


図3 メモリ技術の開発 デジタルメモリは処理系の技術として発展してきたが、これまで、100年を超えるような長期保存用のメモリ技術開発の必要性は認識されてこなかった。

出せたものの、そのフォーマット情報が失われてしまったため、ビットストリームデータの意味が分からなかったというものである。デジタルデータの意味解読には、データの構造を記述するためのデータが必須であることを改めて思い知らされることとなった。この出来事を受けて米国においては、2002年に宇宙データシステム諮問委員会(Consultative Committee for Space Data Systems)がOpen Archival Information System(OAIS)というデジタル情報を保存し利活用するための参照モデル(詳細は後述)を策定し、ISOに登録した⁽⁶⁾。このモデルあるいは類似のモデルを基にデジタルデータを長期に保存し、利活用するための研究が広く行われている。

ヨーロッパにおいては、EUが多年次にわたる助成プログラム、Framework Programを1984年から実施しており、その6期(2002~2006)においてCASPAR(Cultural, Artistic and Scientific Knowledge for Preservation, Access and Retrieval)プロジェクト⁽⁷⁾、7期(2007~2011)にはENSURE(Enabling knowledge, Sustainability, Usability and Recovery for Economic Value)プロジェクト⁽⁸⁾が実施され、OAIS参照モデルに基づくデジタルデータ保存のための解決策について研究されてきた。

SNIA(Storage Networking Industry Association、米国のストレージベンダを中心とした業界団体)は、2009年に276の企業、団体に調査を行い、そのうちの68%が100年以上のデータの保存を望んでいると報告している⁽⁹⁾。

このように、一般に、デジタルデータの長期保存技術に対する認識が不十分なまま、デジタル技術が社会

に浸透する中、デジタルデータの長期保存に関する取り組みは着実に増えているのである。米国においては、博物館、図書館、公文書館、警察・裁判所、病院などにおいて、デジタルデータの長期保存の必要性が叫ばれ、対応が活発化している。例えば、SOX法(Sarbanes-Oxley Act)、FRCP(Federal Rules of Civil Procedure、アメリカ連邦民事訴訟規則)、HIPAA法(Health Insurance Portability and Accountability Act、医療保険の相互運用性と説明責任に関する法律)などデジタルデータ長期保存に備えて社会的法整備にも取り組んでいる。

更に、映画産業においても映像のデジタル化の進展に伴ってデジタルデータ長期保存が必須となっている。しかしながら現実には、デジタルデータを長期間に安定かつ安価に保存する手段を見つけ出すことができないとの認識から、ハリウッドの映画業界においては、いまだにフィルムによる長期保存が行われている⁽¹⁰⁾。

5. デジタルデータの超長期保存の方法

5.1 デジタルデータ保存の現状

歴史上のこれまでのアナログ記録媒体はそれなりの記憶寿命を持っていた。ロゼッタストーンの2,000年を超える記録寿命は象徴的な事例であるが、現実の記憶媒体でも、紙は記録の保証期間が250年、期待寿命は700年といわれ、マイクロフィルムは保証期間が500年のものもある。これに対してデジタルデータの記憶・保存における課題は、メモリ・ストレージの記憶寿命が短いことである。実際上、デジタルメモリ・ストレージは情報システムの一部として使われるため、期待されるメモリ寿命はシステム寿命と同程度で問題ないことになる。

表1に示すとおり、実際の情報システムの寿命が10年程度と短いことから、デジタルメモリの期待寿命も同様に数十年程度と極めて短い状況なのである。

5.2 メタデータの重要性

媒体、システム寿命の課題のほかに、デジタルデータにおいて特に重要なのは意味理解の問題である。デジタルデータ本体はビットストリームなので、これだけでは何の意味も理解できない。そこでメタデータという、データの持つ意味を理解するために必要なデータを付け加えることになる。更にメタデータのメタデータが必要となる場合もある。どのようなメタデータをどのように付与すべきかについて、図4に示す「参照モデル」と呼ばれるメタデータ構造の標準が提案されている。このモデルは、前述のNASAバイキングのデータ消失を受けて作られ、標準化されたOAISの中に概念として定義されている。図から分かるように、記憶すべきデータの本体である内容情報(Content Information)と、その内容に対するメタデータである保存記述情報(PDI: Preservation Description Information)と、この二つを結び付けるパッケージ化情報(Packaging Information)をまとめて一つのパッケージとし、更にパッケージの外側にパッケージの内容の特性を記述している。

表1 保存システムの寿命(ハード, ソフト)

寿命	ハードウェア	ソフトウェア
3~5年	ホストコンピュータ	・アプリケーションソフト ・オペレーティングシステム ・デバイスドライバ
5~10年	物理 インタフェース	・インタフェース ファームウェア
3~5年	メディアドライブ	・ドライブコントロール ファームウェア
5~10年	メディア	・ファイルシステム ・データファイルフォーマット ・物理レコードフォーマット

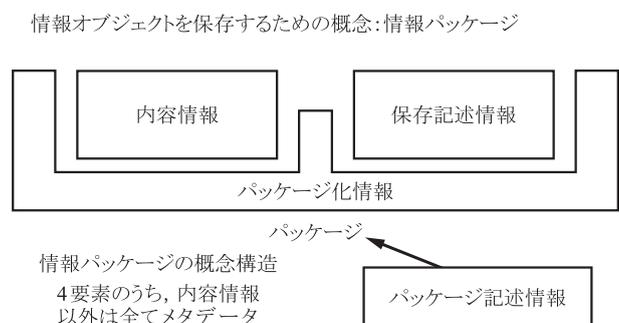


図4 Open Archival Information System (OAIS) デジタルデータの意味理解のために必要なデータであるメタデータの構造を規定する参照モデルの例。

5.3 デジタルデータの長期保存戦略

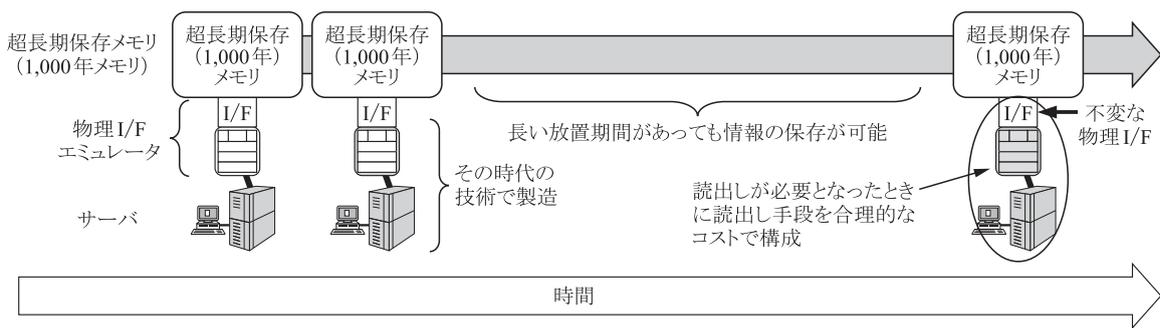
デジタルデータを継続的に保存するための方法を保存戦略と呼び、現状では三つの戦略があるとされている。まず、「リフレッシュ戦略」は、単なるコピーの繰返しである。次に、「エミュレーション戦略」と「マイグレーション戦略」がある。長期保存システムとして現状ではマイグレーション戦略が広く使われており、唯一の長期保存方法となっている。ただ、コンピュータとソフトウェアの世代交替に伴ってデジタルデータを新しいシステムへ移し変えていくという方法であるため大きなコスト負担が生じている。もう一つの方法であるエミュレーション戦略は、新システム上に旧システム上にあったメモリ・ストレージ媒体の読み出し環境を作るという方法である。したがってエミュレーションが成り立つためにはオリジナル媒体に記録が保持されていることが必須条件となる。現状では超長期間にわたって記録を保持できる媒体がないため長期保存戦略として採用されていない。このことは、超長期間の記憶を保持する媒体さえ実現できれば、エミュレーション戦略がデジタルデータ長期保存戦略として有力な方法になり得ることを示唆している。

5.4 課題解決策の一提案

現状のデジタルデータ長期保存には前述のようにマイグレーション戦略が使われている。ただ、現状のマイグレーション法は、コンピュータの世代交替ごとの新システムへの移行コストの負担が必要となる、更に、一度のデータ移行の途切れが回復不可能なデータ消失となってしまうというせい弱さも抱えており、持続的な使用には疑問が多い。そこで、これに代わる経済的で堅ろうな方法として、図5に示すエミュレーション戦略の構築を提案したい。もちろん本提案では既述のように超長期にわたって記憶を保持できるメモリ媒体を開発できることが前提となる。開発するメモリ素子には、100年以上、あわよくば1,000年という超長期の記憶保持性能が求められる。更にデータ保存コストの観点からは、媒体として堅ろうである、保存環境に左右されない、メンテナンスフリーなどの特性が望まれる。また、保存システムとしてはデータ読み出しの技術的手段を合理的なコストで実現できることが重要であり、読み出し用の物理インタフェースを低価格で実現したい。また同時に、意味理解の手段を備えることも必須なので、これを保証するためにデータフォーマットやメタデータの標準化も必要となる。

以上、超長期メモリ開発、及びハード的にもソフト的にも妥当なコストでエミュレータを構築できることが長期保存システム実現のキーである。

・エミュレーション戦略(超長期保存メモリ媒体の存在が前提)



・マイグレーション戦略(現状)

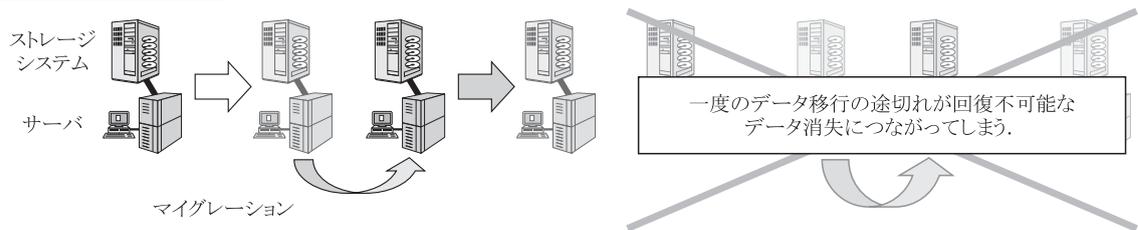


図5 期待される超長期保存システムはマイグレーションかエミュレーションか？ 堅ろうさ、信頼性、コスト等を比較すると、超長期保存メモリ媒体さえ確保できればエミュレーション戦略が有力と考えられる。

6. ま と め

今後、100年以上、望むらくは1,000年にわたって妥当なコストで意味理解を保証する記憶システムの構築は必須である。現在のデータ長期保存で使われているマイグレーション戦略はデータ量とコストの増大に対応できない恐れがあり、新たな保存戦略が必要となる。そこで、超長期保存メモリ媒体を実現できれば、より低コストで堅ろうなエミュレーション戦略が有効な方法となる。またエミュレーション戦略導入のためには、意味理解を保証するハードとソフトのエミュレータがいつでも安価に構築できること、データ再構築に必要なメタデータとメタデータスキーマの構築・標準化が重要となる。

文 献

- (1) G.E. Moore, "Cramming more components onto integrated circuits," Electronics Magazine, vol. 38, no. 8, April 1965.
- (2) R.H. Dennard, F.H. Gaensslen, H.-N. Yu, V.L. Rideout, E. Bassous, and A.R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," IEEE J. Solid-State Circuits, vol. sc-9, no. 5, pp. 256-268, 1974.
- (3) J. Gantz and D. Reinsel, The Digital Universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east, Dec. 2012, sponsored by EMC, <http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf>

- (4) アメリカ議会図書館, Library of Congress, <http://www.loc.gov/index.html>, http://www.digitalpreservation.gov/news/2011/20110308_news_ndiipp_report.html
- (5) J. Palm, The Digital Black Hole, http://www.tape-online.net/docs/Palm_Black_Hole.pdf
- (6) Reference Model for an Open Archival Information System (OAIS), <http://public.ccsds.org/publications/archive/650x0m2.pdf>
- (7) Cultural, Artistic and Scientific knowledge Preservation, for Access and Retrieval, http://cordis.europa.eu/project/rcn/92920_en.html
- (8) Enabling Knowledge, Sustainability, Usability and Recovery for Economic Value, http://cordis.europa.eu/project/rcn/98002_en.html, <http://www.prelida.eu/ensure-enabling-knowledge-sustainability-usability-and-recovery-economic-value>
- (9) SNI, Long-term Preservation of Digital Information, http://www.snia.org/sites/default/files/MaryBaker-RogerCummingsLong-term_Preservation_Digital%20Information.pdf
- (10) The Digital Dilemma, The Academy's Science and Technology Council, <http://www.oscars.org/science-technology/sci-tech-projects/digital-dilemma>

(平成 27 年 6 月 25 日 受付)



いしはら すなお
石原 直 (正員)

昭 48 東大大学院工学系研究科精密機械工学専攻修士了。平 4 工博 (東大), 昭 48 日本電信電話公社 (現 NTT) 電気通信研究所入社, X線リソグラフィの研究, 物性科学基礎研究の推進。平 17 東大・工・機械・教授, ナノメカニクスの研究。平 25 東大大学院工学系研究科 上席研究員 (現職)。