インターネットの発展に数理科学が果たした役割

2006年12月20日

株式会社インターネット総合研究所代表取締役所長
東京大学大学院数理科学研究科客員教授
（情報理論901-35担当）
藤原　洋

目　　次

1. 産業革命と数理科学

産業革命において数理科学が果たす役割

IRI Ubilabs Innovations

インターネットによる情報革命は五大産業革命の一つである！

（ブライアン・アーサー：サンタフェ研究所【複雑系研究】、複雑系経済学）

- 1780～1830： イギリス　紡績機械（水力）
- 1830～1880： イギリス　鉄道（蒸気機関）

原理：力学
⇒動力機関

- 19世紀末　：　ドイツ　重工業（電動機、鉄鋼）
- 1913～1970代： アメリカ　T型フォード（1913）からの製造業革命
  ⇒大量生産、自動車産業、石油の時代

原理：物質科学⇒重化学工業

- 20世紀末～： アメリカ、（日本？） ディジタル情報革命

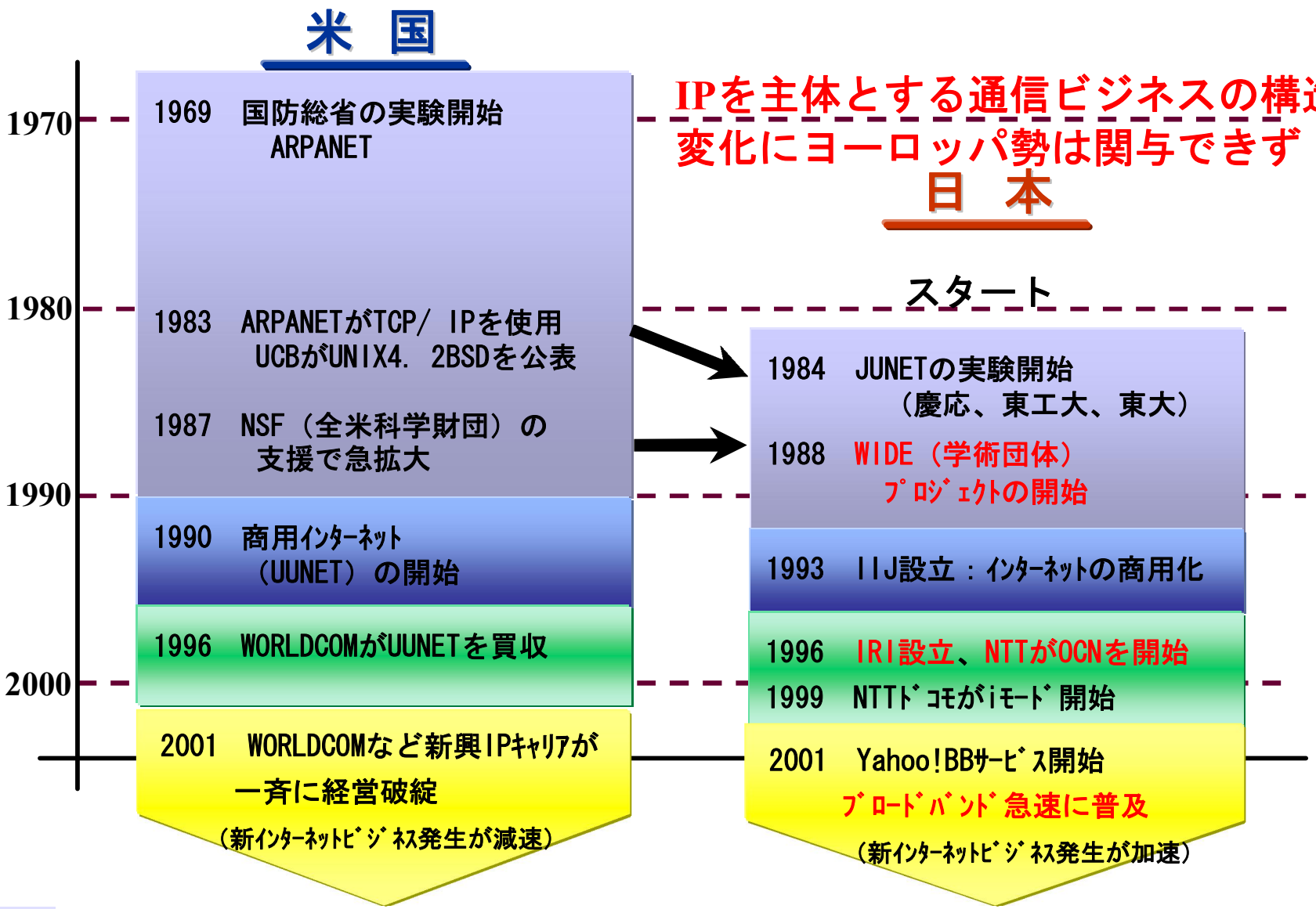
IRI Ubilabs Innovations

原理：数理科学⇒情報産業

産業の構造変化：通信、金融、ｺﾝﾋﾟｭｰﾀ、放送、家電、新聞、広告、出版、流通etc.

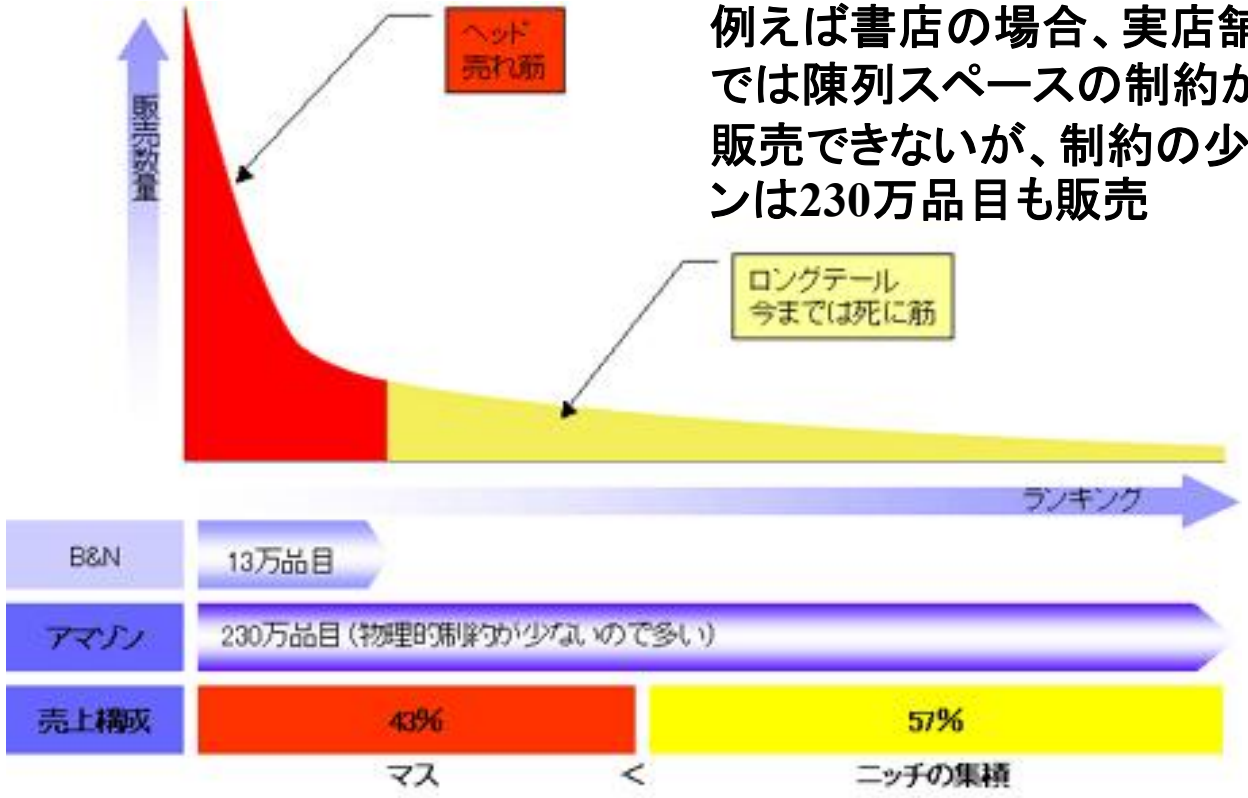
既存サービスの仮想世界への写像は起こったがネットワークそのものへの移行はこれから！

学術研究が産み出した21世紀の技術革新＝インターネット



アマゾン・ドット・コムが常識を変えた：ロングテール

ロングテールとは、ネット販売において、ほとんど売れないニッチ商品の販売額の合計が、ベストセラー商品の販売額合計を上回るようになる現象のこと。雑誌『ワイヤード』編集長のクリス・アンダーソンが提唱したもので、販売ランキング順に販売額の曲線を描くと、ベストセラーが恐竜の高い首(ヘッド)で、ニッチ商品が長い尾（テール）のようになっているところから名づけられた。



例えば書店の場合、実店舗のバーンズ&ノーブルでは陳列スペースの制約があるので13万品目しか販売できないが、制約の少ないネット書店のアマゾンは230万品目も販売

このロングテール革命こそが今日の新潮流の始まり！

2. インターネットそのもの/こちら側/あちら側を支える数理科学とは？

現在のインターネットを支える数理科学

IRI
Ubilabs Innovations

③（ネットワークの）あちら側

⇒「グラフ理論」に基づくデータベース、検索エンジン、データベース
⇒「金融工学理論」ネット金融サービス

①ネットワークそのもの

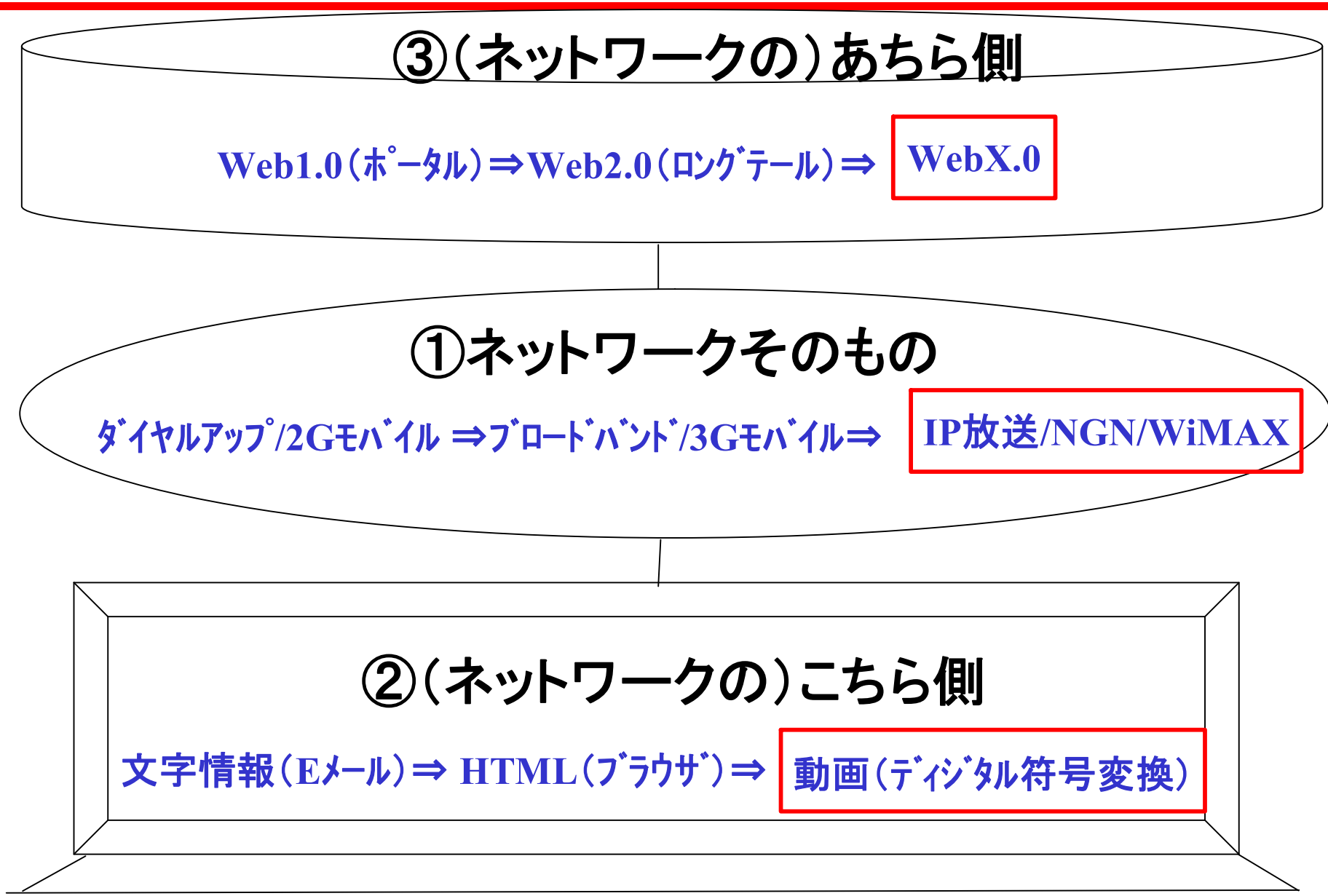
⇒「グラフ理論」による動的ルーティング、帯域制御、放送型ルーティング
⇒「ディジタル信号処理理論」に基づく変復調技術

②（ネットワークの）こちら側

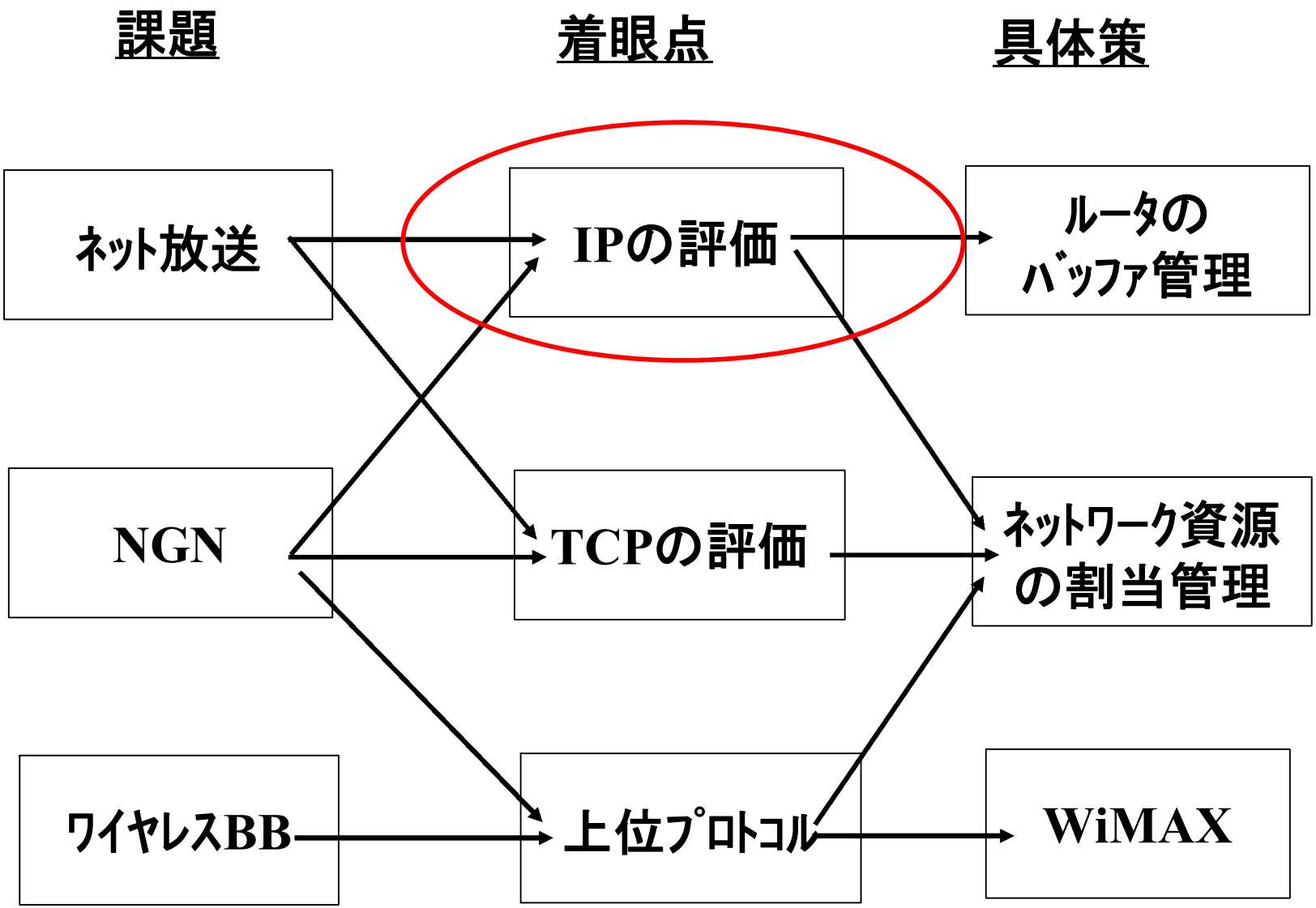
⇒「HTML/XML」に基づくブラウザ技術

⇒「ディジタル信号処理理論」に基づくコンテンツ符号化技術

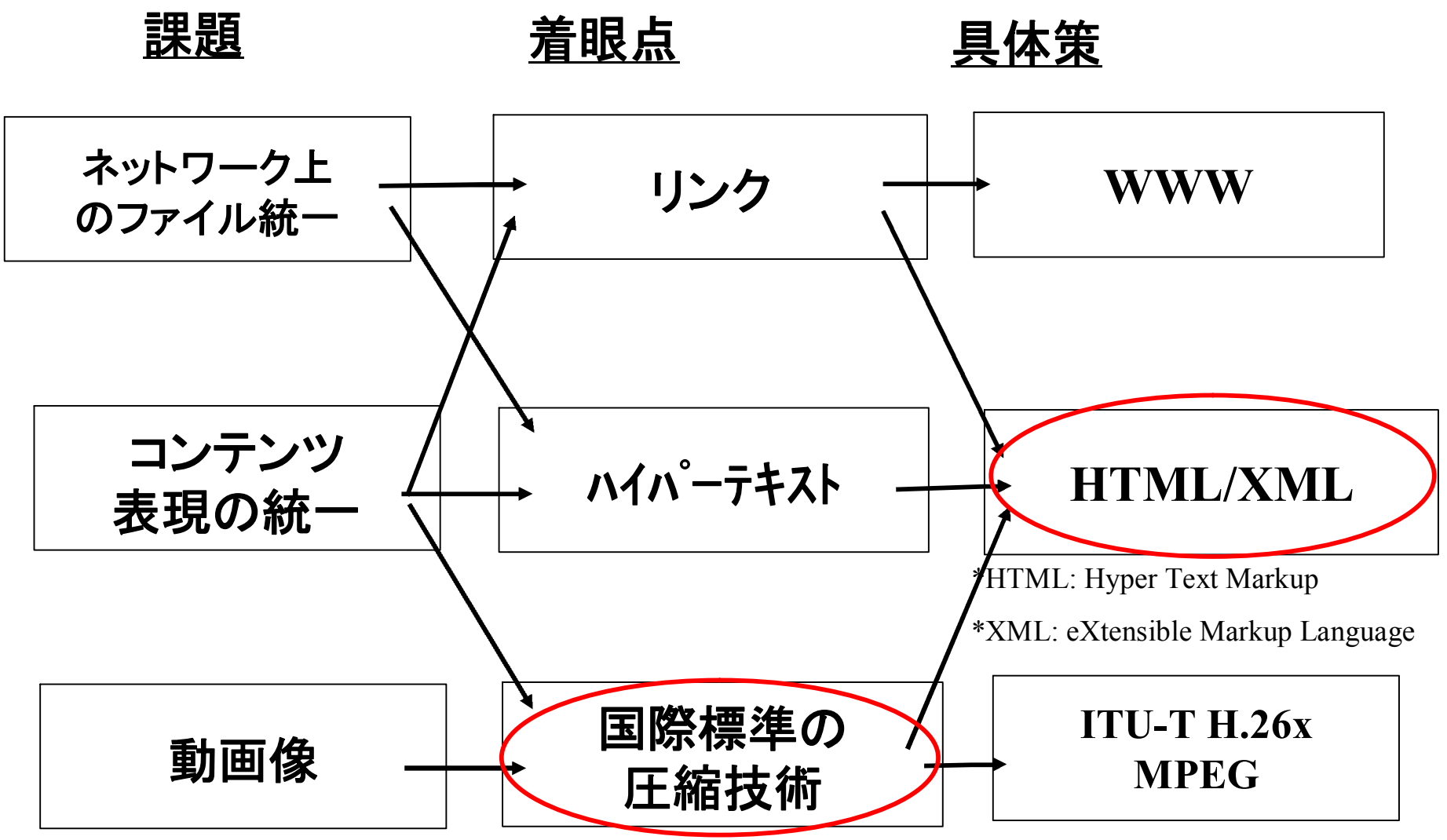
インターネットにおける技術革新と最先端のテーマ



「インターネットそのもの」を支える数理科学の課題



「インターネットのこちら側」を支える数理科学の課題



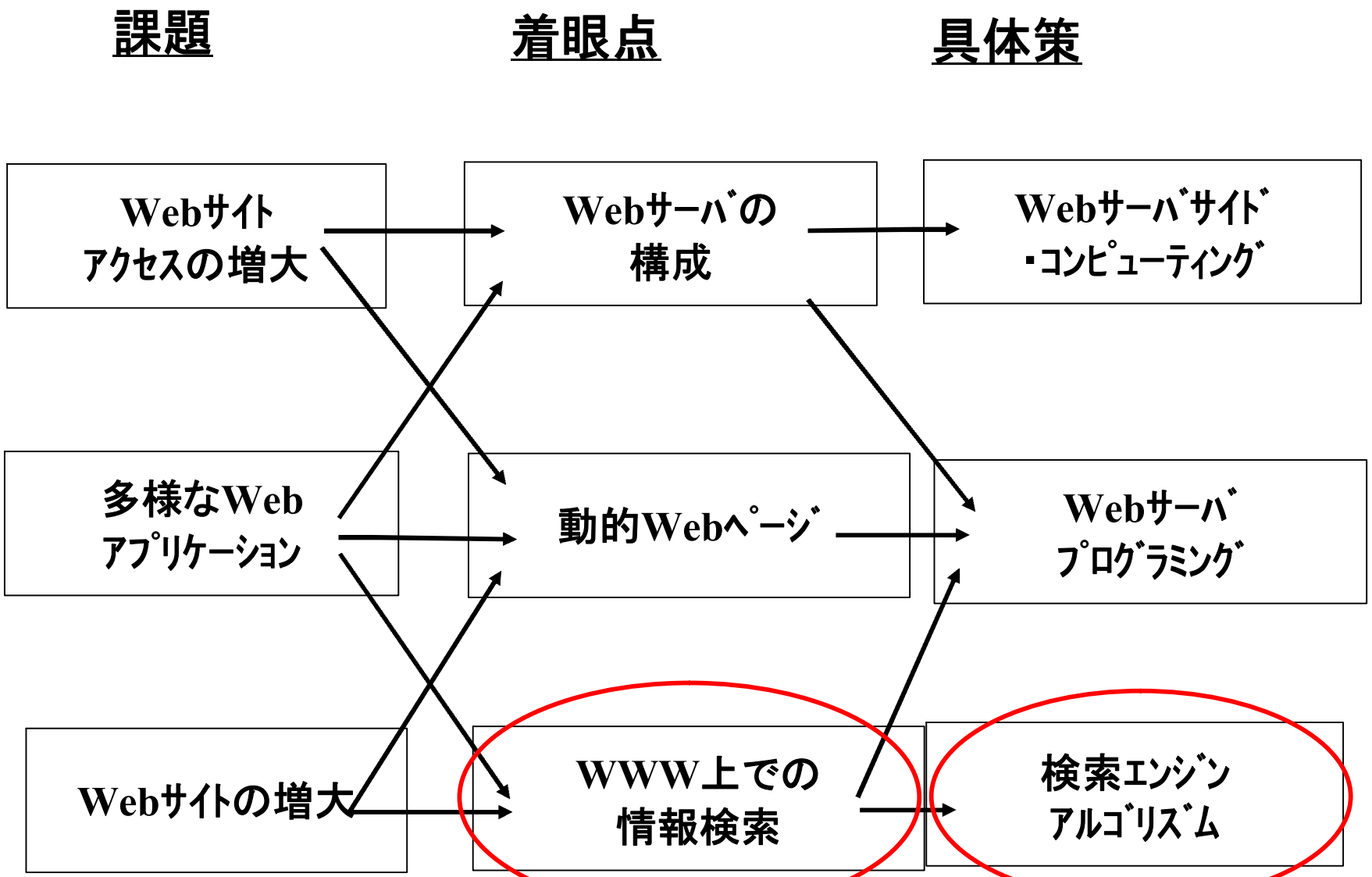
*HTML: Hyper Text Markup

*XML: eXtensible Markup Language

*ITU-T: International Telecommunications Union-Telecommunication Sector

*MPEG: Moving Picture Experts Group

「インターネットのあちら側」を支える数理科学の課題



3．インターネットの定義，基本思想とIPとは？

インターネットとは？

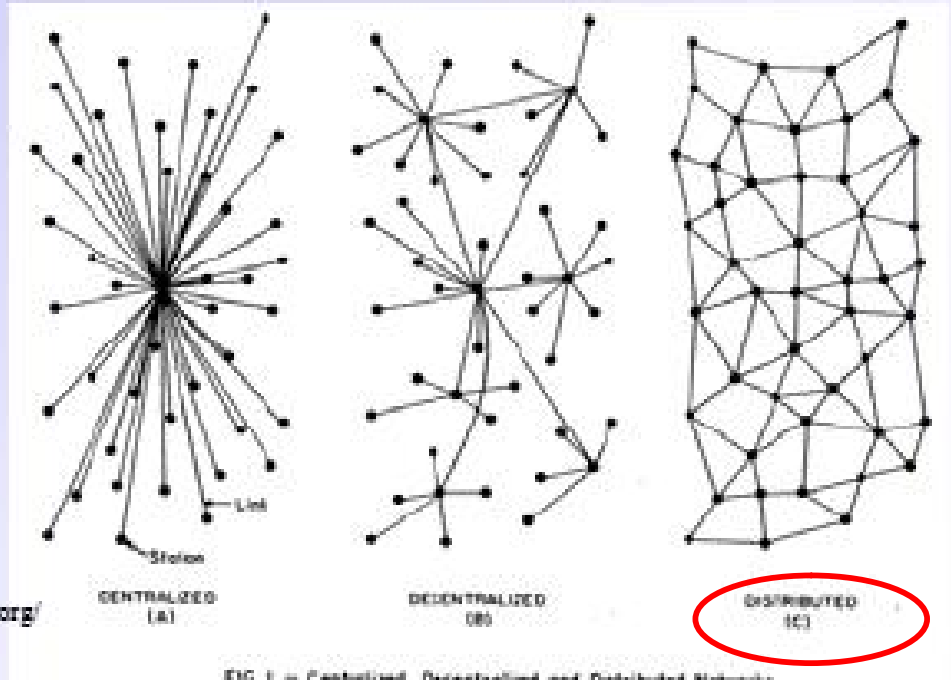
言葉の由来
　ネットワークのネットワーク⇒インターネット

広義：
　複数のコンピュータ・ネットワークをインターネットワーキングと呼ばれる技術により相互接続したネットワーク広義のインターネット(an internet)。普通名詞。

狭義：
　前述の広義のインターネットに該当するもの同士が非常に大きな規模で国際的に広く相互接続されている状態。またそれ全体をネットワークとみなしたときの呼称。
　狭義のインターネット(The Internet, The Net)。
　現在のところ唯一無二のため固有名詞として扱われる。一般に「インターネット」と呼ぶ場合はこちらを指す場合が多い。
　本講では、こちらを指すこととする。

インターネットの基本的な分散概念

- Distributed communications network
  - リンクやノードを目標とする攻撃に耐えるネットワーク
  - 広範囲、異なる要求をもつユーザへのサービスインフラ



出展:
http://www.cybergeography.org/atlas/historical.html

OSIとTCP/IPの歴史

国家間の条約に基づくISO規格のOSIか？

●1970年代中頃、ネットワーク機器各社独自のネットワークアーキテクチャが次々に発表された。機器を一つのメーカー製で揃えられるのであれば問題は無いが現実的には難しく、異なる機種同士を接続する為の標準化が急がれていた。

●ISO(国際標準化機構)の情報処理システム技術委員会は1977年3月にSC16を設置、OSI (Open Systems Interconnectionの国際標準化を開始。

●CCITT(国際電信電話諮問委員会)がOSI参照　モデル案を参考として独自の検討を開始。CCITTとSC16での意見のすり合わせを行い、基本合意。1982年にトランスポート層の標準、1983年にセッション層の標準の草稿が完成。

●1984年、情報処理システム技術委員会はSC16からSC21にOSIの標準化を引き継がせ、1985年に応用層の新プロトコルを標準化項目に追加。その後現在まで、拡張や新たなプロトコルの制定が継続。

米国のインターネット標準のTCP/IPか？

●TCP/IPの基本仕様は1982年頃にはほぼ固まっており、OSI参照モデルは1984年に完成。当初の予定ではOSI参照モデルを基に、準拠した通信機器やソフトウェアが開発・製品化していくはずであったが、TCP/IPが1980年代後半から急速に普及し、個別のOSIプロトコルに準拠した製品は普及しなかった。
●OSI参照モデルはネットワークの基本モデルとして残り、補完関係に落ち着いた。

OSI参照モデルの各層（レイヤ）

コンピュータの持つべき通信機能を階層構造に分割したモデル。OSI基本参照モデルとも呼ばれる。
1978年に、国際標準化機構(ISO)によって制定された、異機種間のデータ通信を実現するためのネットワーク構造の設計方針「OSI(Open Systems Interconnection)」に基づいて通信機能を以下の7階層に分割する。

第1層 - 物理層
　　電気信号の変換等。
第2層 - データリンク層
　　隣接する通信機器間の直接的な信号の受け渡し。
第3層 - ネットワーク層
　　ネットワークにおいて通信経路の選択。
第4層 - トランスポート層
　　エンド・トゥ・エンドのネットワークにおける通信管理。
第5層 - セッション層
　　通信プログラム（プロセス）間の通信の開始から終了までの手順。
第6層 - プレゼンテーション層
　　データの表現方法。
第7層 - アプリケーション層
　　ユーザー・アプリケーションが操作するインターフェース。

OSI参照モデルの実際例

7	アプリケーション層	HTTP, SMTP, SNMP, FTP, Telnet, AppleTalk, X.500
6	プレゼンテーション層	SMTP, SNMP, FTP, Telnet
5	セッション層	NetBIOS, NWLink, PAP，名前付きパイプ
4	トランスポート層	*TCP*, UDP, SPX, NetBEUI
3	ネットワーク層	*IP*, ARP, RARP, ICMP, DHCP, IPX, NetBEUI
2	データリンク層	イーサネット, トークンリング，PPP, フレームリレー
1	物理層	電話線, 無線, 光ケーブル

OSIモデルは仕様ではなく指針であるため、全てのプロトコルやネットワークがOSIモデルに沿って実装されているとは限らない。従って、一部のプロトコルやサービスに関しては、OSIモデルのどの層に属するかについて、幾つかの異なる見解が存在する。複数層に跨っている物もある。図示の例はあくまでも例に過ぎない。

IPのメリット

IRI
Ubilabs Innovations

1．上位レイヤを規定しない

⇒アプリケーションを自由に作れる

2．下位レイヤを規定しない

⇒どのような通信網（銅線、無線、光ファイバ）も使える

今後は、何でも つながる/つなげる

IRI
Ubilabs Innovations

電話線　無線　ファイバ　電力線　・・・・　・・・・

ipv6
IPv6 Promotion Council

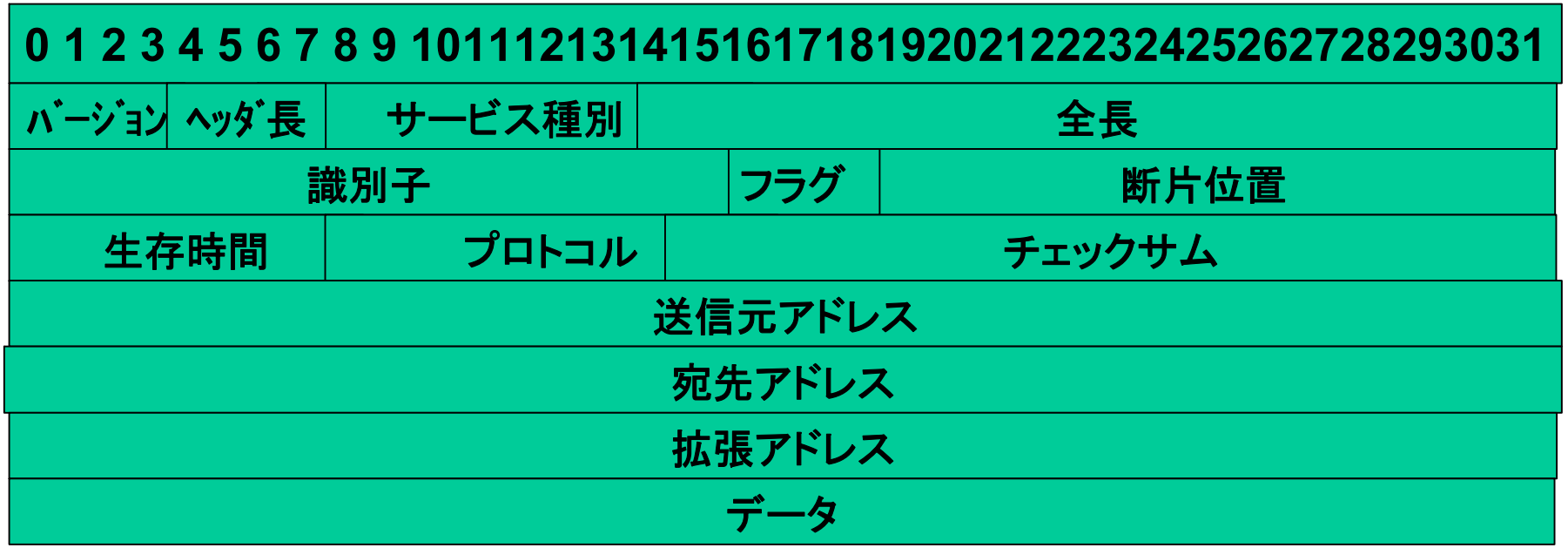
32ビットアドレス長のIPv4

Internet Protocol（インターネットプロトコル、IP、IPv4）は、OSI参照モデルにおいてネットワーク層に位置付けられるプロトコルである。
転送の単位であるパケットの経路選択と、その断片化と再統合を主な機能とする。
TCP/IPの基本機能としてインターネットなどで世界中広く用いられている。

IPパケット

IPパケットの先頭には必ずIPヘッダが付加され、それにより経路選択などのIPの機能が実現されている。以下にパケット形式図とそれぞれの領域の役割などを記す。

パケット形式図



IPに代表されるインターネットの精神

IRI
Ubilabs Innovations

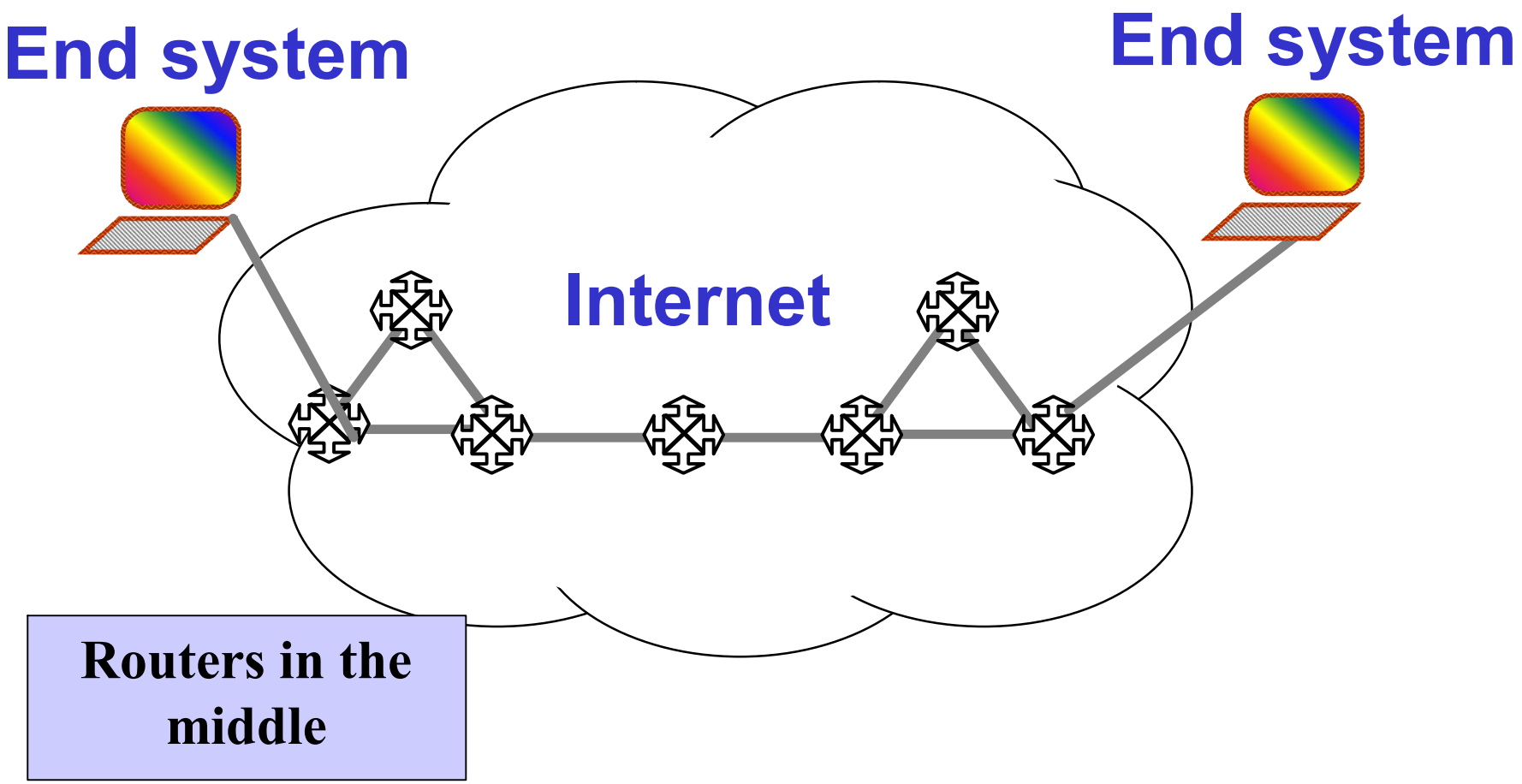
IETF (Internet Engineering Task Force) でRFC (Request for Comments) を作成してインターネットの標準を作っていく精神

"We reject kings, presidents, and voting.

We believe in
rough consensus and running code."

-Dave Clark (1992)

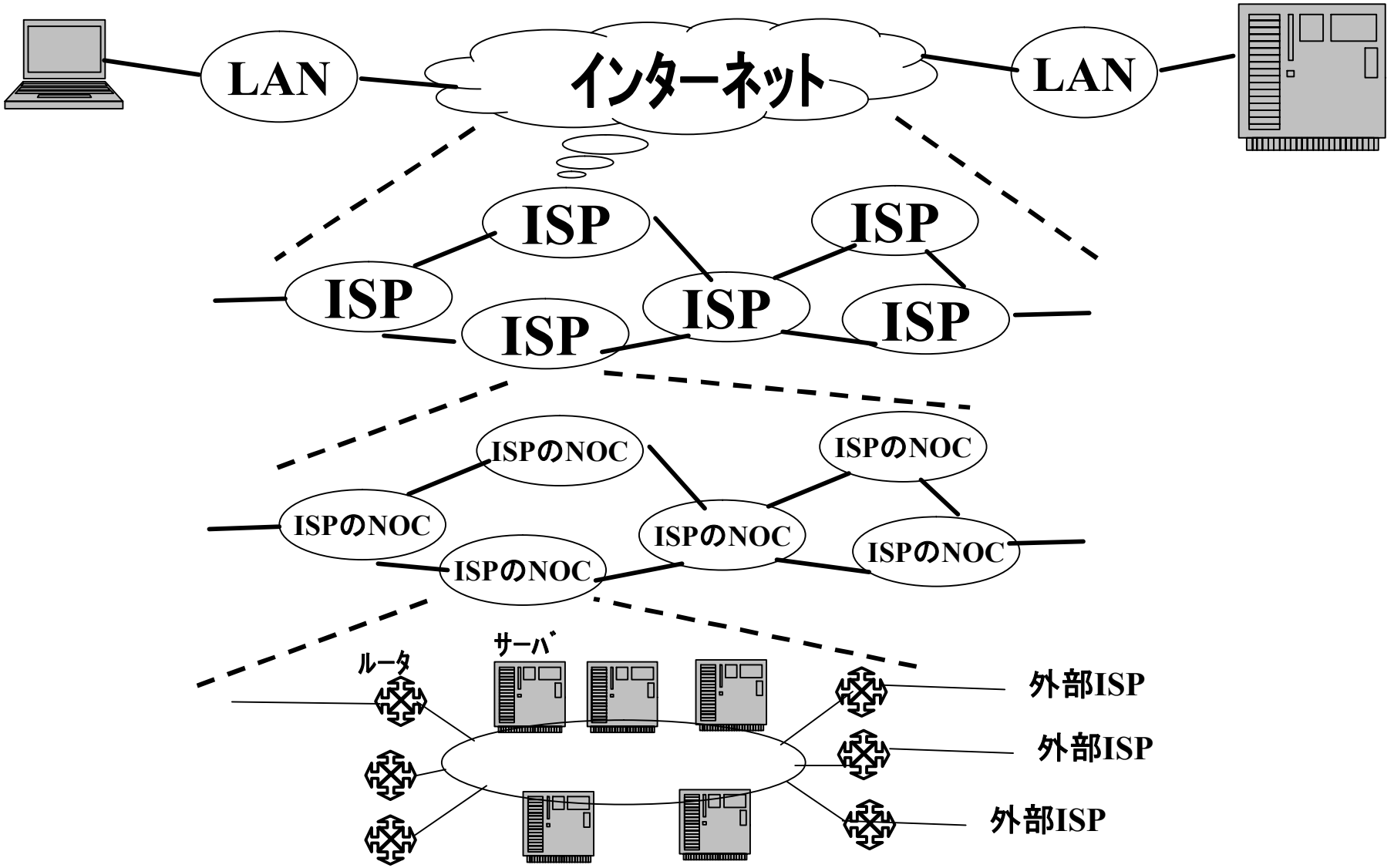
インターネットの基本は“End-to-End Model”である



出典：江崎浩（東京大学大学院情報理工学系研究科教授）　Interop2006 Executive Summit

4. IPルーティングのための数理科学

ルータによるインターネットの基本構成



NOC: Network Operation Center

経路制御（ルーティング）とルータとは？

IRI
Ubilabs Innovations

1．目的地までの接続経路＝ルート

2．どの接続経路かを決定する装置＝ルータ

3．接続経路を決定すること＝経路制御

4．経路制御の実際＝ルータXに接続される
Webサーバをアクセスする
→DNSからWebサーバのIPアドレスを見つける
→最寄のルータにIPアドレスを渡すだけで
自動的にアクセスする

最短経路アルゴリズム（Shortest-Paths Algorithms）

●グラフ理論における基本的な問題のひとつは、グラフ中の2頂点間を結ぶ最短経路を見つけることである。形式的に経路はグラフ G = (V, E) 中の頂点のシーケンス <v0,v1,...,vk> で表される（辺 (vi,vi+1) for i=0,1,...,k-1 は 辺の集合 E ）。
シーケンスにおいて各頂点は次の頂点へ接続される。最短経路問題において、各辺は重みを数値として与えられている。それゆえ、経路の重み(weight of a path) について記す
w(p) = i=1..k of w(vi-1,vi) の合計

●頂点 u から v に至る最短経路の重み (shortest path weight) は
delta (u,v) = min { w(p) : u --> v } もし 頂点 u から v に至る経路が存在すれば
delta (u,v) = 無限(infinity) そうでなければ（ u から v に至る経路が無ければ）
最短経路は重みの合計が最小となる経路といえる。

●最短経路問題には、いくつかの変形された問題がある。ここでは単一ペアの問題を定義した、しかし、さらに単一出所問題(グラフ中の1つの頂点から各頂点ごとまでの最短のパス)があり、等価な単一目的地問題、全ペア問題、等である。単一出所の問題を解決するアルゴリズムより漸近的に速い、単一ペアの問題を解決するアルゴリズムは存在しない。

●最短経路木(shortest-paths tree) は、グラフ G=(V,E) 中のある頂点を原点とした有向サブグラフである。V‘ を V のサブセット、E’ を E のサブセットとし、V‘ はG’ から到達可能な頂点のセット、G‘ は原点から連なる経路木を成すものとすれば、V’ 中の全ての頂点 v は G‘ 中の頂点v から唯一の経路を持つ。再帰的に、単一頂点アルゴリズムによる結果は最短経路木である。

3つのルーティング手法

IRI
Ubilabs Innovations

1. スタティック・ルーティング
   - ルータに手作業で宛先を設定
   - ネットワークの状態は不問

2. デフォルト・ルーティング
   - ルータの知らない宛先のトラフィックを所定の出口へ送る
   - 唯一の出口で接続されるドメインで使用

3. ダイナミック・ルーティング
   - インテリア・ルーティングやエクステリア・ルーティングから学習したルート
   - ネットワークの状態に依存

ダイナミック・ルーティングの基本的考え方

IRI Ubilabs Innovations

　指示された経路が有効でなくなっている場合、現存するノードを使った別の経路を決めなければならない。
　これは通常ルーティングプロトコルと経路決定アルゴリズムによってなされる。経路決定アルゴリズムには二種類ある。

①距離ベクトルアルゴリズム (distance vector algorithm, DVA)

RIP (Routing information Protocol)

②リンク状態アルゴリズム (link state algorithm, LSA)

OSPF (Open Shortest Path Fast)

IS-IS (Intermediate System-to- Intermediate System)

この内どちらか一方が用いられる。
　インターネット上の経路決定問題は、これら2つに尽きる。
以下用いられる*「コスト」*ないし*「距離」*は経由するルータの数（「ホップ数」）や回線速度を数値化したもので、*「メトリック metric」*と呼ばれる。メトリックの決定法はプロトコルによって異なる。

①距離ﾍﾞｸﾄﾙｱﾙｺﾞﾘｽﾞﾑ：DVA (distance vector algorithm)

DVAは Bellman-Ford アルゴリズムを用いている。この方法では、各ノード間に「コスト」と呼ばれる数値が割り振られる。二点間を結ぶ経路のコストは、その間に経由するノード間のコストの総和であり、その情報はノードから得られる。

アルゴリズムは極めて単純である。最初の段階では、各ノードは直近のノードがどれかという情報と、それらの間とのコストだけを知っている（このような、「行き先リスト」とそれぞれの総コスト、やりとりするべき「次の相手(next hop)」を集めたものがルーティングテーブルないし、ディスタンステーブルである）。

代表例としてRIP (Routing information Protocol)がある！

定期的にノード間でやりとりがなされ、互いにルーティングテーブルのデータを交換する。もし隣から渡されたデータに、自分のルーティングテーブルより優れたもの（同じ行き先に到達するのに、コストが少ない）があれば、それを用いてテーブルを更新する。自分のテーブルにない相手への情報が入っていた場合も同様である。時間をかけると、全てのノードがあらゆる宛先についての最良の「次の相手」と最良の「コスト」を見つけだす。

あるノードが脱落した場合は、そこを「次の相手」としていたノード全てにおいて、ルーティングテーブルの破棄と再構築が行われる。この情報は隣のノードに次々 伝えられて行き、最終的には到達可能な全てのノードについて最良の経路が発見されることになる。

②リンク状態アルゴリズム：LSA（ link state algorithm ）

LSAでは、各ノードが用いるのはネットワークのマップであり、それはグラフの形で格納されている。このマップをつくるために、全てのノードがネットワーク全体に「自分が接続しているノード」をブロードキャストする。各ノードはそのデータをもとに、個々独立してマップを計算し生成する。自分で生成したマップをもとに、各ノードは他のノードへの最短経路を決定する。
最短経路の計算にはダイクストラのアルゴリズムが用いられる

このアルゴリズムはネットワーク全体を木構造で表現する。木の根（最初の要素）は各ノードそれ自体である。次いで、ノードの集合から未登録のノードを一つずつ木に加えていく。

加えるノードは既に木に存在するノードのどれかから到達できるノードのうち、最も少ないコストで到達できるものである。ネットワーク上の全てのノードを登録するまでこれを繰り返す。

木構造ができあがったら、それを用いて、ルーティングテーブルをつくる。最良の「次の相手」等がそこに登録される。

代表例としてOSPF (Open Shortest Path Fast)がある！

ダイクストラ法（構造化ﾌﾟﾛｸﾞﾗﾐﾝｸﾞの提唱者ｴﾄﾞｶﾞｰ･ﾀﾞｲｸｽﾄﾗ が考案）

Ubilabs Innovations

グラフ理論における最短経路問題(与えられた重み付きグラフの2頂点間を結ぶ路の中で最小の重みを持つ路を求める問題)を解くためのアルゴリズム。重み付きグラフにおいて始点sから他の点への最短経路を求める。まず、

S: = {s}

$T := V \setminus \{s\}$

p(s): = 0

とする。さらに、各 $i \in T$ に対して、もし辺siが存在すれば

p(i): = wsi

q(i): = s

とする。ただしここでwijは辺ijのコストとし、辺ijが存在しない場合は +∞とする。Vは頂点集合である。辺siが存在しない場合は、$p(i) := +\infty$ とする。

次に、以下の操作を、Tが空集合となるまで繰り返す。

$p(k) = \min\{ p(j) \mid j \in T \}$

と $k \in T$ を一つ取り、

$S := S \cup \{k\}$

$T := T \setminus \{k\}$

とする。

$j \in T$ に対して、もしp(k) + wkj < p(j)ならば

p(j): = p(k) + wkj

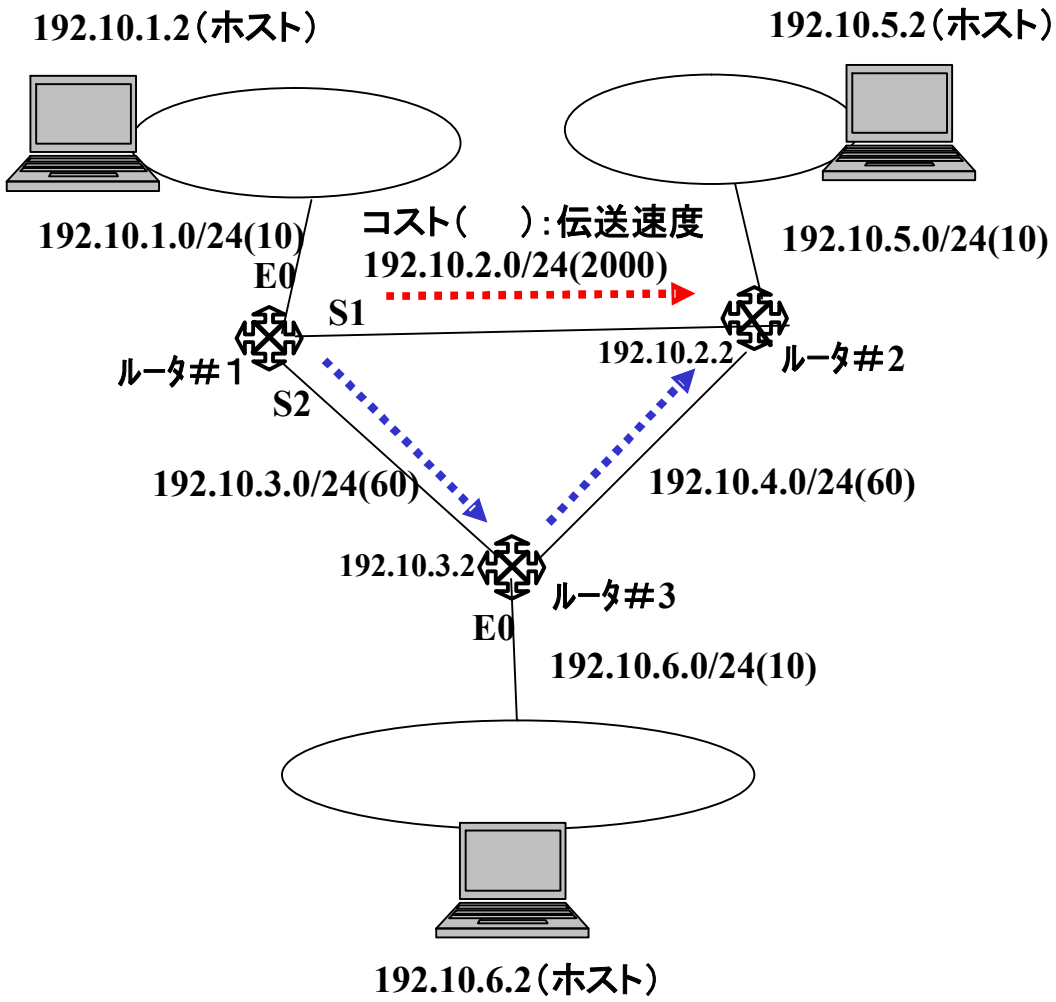
q(j): = k

とする。

これが終了したとき、pjはsからjへの最短距離となっている。また、点列 $\{j, q(j), q(q(j)), \ldots\}$ は、その最短経路での経由点を逆順にしたものになっている。

DVAかLSAどちらが有利か？

RIPまたはOSPFで



ルータ＃1のルーティングテーブル（RIP）

宛先	ネクストホップ	ホップ数
192.10.1.0	接続（E0）	—
192.10.2.0	接続（S1）	—
192.10.3.0	接続（S2）	—
192.10.4.0	192.10.2.2 (S1) 192.10.3.2 (S2)	1 1
192.10.5.0	192.10.2.2 (S1)	1
192.10.6.0	192.10.3.2 (S2)	1

ルータ＃1のルーティングテーブル（OSPF）

宛先	ネクストホップ	コスト
192.10.1.0	接続（E0）	—
192.10.2.0	接続（S1）	—
192.10.3.0	接続（S2）	—
192.10.4.0	192.10.3.2 (S2)	120
192.10.5.0	192.10.3.2 (S2)	130
192.10.6.0	192.10.3.2 (S2)	70

LSAの特徴とメリット

IRI Ubilabs Innovations

1. 反復分散ﾃﾞｰﾀﾍﾞｰｽﾓﾃﾞﾙ使用したより複雑なﾌﾟﾛﾄｺﾙ
2. ﾘﾝｸｽﾃｰﾄ（情報要素：ﾄﾞﾒｲﾝ内ﾘﾝｸやﾉｰﾄﾞの情報）をﾙｰﾀが交換
3. ﾙｰﾃｨﾝｸﾞ・ﾃｰﾌﾞﾙは、交換せず、隣接ﾙｰﾀ、接続に関するﾒﾄﾘｯｸ情報を保有。
4. ｼﾞｸﾞｿｰ・ﾊﾟｽﾞﾙ型ｱﾙｺﾞﾘｽﾞﾑで、ﾄﾞﾒｲﾝ内の全ﾉｰﾄﾞが、ﾊﾟｽﾞﾙﾋﾟｰｽ情報のｺﾋﾟｰを受信。
5. ﾈｯﾄﾜｰｸ内の各ﾙｰﾀは、個々にﾊﾟｽﾞﾙを組み立てる。

→ホップ数に制約されない
→リンクの帯域と遅延を計算可能
→収束の速さ
→VLSMとCIDRをサポート
→階層化に優れる

しかし、LSAは、ﾄﾞﾒｲﾝ間ﾙｰﾃｨﾝｸﾞへの適用不可！

そこで、DVA型の*BGP*をﾄﾞﾒｲﾝ間ﾙｰﾃｨﾝｸﾞへ適用！

ドメイン間(ポリシーで決まる)ルーティングの小史

- 1987年完成したNSFNETでは初期のEGP (一般的なExterior Gateway Protocolではない) を使用
  - ⇒ ・ルーティンググループ上の制約
    ・トポロジー上の制約
- 現在：
  インターネット上の事実上の標準＝BGP-4

AS(autonomous system)

●インターネットにおける自律システム(autonomous system)（AS）とは、インターネットに接続する1つ（時に複数）のルーティングポリシー配下にあるIPネットワークやルータの集合。

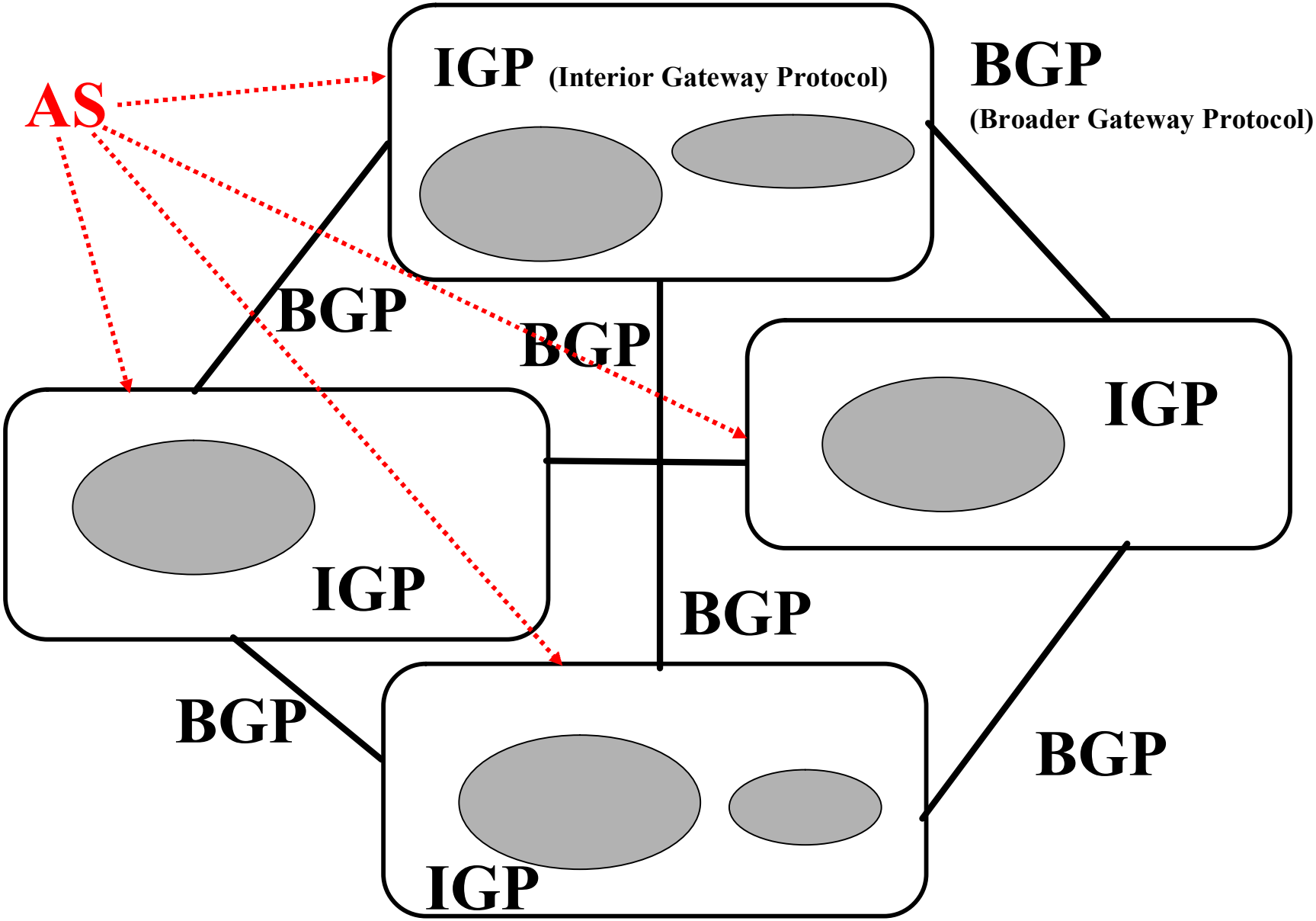
●元来インターネットサービスプロバイダや複数のネットワーク接続されたに巨大な　組織など、１つのルーティングポリシーによって制御されているネットワークと定義。

●RFC 1930で新しく定義されたのは、グローバルAS番号を持ったISPに接続される複数の組織においてプライベートAS番号を使用してBGPを走らせ、インターネット接続可能になった。

●例えそのISPが複数のASを抱えていても、インターネットからはISPのルーティングポリシーが見えるだけ。ユニークなAS番号（AS number）はBGPのルーティングを行うのに必要なため、各ASごとに割り振られている。

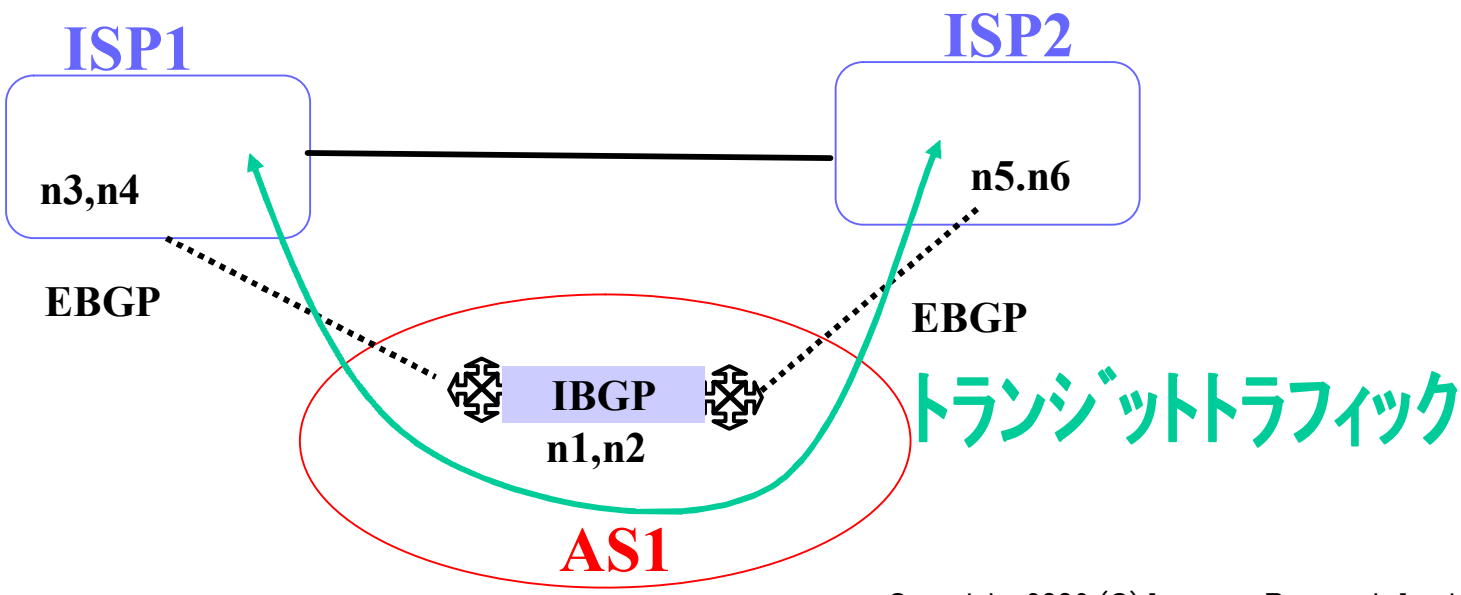
●BGPにおいて、AS番号はそのユニークさをもってインターネット上の各々のネットワークを認識する。

AS，EGP，IGPの関係



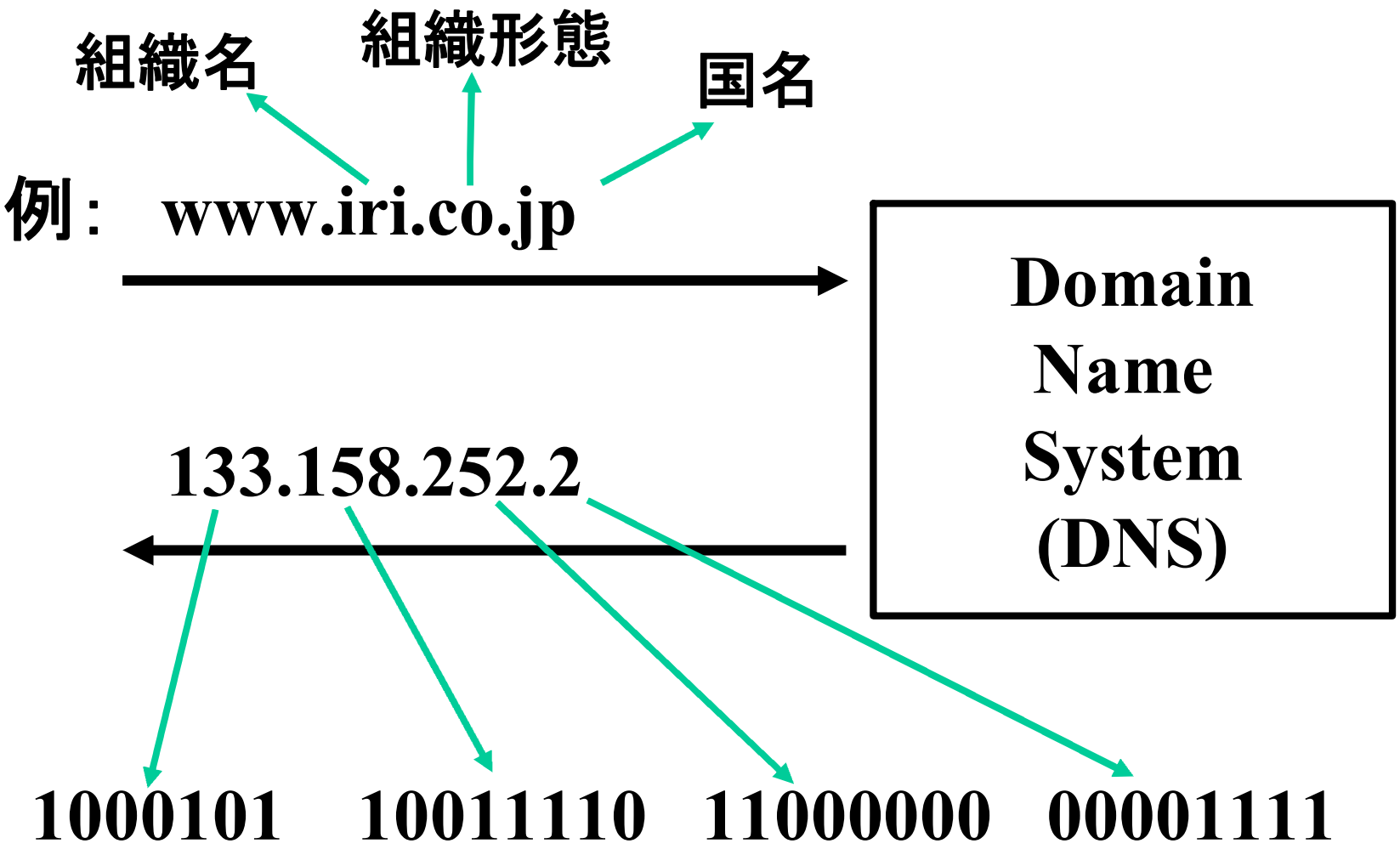
各種のASとトランジットの関係例

- マルチホーム・トランジットAS
  - ・マルチホーム・トランジットAS＝外部に複数の通信路があり、他のASからのトランジットトラフィックに使用可能
  - ・トランジットトラフィック＝ローカルASに属さない起点/宛先を有するトラフィック
  - ・BGP-4は、エクステリアゲートウェイプロトコルだが、BGPの更新情報交換用パイプとしてAS内で利用可能
    ＝内部BGP（IBGP）
  - ・外部とのルータ間通信路＝外部BGP（EBGP）



5. 情報アクセスと圧縮のための数理科学

実際のIPアドレスはドメイン名で利用される



グラフの例としてのWWW

●WWWにおけるウェブページの、リンク・被リンク関係がなす構造は、有向グラフの一種である。
●World Wide Web（WWW、ワールド・ワイド・ウェブ）は、インターネットで提供されるハイパーテキストシステム。単にWeb（ウェブ）と呼ばれることも多い。インターネットは本来、TCP/IPで接続されたコンピュータ・ネットワークを指す言葉であるが、日常用語ではWWWのことだと解釈している人も多いがそれは違う。

Webの歴史

●Webの根底にある考え方は1980年にティム・バーナーズ＝リーが Robert Cailliau と構築したENQUIREに遡ることができる（ENQUIREは一般に公表されるまでいかなかった。その名称は Enquire Within Upon Everything というビクトリア朝時代の日常生活のハウツー本に由来していて、バーナーズ＝リーが幼少のころを思い出して付けたものである）。それは現在のWebとは大分違うが、根本的なアイデアの多くを含んでいる（さらには、バーナーズ＝リーのWWW後のプロジェクトである Semantic Web の考え方も含んでいる）。

●1989年3月、欧州原子核研究機構(CERN) のティム・バーナーズ＝リーは「Information Management: A Proposal（情報管理：提案）」を執筆し、ENQUIREを参照しつつさらに進んだ情報管理システムを描いた。彼は1990年11月12日、World Wide Web をより具体化した提案書を発表した。実装は1990年11月13日から開始され、バーナーズ＝リーは最初のWebページを NeXTワークステーション上に置いた。

同年のクリスマス休暇の間に、バーナーズ＝リーは Webに必要な全ツールを構築した。世界初のWebブラウザ（Webエディタでもある）と世界初のWebサーバである。

●1991年8月6日、彼は後述のWorldWideWeb - Executive Summaryを alt.hypertextニュースグループに投稿した。この日が Webがインターネット上で利用可能なサービスとしてデビューした日となる。

●ハイパーテキストの概念は1960年代まで遡ることができる。テッド・ネルソンの Project Xanadu、ダグラス・エンゲルバートの oN-Line System（NLS）などである。

Web前史

●ヴァネヴァー・ブッシュのマイクロフィルムベースの「memex」にインスパイアされたものであり、memex は1945年の論文「As We May Think」で描かれている。
●テッド・ネルソンによる1965年のProject Xanadu、1987年アップルのHypercard。
●バーナーズ＝リーのブレイクスルーはハイパーテキストとインターネットを結合したこと。彼の著書『Weaving The Web』では、このふたつの技術の結合は双方の技術コミュニティの協力によって成立することを強調しているが、誰もこの提案を取り上げることはなく、彼は最終的に自分でプロジェクトを実行した。この過程で彼はURIと呼ばれるグローバルな資源識別子を開発した。
●World Wide Web は当時実現していた他のハイパーテキストシステムとはいくつかの点で異なる。ネットワークをまたがりコンピュータネットワーク上に実現した。

バーナーズ＝リーがCERNで使用していた
NeXTcube。最初のWebサーバとなった

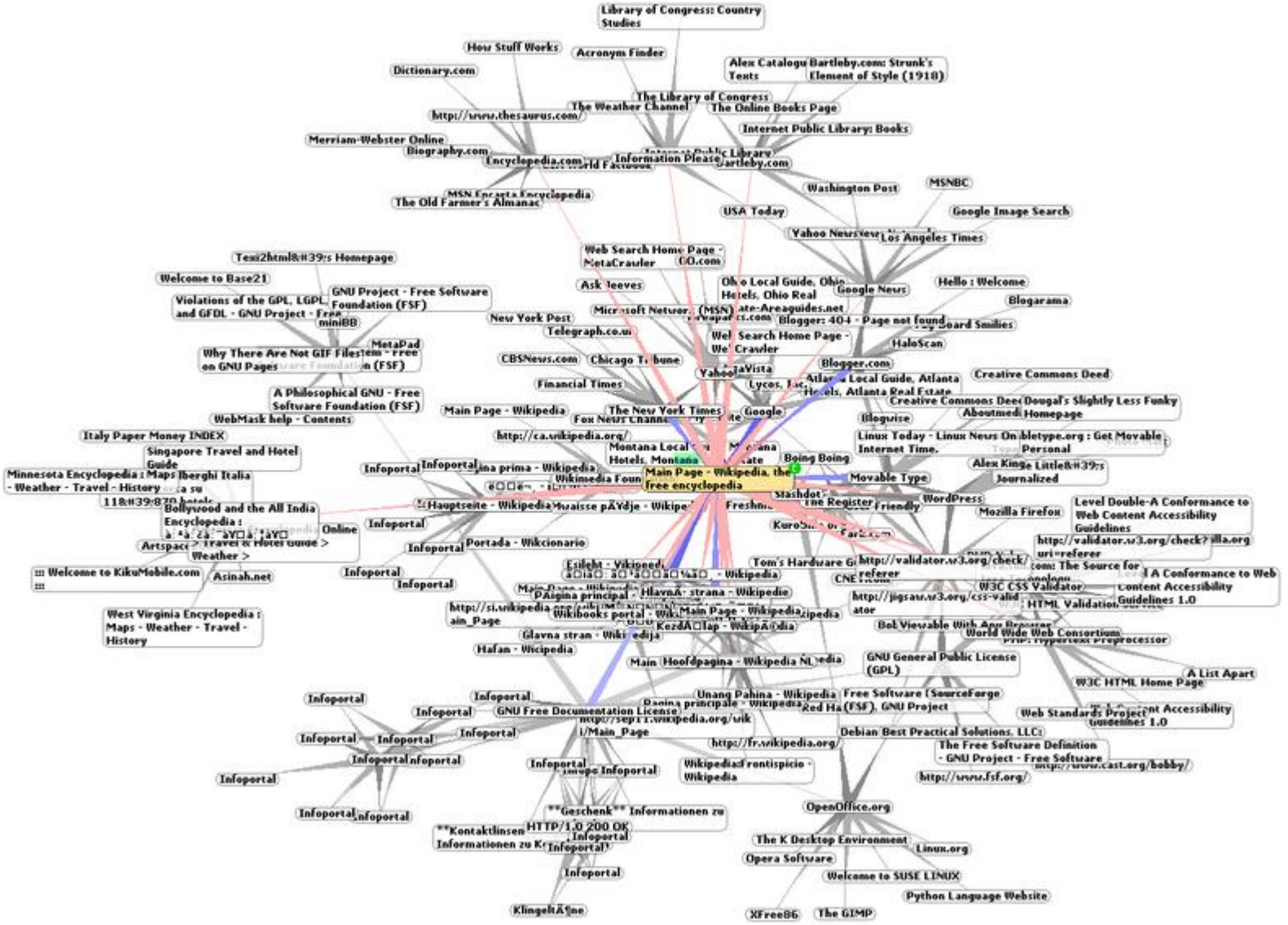
●WWWは、双方向ではなく単方向のリンクを使用する。これにより、何らかの資源の所有者と連絡を取らなくてもリンクすることが可能となった。これによって Webサーバやブラウザの実装も簡単になっているが、同時にリンク先の資源がいつの間にか無くなるという問題も発生させることとなる。

●HyperCardやGopherとは違い、World Wide Web は、一企業に独占されておらず、サーバや　クライアントを独自に開発し拡張するのも自由にできてライセンスを得る必要も無い。 開発当初、WWWは文字情報を扱うだけの比較的単純なものであった（NeXT上で開発されたためOS自身が文字以外を適切に扱うため、WWWは情報を区別しなくてもよかったというのが真相）。

●1992年、現在のような画像なども扱えるWWWにしたのが、イリノイ大学に設置されている米国立スーパーコンピュータ応用研究所 (National Center for Supercomputing Applications・NCSA) である。ここの学生であったマーク・アンドリーセンらは、文字だけでなく画像なども扱える革新的なブラウザ「Mosaic」を開発。そしてこのソフトに改良を加えるために無料でソースコードを公開したため、Mosaicはたちまち普及し、WWWは誰でも手軽に使うことのできる世界的なメディアとなった。

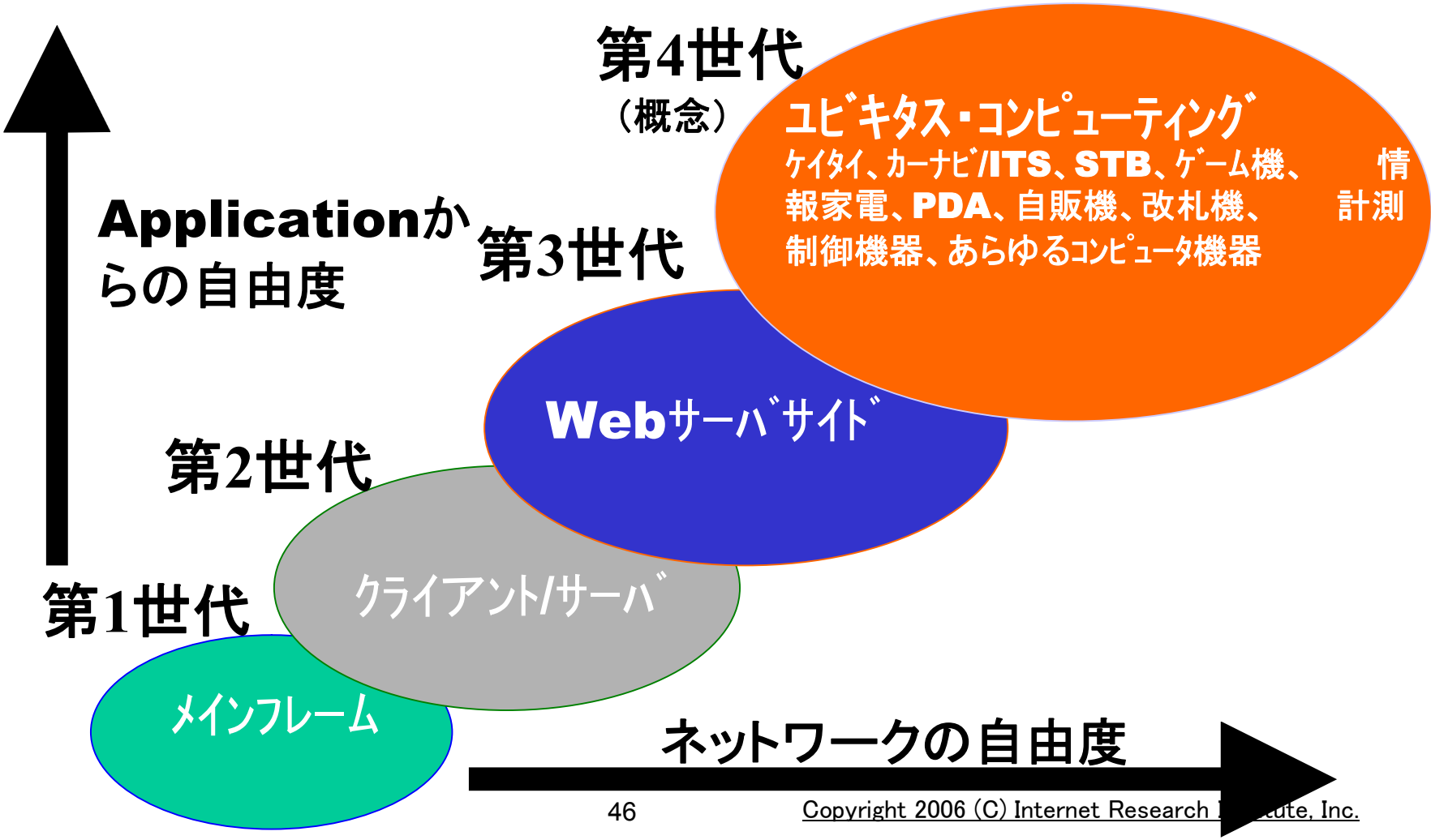
●1993年4月30日、CERNは World Wide Web を無料で誰にでも開放することを発表。日本最初のホームページを開設したのは、高エネルギー加速器研究機構所属の森田洋平氏。

Wikipediaの周辺の World Wide Web



“The Network is The Computer.”　Bill Joy.

“C&C：Computer & Communication”小林宏治



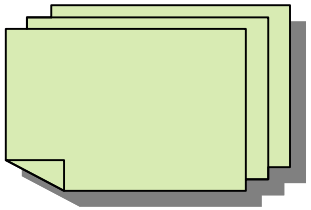
Webの4大要素のよる各種サービスの実現

IRI Ubilabs Innovations

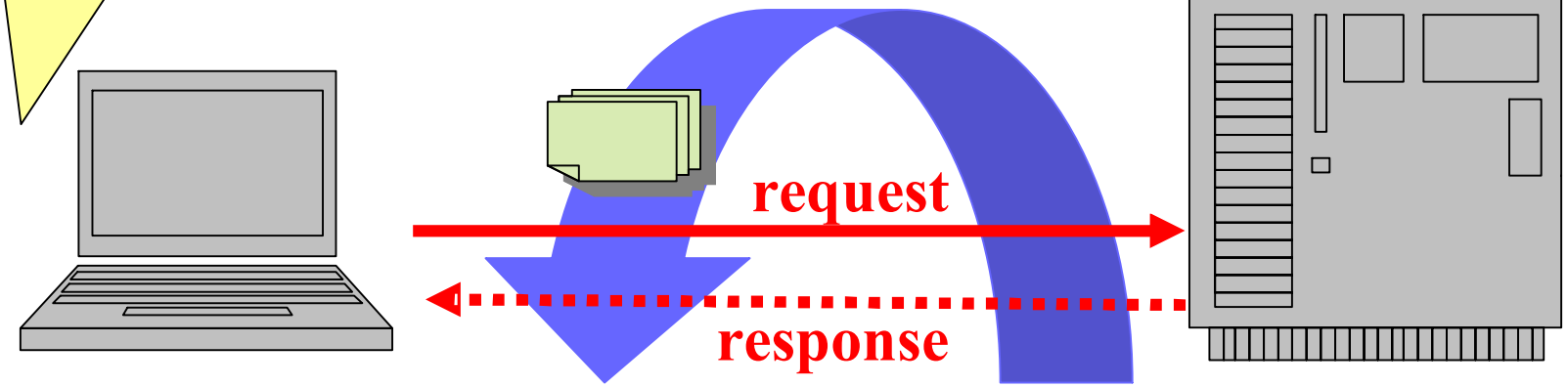
情報検索
予約サービス
電子商取引
Web掲示板

これが基本！

①記述言語：HTML



④送受信プロトコル：HTTP



③情報受信：Webブラウザ
(Internet Explorer, Navigator,etc.)

②情報発信：Webサーバ
(IIS, Apache,etc.)

WebブラウザにURLを入力するか、
Webページのリンクをたどればよい！

CGI,SSI,Java

CGI: Common Gateway Interface

SSI: Server Side Include

最近のWebサーバは、ユーザー入力で
起動するWebアプリケーションを実装する！

Webサーバサイドとクライアントサイドの処理分担

- Webサーバサイド
  ・データベース処理等の機密性の高い処理
  ⇒暗号化、アクセス制限（クライアントと協調）
  ⇒ここにも多くの数理科学的手法が存在

- クライアントサイド
  ・マウスクリックに即座に反応すべき処理
  ・アニメーション、ポップアップメニュー等

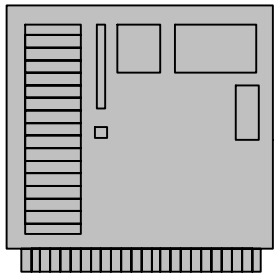
- Webサーバサイドコンピューティングへ収束
  ・クライアント側　：　汎用Webブラウザだけを実装
  ・Webサーバ側：　専用ハードウェアやソフトウェアも実装
  ・前提条件　：　HTTP以外の通信プロトコルは使用せず

Webサーバサイド・コンピューティングの進化と3層化

【クライアント】

request

response



【サーバ】

2層モデル

第1層：クライアント層 ..... 第2層：サーバ層

3層モデル

第1層：ユーザー対話層 ..... 第2層：ビジネスロジック層

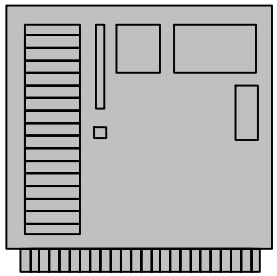
第3層：データベース層

進化

【Webブラウザ】

request

response



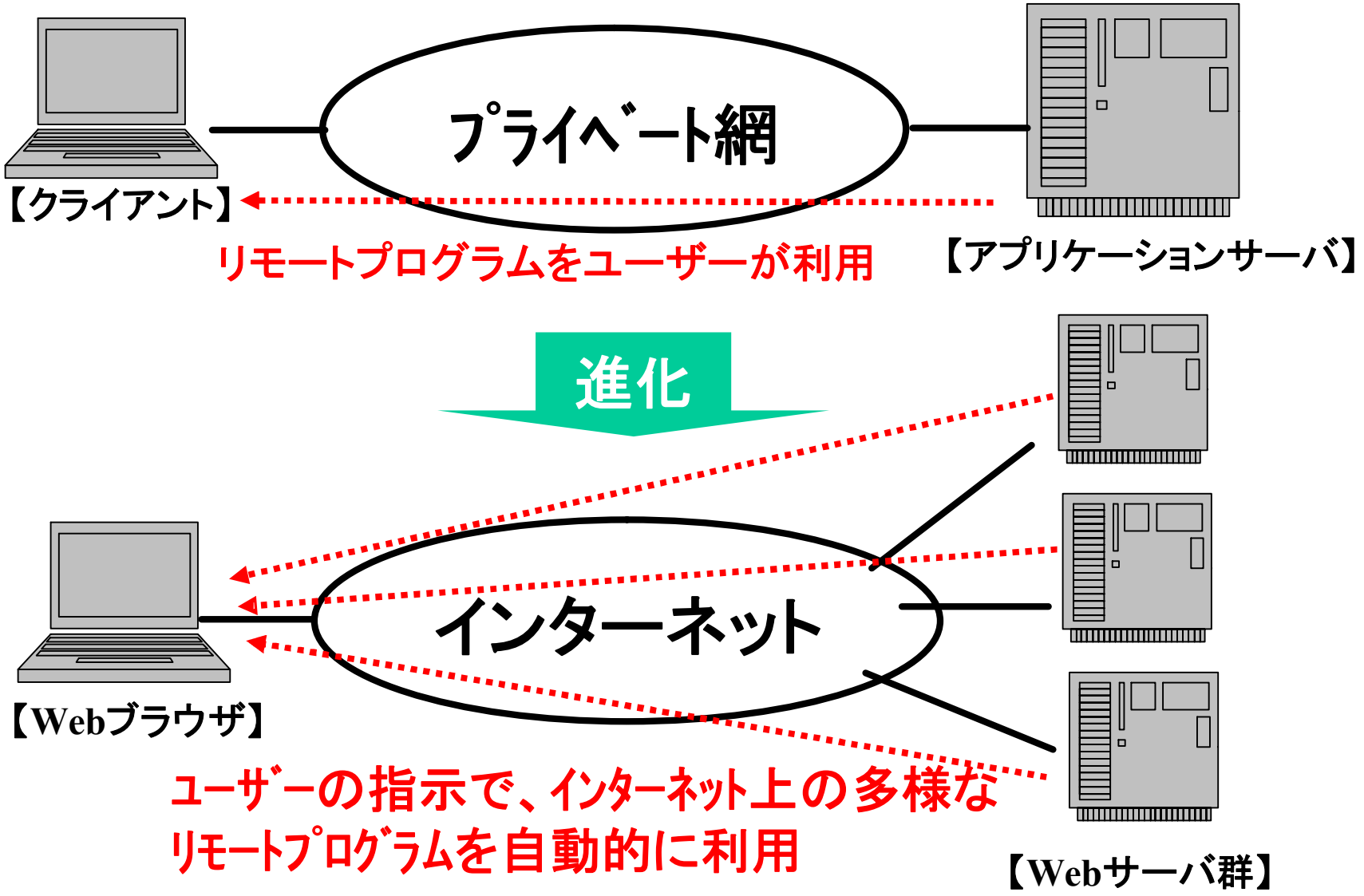
【Webサーバ】

3層モデル

第1層：汎用ブラウザ ..... 第2層：ビジネスロジック層

第3層：データベース層

Webサーバサイド・コンピューティングの進化とWebサービス



Webサーバサイド・コンピューティングの性能向上策

IRI
Ubilabs Innovations

- 性能評価指標: **Webサーバ群の性能を定量的に見積もる数理科学的手法**

  ① スループット
  ② アクセス数
  ③ 接続数

- 負荷分散方式: **サービス実現のWebサーバ群の構成方式**

  ① ラウンドロビン
  ② ロードバランサー

帯域保証ネットワーク

端末AとBとの間に電話番号を発呼して交換機が動作し通信中は回線が接続状態になる　⇒「回線交換」:帯域保証ネットワークの原型

＊コネクション型ネットワーク

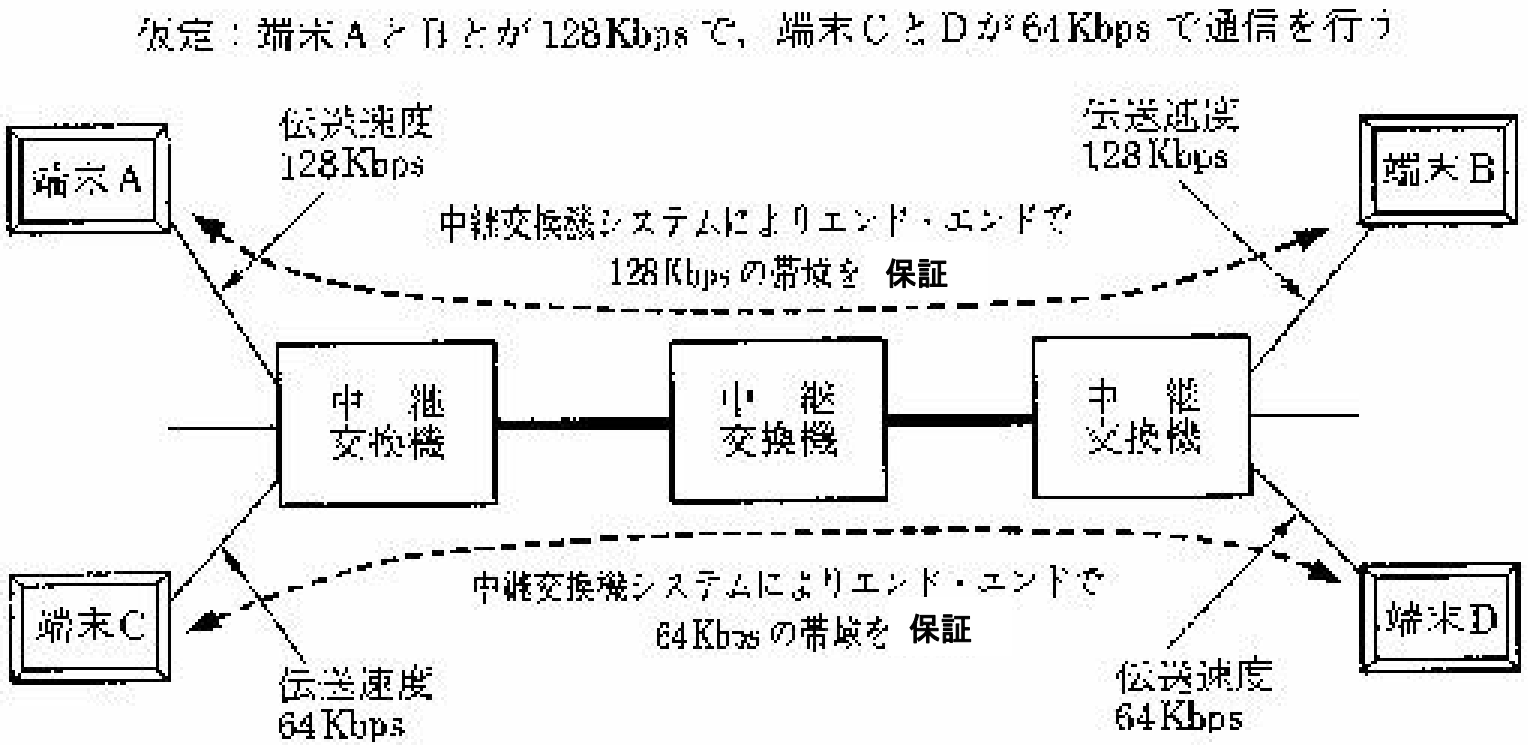


図 1.1　帯域保証ネットワーク

パケット交換によるﾍﾞストエフォート・ネットワーク

端末AからBへはひとかたまりのデータ（パケット）の先頭に宛先アドレスと送元アドレスを付加して送受信する。パケット交換機の蓄積交換機能によって正しく宛先に届く　⇒「パケット交換」:　ベストエフォート・ネットワークの原型

＊ｺﾈｸｼｮﾝﾚｽ型が基本

＊ｺﾈｸｼｮﾝ型もある

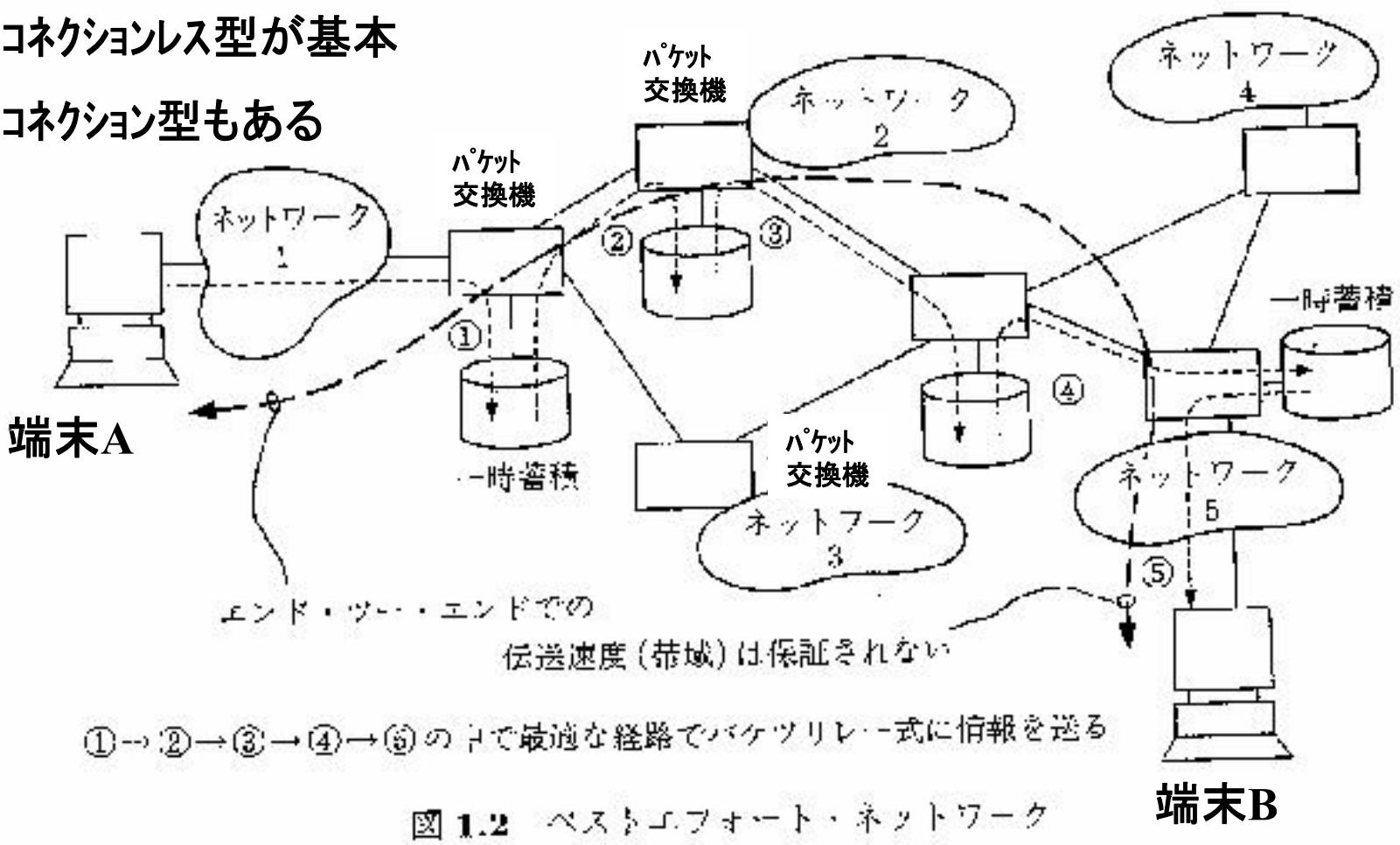


図 1.2　ベストエフォート・ネットワーク

ネットワーク伝送速度と圧縮の関係

表 1.2　ネットワーク伝送速度と画像・音声情報の伝送速度との関係

	品質	信号帯域 (kHz)	サンプリング周波数 (kHz)	間引き後ビット・レート (bps)	圧縮後ビット・レート (bps)
オーディオ	電話音声	3.4	8	64k	8.5k
	AM 放送	7	16	130k	24k
	FM 放送	7(ステレオ) または 14(モノラル)	16 または 32	510k	56k
		10(ステレオ) または 20(モノラル)	22.05 または 44.1	700k	64k
	音楽 CD	20(ステレオ)	44.1	1.4M	112k～224k

	品質	空間解像度 (縦×横)	フレーム・レート/秒	間引き後ビット・レート (bps)	圧縮後ビット・レート (bps)
ビデオ	ビデオ・クリップ	80 × 60	1	55k	1.8～2.8k
			3	165k	6～9k
			10	550k	18～28k
	1/4 画面	160 × 120	3	1.1M	20～33k
			10	2.2M	70～110k
			30	6.7M	220～335k
	VTR (VHS)	360 × 240	10	10M	330～500k
			30	30M	1～1.5M
	テレビ放送	720 × 480	30	120M	4～6M

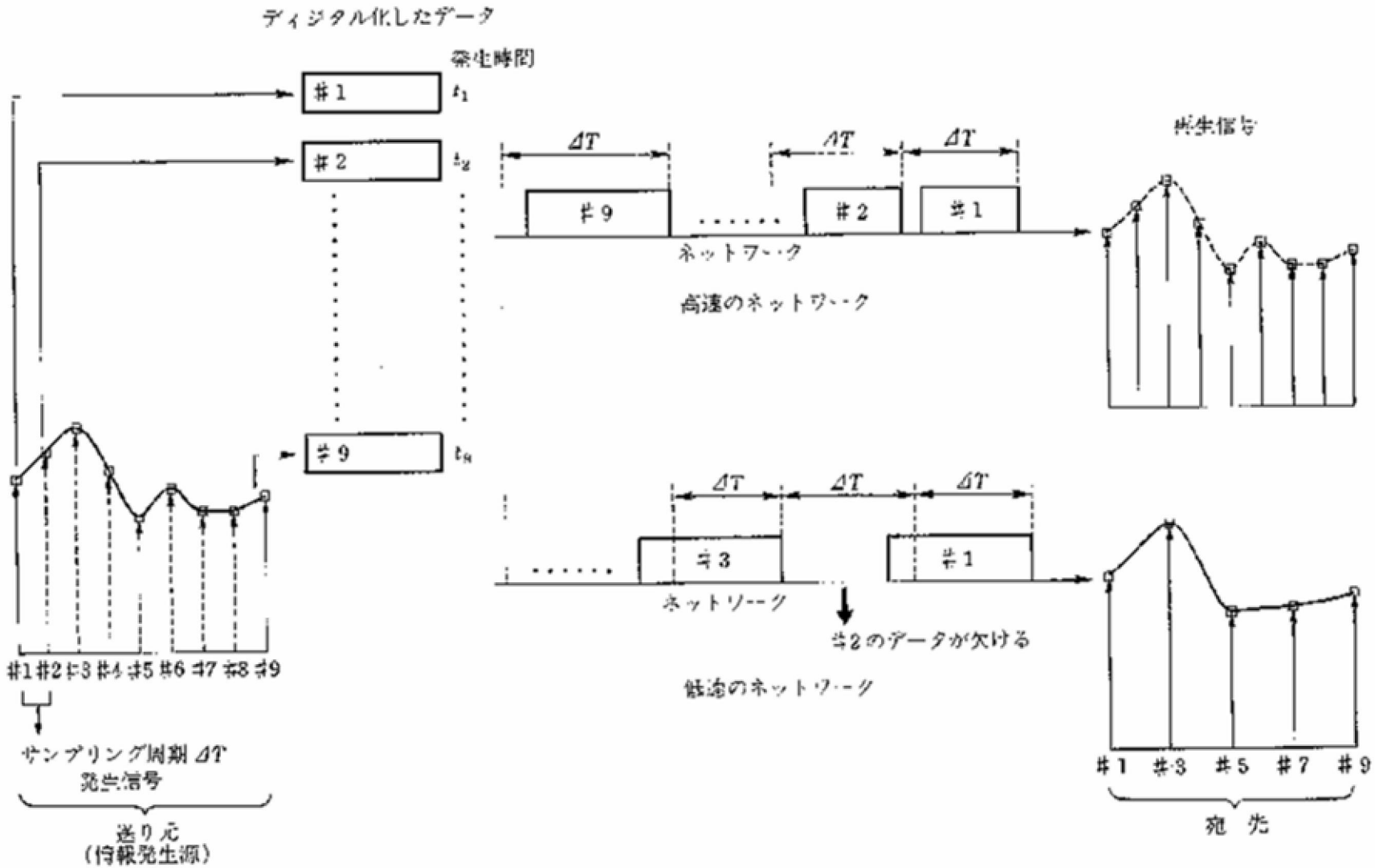


図 1.3 画像・音声情報の発生速度と伝送時間

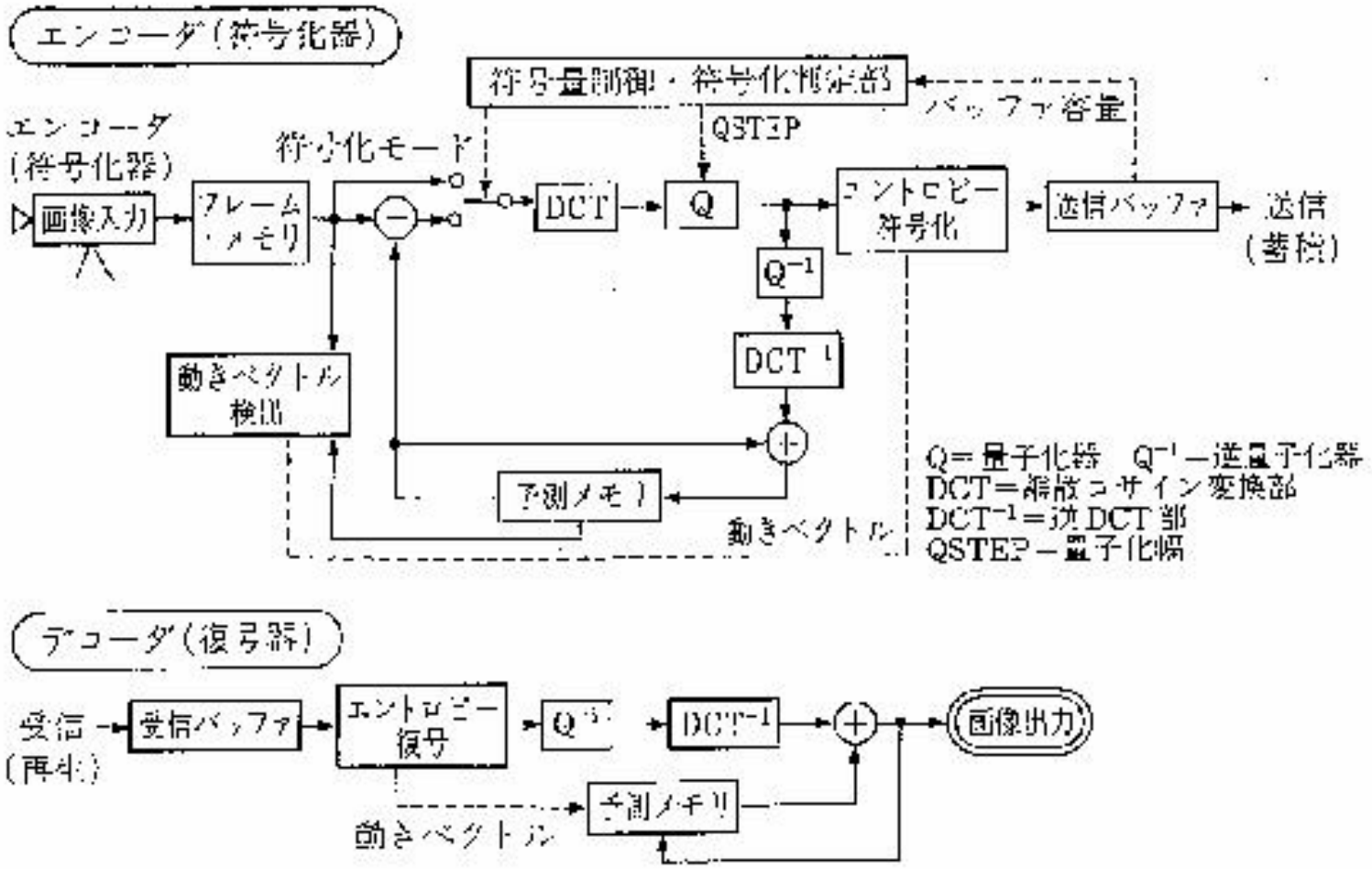


図 1.5 ハイブリッド動画像情報圧縮処理モデル

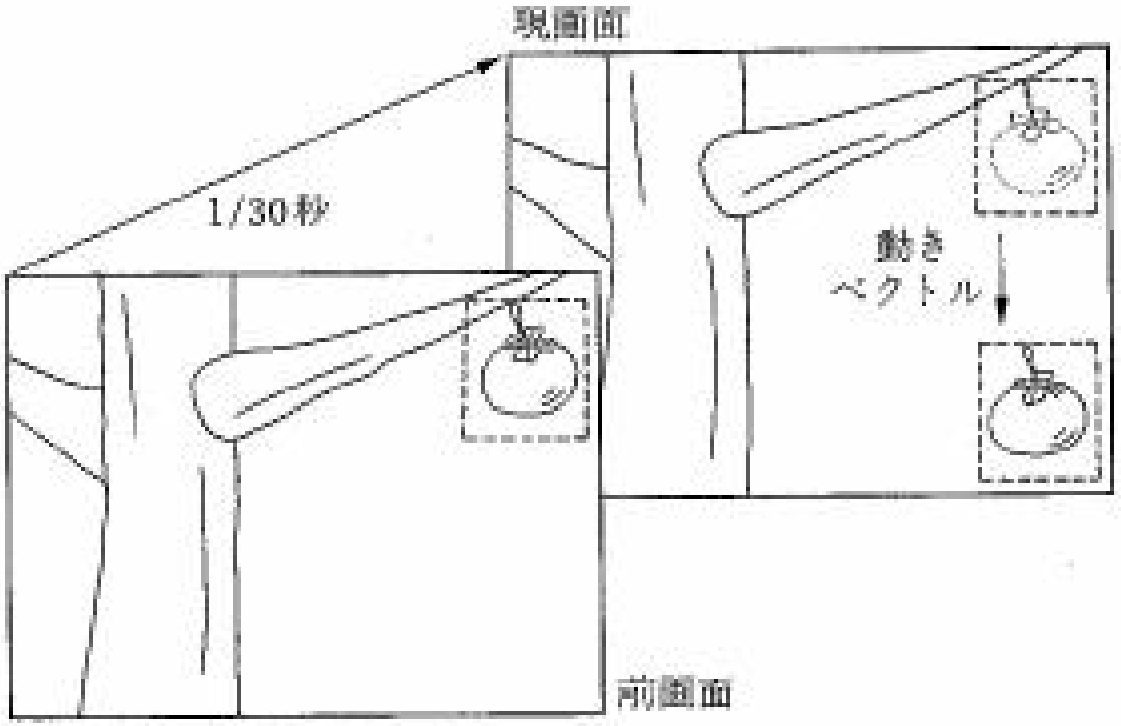


図 1.6　動き補償の原理

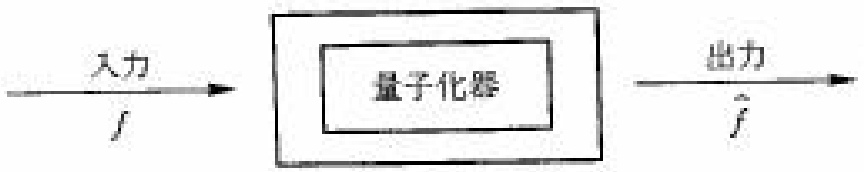
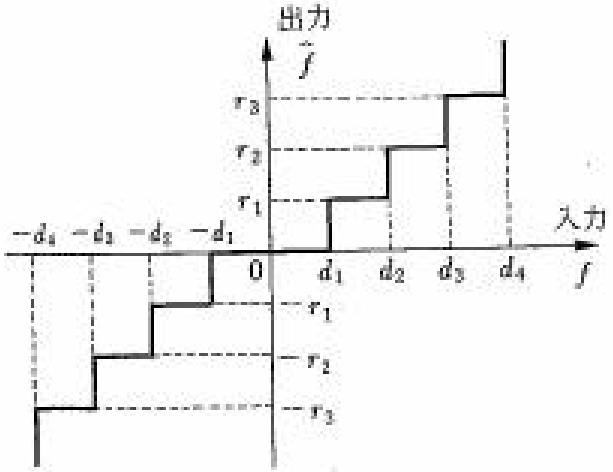
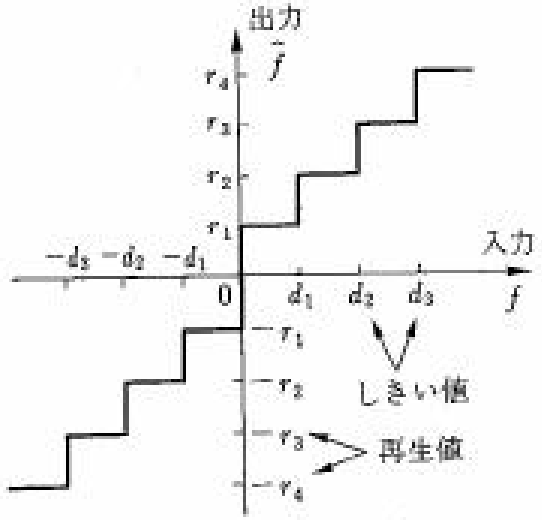


図 1.7　量子化器



(a) ミッドトレッダ型



(b) ミッドライザ型

図 1.8　一様量子化器での入出力関係

量子化器設計における設計手法としてまず重要なのは，量子化誤差を抑えることである．これには，主として式 (1.1) から (1.4) に示す 4 つの誤差表現が用いられている．

① 平均二乗量子化誤差（MSQE： Mean Square Quantization Error）

$$E[(f-\hat{f})^2] = \int_{a_L}^{a_U} (f-\hat{f})^2 \, p(f) df \qquad (1.1)$$

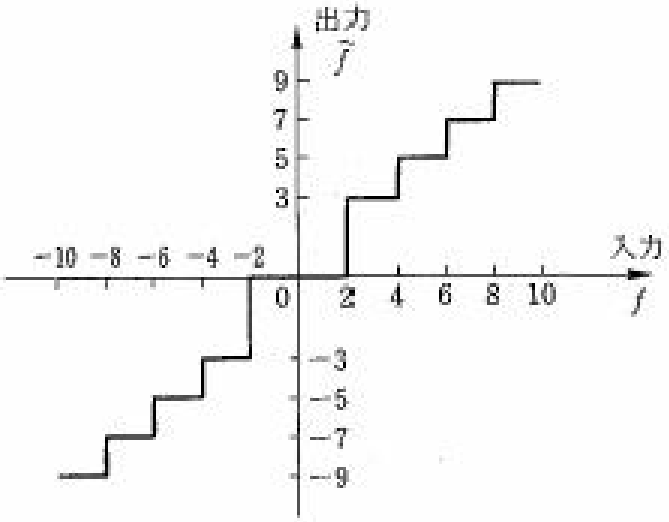


図 1.9　準一様量子化器での入出力関係

② 平均絶対値量子化誤差（MAQE： Mean Absolute Quantization Error）

$$E[|f - \hat{f}|] = \int_{a_L}^{a_U} |f - \hat{f}|\, p(f) df \tag{1.2}$$

③ 平均 Ln 正規量子化誤差

$$E[|f - \hat{f}|^N] = \int_{a_L}^{a_U} |f - \hat{f}|^N\, p(f) df \tag{1.3}$$

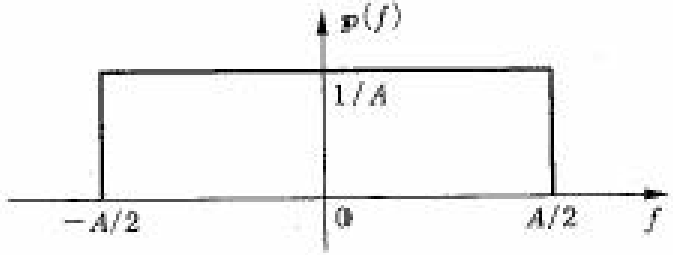
④ 重みづけ量子化誤差

$$\int_{a_L}^{a_U} w(f)|f - \hat{f}|\, p(f) df \tag{1.4}$$

以上の量子化誤差を一様量子化器の例について評価すると以下のようになる．

まず，一様量子化器の入力信号の入力信号に対する確率密度関数を図 1.10 のように仮定すると式 (1.5)～(1.7) が成立する．

$$p(f) = \frac{1}{a_U - a_L} = \frac{1}{A} \tag{1.5}$$



1.10　一様確率密度関数

$$r_k = \frac{\int_{d_k}^{d_{k+1}} f\, p(f) df}{\int_{d_k}^{d_{k+1}} p(f) df} = \frac{1}{2}(d_{k+1} + d_k) \tag{1.6}$$

$$d_{(k)} = \frac{r_k + r_{k-1}}{2} = \frac{d_{k+1} + d_{k-1}}{2} \tag{1.7}$$

したがって，

$$d_k - d_{k-1} = d_{k+1} - d_k = \text{ステップサイズ一定} = \frac{a_U - a_L}{J} \tag{1.8}$$

1.2.3　変換符号化

変換符号化は，信号領域をある関数によって変換し，別の領域へ写像することである．一般には，直交関数を用いるため可逆である．直交変換では変換後もエネルギーが保存され，逆変換による完全再生が可能である．変換は，一次元でも多次元でもあり得るが，多次元については一次元への分解が可能な変換が実装上有利であるため，可分多次元変換を情報圧縮処理の対象とする．

情報圧縮に適用可能な直交変換は，1970 年代初頭から今日まで実に多くの変換が考えられてきたが，実質的に有効なのは DCT である．このため，ここでは，DCT を中心に考え方を示す．DCT は，現在最も広く情報圧縮アルゴリズムとして採用されているものである．

(1) DCT の基礎

DCT の議論に入る前に，もとになる KLT（Karhunen-Loeve 変換）についてふれておく．KLT（式 (1.12)）は，カルーネン・レーベ変換と呼ばれ，定常確率過程において理論的に最適変換である．最適の意味は，無相関化，最大エネルギー寄与率，および最大符号化利得の 3 つである．しかしながら，KLT は理論的には最適だが，信号に依存する変換であり高速アルゴリズムが存在しないことから非実用的である．

DCT は，現在知られている KLT に最も近い特性をもつ極めて実用的な変換である．画像符号化を例にとると，8×8 の正方画素ブロック単位に 64 個の画素について DCT 演算を実行する．DCT の系統は，以下の式 (1.13)~(1.18) に示す 4 系統がある．

$$
\begin{aligned}
\left[C_{N+1}^{\mathrm{I}}\right] &= \sqrt{\frac{2}{N}}\left[K_m K_n \cos\left(\frac{mn\pi}{N}\right)\right],\ m,n=0,1,\cdots,N \\
\left[C_{N}^{\mathrm{II}}\right] &= \sqrt{\frac{2}{N}}\left[K_m \cos\left(m\left(n+\frac{1}{2}\right)\frac{\pi}{N}\right)\right],\ m,n=0,1,\cdots,N-1 \\
\left[C_{N}^{\mathrm{III}}\right] &= \sqrt{\frac{2}{N}}\left[K_n \cos\left(\left(m+\frac{1}{2}\right)\frac{n\pi}{N}\right)\right],\ m,n=0,1,\cdots,N-1 \\
\left[C_{N}^{\mathrm{IV}}\right] &= \sqrt{\frac{2}{N}}\left[\cos\left(\left(m+\frac{1}{2}\right)\left(n+\frac{1}{2}\right)\frac{\pi}{N}\right)\right],m,n=0,1,\cdots,N-1
\end{aligned}
\tag{1.13}
$$

ここで，

$$
K_j = \begin{cases} 1, & j \neq \text{or } N \\ 1/\sqrt{2}, & j \neq \text{or } N \end{cases}
$$

である．DCT の同一性は次のように示される．

$$\left[C_{N+1}^{\mathrm{I}}\right]^{-1} = \left[C_{N+1}^{\mathrm{I}}\right], \text{ 順変換=逆変換}$$

$$\left[C_N^{\mathrm{II}}\right]^{-1} = \left[C_N^{\mathrm{III}}\right] = \left[C_N^{\mathrm{II}}\right]^{-T} \qquad (1.14$$

$$\left[C_N^{\mathrm{III}}\right]^{-1} = \left[C_N^{\mathrm{II}}\right] = \left[C_N^{\mathrm{III}}\right]^{-T}$$

$$\left[C_N^{\mathrm{IV}}\right]^{-1} = \left[C_N^{\mathrm{IV}}\right], \text{ 順変換=逆変換}$$

DCT の順変換および逆変換の系列は以下のように定義できる．

DCT-I

$$X^{c1}(m) = \sqrt{\frac{2}{N}} K_m \sum_{n=0}^{N} K_n x(n) \left[\cos\left(\frac{mn\pi}{N}\right)\right], \quad m = 0, 1, \cdots, N \tag{1.15}$$

IDCT-I

$$x(n) = \sqrt{\frac{2}{N}} K_n \sum_{m=0}^{N} K_m X^{c1}(m) \cos\left(\frac{mn\pi}{N}\right), \quad m = 0, 1, \cdots, N$$

DCT-II

$$X^{c2}(m) = \sqrt{\frac{2}{N}} K_m \sum_{n=0}^{N-1} x(n) \cos\left[\frac{m(2n+1)\pi}{2N}\right], \quad m = 0, 1, \cdots, N-1 \tag{1.16}$$

IDCT-II

$$x(n) = \sqrt{\frac{2}{N}} \sum_{m=0}^{N-1} K_m X^{c2}(m) \cos\left[\frac{m(2n+1)\pi}{2N}\right], \quad m = 0, 1, \cdots, N-1$$

DCT-III

$$X^{c3}(m) = \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} K_n x(n) \cos\left[\frac{n(2m+1)\pi}{2N}\right], \quad (1.17)$$
$$m = 0, 1, \cdots, N-1$$

IDCT-III

$$x(n) = \sqrt{\frac{2}{N}} \sum_{m=0}^{N-1} K_n X^{c3}(m) \cos\left[\frac{n(2m+1)\pi}{2N}\right],$$
$$m = 0, 1, \cdots, N-1$$

DCT-IV

$$X^{c4}(m) = \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} x(n) \cos\left[\frac{(2n+1)(2m+1)\pi}{4N}\right], \quad (1.18)$$
$$m = 0, 1, \cdots, N-1$$

IDCT-IV

$$x(n) = \sqrt{\frac{2}{N}} \sum_{m=0}^{N-1} X^{c4}(m) \cos\left[\frac{(2n+1)(2m+1)\pi}{4N}\right],$$
$$m = 0, 1, \cdots, N-1$$

この中で，最も多用されているのが DCT-II である．これは，1974 年に K.R.Rao らによって開発されたもので，MPEG など多くの国際標準で用いられている．I から IV の型は，いずれも分解可能で高速アルゴリズムが見つかっている．ここで，この DCT-II を前提に議論を進めると，実際には多点 DCT 演算が用いられる．

このようにして得られるDCTの特性を理解するために，図1.12に二次元8×8DCTの基底画像を示す．

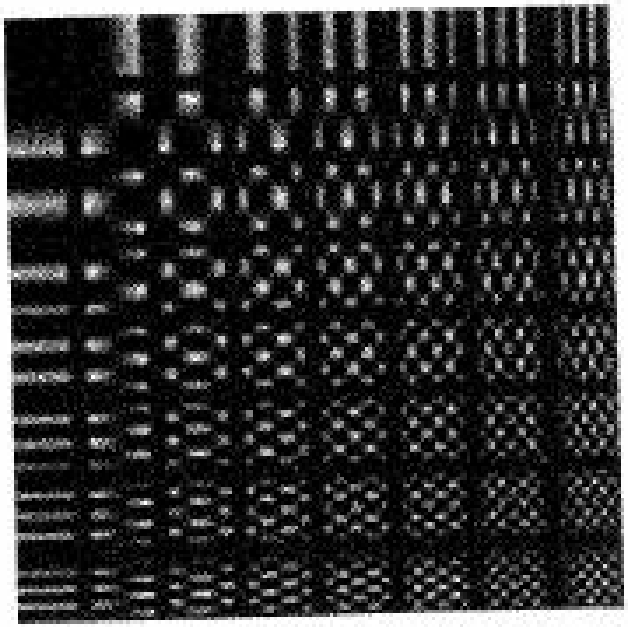


図 1.12　二次元 (8×8) DCT の基底画像

1.2.7　動き補償

動き補償とは，動画像圧縮だけに有効な手法である．これは，1画面中にある対象物体が異なる時刻にどの位置にあるかを探索し，移動量を求めることで情報量を圧縮する考え方である．アイデアとしては，画素ごとに追跡するPRA（Pel-Recursive Algorithm，画素漸化型アルゴリズム）と画素ブロックごとに追跡するBMA（Block Matching Algorithm，ブロック・マッチング・アルゴリズム）とが知られているが，PRAは実用的でないためもっぱらBMA（図1.6参照）が用いられている．

図1.16は，現フレームの $M \times N$ 画素ブロックが，前フレーム中の探索範囲（サーチウィンドウ）内で，探索する様子を示している．

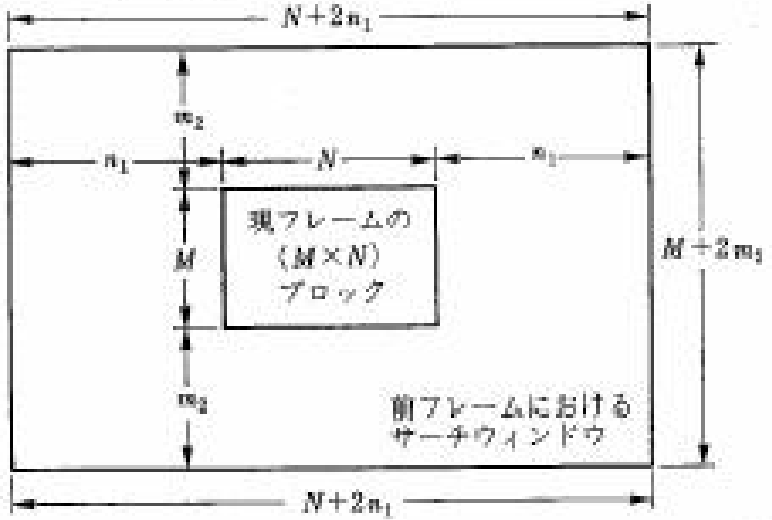


図1.16　現フレームの $M \times N$ 画素ブロックと前フレーム中の探索範囲の関係

この対象ブロックの画素値と比較ブロックの画素値間の一致度を測る評価尺度としては，以下に示す3つが主として用いられている．各々一長一短があるが，実装の容易なMAEが多く用いられている．

① 平均二乗誤差 (MSE)

$$M_1(i,j) = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}(X_{m,n} - X^R_{m+i,n+j})^2, \qquad (1.52)$$

$$|i| \le m_2,\ |j| \le n_2$$

② 平均絶対値誤差 (MAE)

$$M_2(i,j) = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}|X_{m,n} - X^R_{m+i,n+j}|, \qquad (1.53)$$

$$|i| \le m_2,\ |j| \le n_2$$

③ 相互相関関数

$$M_3(i,j) = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N} X_{m,n}X^R_{m+i,n+j}}{\left[\sum_{m=1}^{M}\sum_{n=1}^{N} X^2_{m,n}\right]^{1/2}\left[\sum_{m=1}^{M}\sum_{n=1}^{N}(X^R_{m+i,n+j})^2\right]^{1/2}}, \qquad (1.54)$$

$$|i| \le m_2,\ |j| \le n_2$$

1.2.8　エントロピー符号化

エントロピー符号化は，これまで述べてきた様々な情報圧縮符号化の最終段階で適用するのが常である．これは，前にも述べたように符号化データの発生確率の偏りを利用した可変長符号化である．例えば，動きベクトルの移動量は，小さい値をとる確率が高い．このため，符号化する際には近傍での動きベクトルにより短い符号を，遠隔の動きベクトルにはより長い符号を割り当てることで，全体の情報量を抑えることができる．このような発生確率を基本とした可変長符号化をエントロピー符号化と呼んでいる．

実際例として Huffman 符号と二次元 Huffman 符号とについて示す．

(1) Huffman 符号

以下に示すように 4 値をとる例では，固定長 2 進データでは 2 ビット長となるが，可変長 Huffman 符号では 3 ビット長となる．しかし，ビット長の短い 0 の発生確率が極端に高いと統計的なデータ長は，2 ビット以下にすることができる．

元データ （固定長） ： 10 進数で 0 から 3 とする
2 進データ （固定長） ： 2 進数で 00，01，10，11 となる
Huffman 符号（可変長）： 発生確率順位 1 位： 0 → 0 （1 ビット）
2 位： 1 → 10 （2 ビット）
3 位： 2 → 110 （3 ビット）
4 位： 3 → 111 （3 ビット）

(2) 二次元 Huffman 符号

上記の (1) に加えてランレングス符号化を併用するものである．

この場合は，ゼロラン（連続するゼロの個数）とその次に現れる値のマトリクスに対して Huffman 符号を割り当てるため，二次元 Huffman 符号と呼ばれる．例を表 1.3 に示す．

表 1.3　二次元 Huffman 符号

元データ値		0	1	2	3
ゼロラン	0	—	00	110	1110
	1	—	010	1010	01111
	2	—	0110	01110	111101
	3	110	1011	11110	111100

10 進 元データ　：0 2 0 1 0 0 0 1 1 0 0 0 0 3

2 進 元データ　：00 10 00 01 00 00 00 01 01 00 00 00 00 11 = 28 ビット

Huffman 符号　：0 110 0 10 0 0 0 10 10 0 0 0 0 111 = 21 ビット

二次元 Huffman：1010 010 1011 00 110 1110 = 20 ビット

以上の例においてわかるように 2 進元データに対して Huffman 符号と二次元 Huffman 符号で符号化したデータは，全体のデータ長を圧縮することができる．

(3) Huffman 符号と二次元 Huffman 符号で符号化したデータの生成法

以下のような手順で符号化する．

① Huffman 符号の場合は，符号化対象とする元データの，また二次元 Huffman 符号の場合は，マトリクスの各値の発生確率を求める．

② 対象値の数が奇数の場合は，発生確率の最低のものに 0 か 1 を割り当てる．偶数の場合は，最低の 2 つを組み合わせて 0 と 1 を割り当てる．

③ 上記のペアと次に低確率の符号と組み合わせて 0 と 1 を割り当てる．

④ 順次組み合わせて符号化の木を生成する．

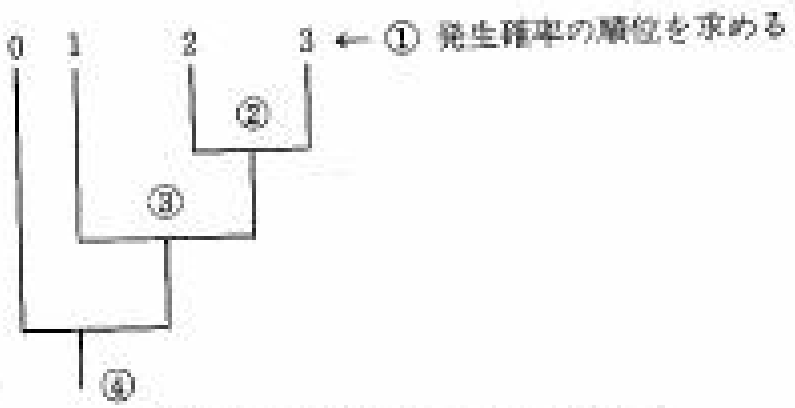


図 1.17　Huffman 符号の生成法

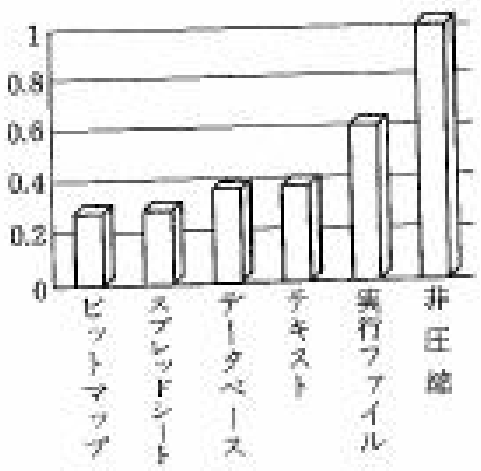


図 1.21　Addstor 社による実験例

現在，インターネット上で多くのデータ交換が行われるようになってきたが，これらは上記の基本原理に基いて，OS 種別に応じて多様なファイル圧縮形式が用いられている（表 1.5 参照）．

表 1.5　OS 種別に応じて多様なファイル圧縮形式

OS 種別	圧縮形式（拡張子）
UNIX	compress(Z) 他
Windows	Zip (zip，32 ビット)，LZH (lzh，16 ビット) 他
Macintosh	StuffIt (sit,sea)，Compact Pro (cpt) 他

6. 情報検索のための数理科学

情報検索の概要

インターネットを介した情報検索の基本技術

①データの管理および入出力のためのデータベース

②文書データ処理のための自然言語処理や計算言語学

③画像や音声を扱うためのディジテル信号処理と認知心理学を背景とするパターン認識技術

④メタデータの考え方の基本となる図書館情報学

⑤検索アルゴリズム設計

⑥情報検索システムの評価尺度理論

情報検索の過去・現在・未来

●「情報検索」に関する研究が、開始されたのは、大規模に蓄積される学術文献や論文等の管理をコンピュータ上で行うために、大規模の図書館でデータの管理と検索が行われるようになった1970年代のことである 。

●インターネットが登場した1990年代からは、Yahoo!やGoogleのようなWorld Wide Web上のデータを対象にした検索エンジンが登場し、現在では一般生活にまで普及し、これを支える数理科学の役割が大きくなっている。
　今後の情報検索の課題は、以下のようなものがある。
①Deep Web（ショッピングサイト等大規模なデータベースが動的Webサイト）を対象にした検索
②より直感的で操作性に優れたユーザインタフェース
③より人間的なマルチメディア情報検索
④さまざまなメディアを統合的かつ横断的に扱うクロスメディア情報検索
⑤格納されるデータや検索入力が言語に依存しないマルチリンガル(クロスリンガル)検索
⑥P2Pネットワーク等の大規模分散データを対象にした情報検索

情報検索のシステム要素

情報検索のシステム要素を以下のようになる。

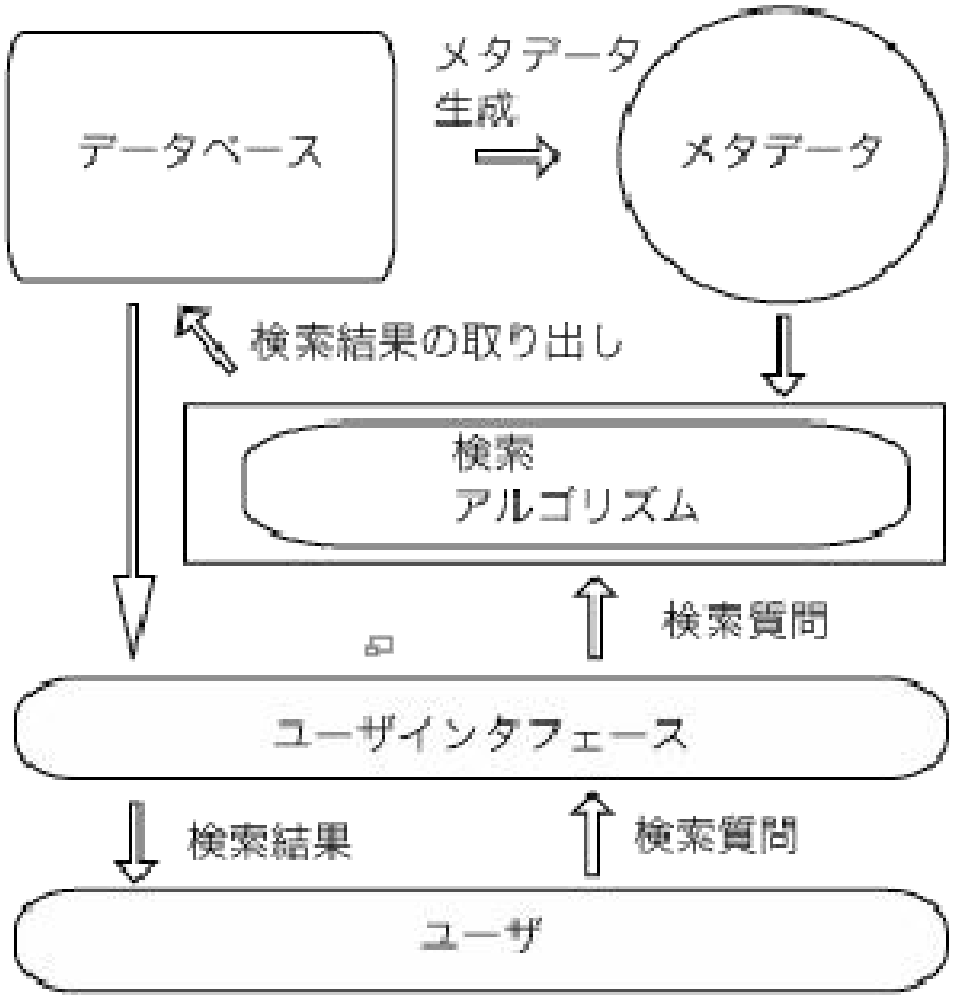
①検索対象のデータ

②データベース（検索対象データの蓄積・管理）

③索引語としてのメタデータ

④ユーザインタフェース

⑤検索アルゴリズム



情報検索システムの構築法

情報検索システムの構築は以下の各フェーズによって実行する。

①検索対象データの収集

検索対象データの収集方針の決定が重要。
*World Wide Web上のハイパーテキストを収集して対象とする場合にはクローラ(ロボット、スパイダー等)を用いて自動収集*するのが一般的。
Web上には、膨大なデータが存在し、データ自身が急激に変化するため、網羅的に収集することは困難。
そのため、いかにして多くの対象のデータを収集するかが　重要課題となっており、*World Wide Web検索エンジンのサービスでは何ページのデータか検索が可能であるかが重要な性能指標*となっている。

②検索対象データからメタデータを作成

メタデータの形式および作成方法は、データベースの構造、検索アルゴリズム、およびデータ収集方針との関連性深い。
データ収集を広範に継続的に行うような場合、人海戦術によるメタデータ作成は、コストの大幅増大を招く。

③検索アルゴリズムの設計

作成した*メタデータを用いてどのような計算*によって、データを出力するか決定する。

④検索性能の評価

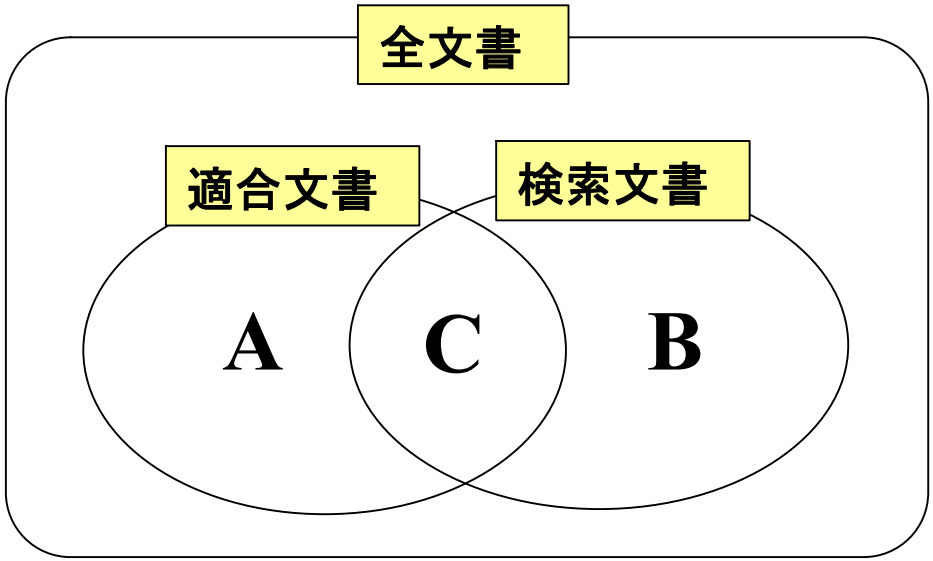
情報検索システムの検索性能の評価を行う。情報検索システムの検索性能は主に正確性と網羅性の質的な観点から適合率(precision;精度ともいう)と再現率(recall)を、処理性能の量的な観点からスループットを測定することにより判定するのが一般的。

ⅰ)適合率　P(Precision)
検索結果として得られた集合中にどれだけ検索に適合した文書を含んでいるかという正確性の指標　　検索ノイズの少なさの指標！
【適合率:P＝C/B(C:検索された適合文書数、*B*:検索結果の文書数)】

ⅱ)再現率　R(Recall)
検索対象としている文書の中で検索結果として適合文書のうちでどれだけの文書を検索できているかという網羅性の指標　検索漏れの少なさの指標！
【再現率:R＝C/A(C:検索された適合文書数、A:全対象文書中の適合文書数)】

再現率と適合率

適合率(P)をあげれば再現率(R)が下がり、再現率を上げれば適合率が下がる傾向にあるため、F値(F-measure)、E値(E-measure)という尺度も用いられる。



＊F値＝2/(1/R+1/P)によって求められる。 F値(RとPとの調和平均:逆数の平均値の逆数)が大ききほど、性能が良いことを意味。再現率と適合率双方の値(0と1の間)が大きい時に大きくなる。

＊E値＝1-(1+b^2)/(b^2/R+1/P)によって求められる。bはパラメータで再現率と適合率のどちらを重視するかを決定。b>1、b<1の時、それぞれ、適合率、再現率を重視する。小さいほど性能が良いことを意味。

情報検索技術の分類

IRI
Ubilabs Innovations

1．検索対象データの抽象度

●直接検索

メタデータを介さずデータそのものを直接計算アルゴリズム上で処理する検索方法。例としてハミングによる検索の入力を行い類似する音程の音楽を検索するもの等。実用上は、前処理としての索引の生成を事前におこなう方式も多いが、このような場合もデータに含まれる表現をそのまま用いて検索を行うため検索モデルとしては直接検索に分類される。

＊全文検索

直接検索の一種で、文書データの全文自動走査により、メタデータを作成・保管し、検索の入力に合致するデータを検索結果とする方法。

●間接検索

データベースに蓄積されたデータからメタデータを生成・保管し、検索入力時に、内部表現に変換された検索入力と保管されたメタデータを比較することで検索結果を生成する方法。

情報検索技術の分類

2. 検索入力の種類

●単語（キーワード）

単語（キーワード）を指定することによって検索を行う。最も単純な形式。

●検索言語

システム特有の検索言語を用いて検索を行う方法。論理和・論理積などのブーリアンモデルの演算を検索の絞り込みに用いる。研究者や法律・医学等の専門的な実務家など、特定の分野の専門家を対象にした検索システムなどに用いられる。SQLのようなデータベース・マネージメント・システムで標準的言語を用いることもあるが、特定分野の検索エンジン特有の検索言語を用いているシステムも多い。実現例としてはIEEE Xploreなどがある。

●直接入力

検索のパラメータとなる関連するデータを直接入力する方法。画像入力による類似画像を検索や、音楽クリップ検索などが研究されている。

●自然文

検索に関わるユーザインタフェースの研究として古くから研究されてきた。
近年ではGoo ラボによって開発された「日本語自然文検索」が話題。

情報検索技術の分類

IRI
Ubilabs Innovations

3．検索アルゴリズム（問題を解くための手順、算法）
　一般に情報検索システムの構築時にはメタデータ生成時に索引を同時に作成し検索アルゴリズムによる検索結果の評価の際に索引を用いた最適化を行う。

①*パターンマッチング*

クエリ（検索質問）として入力された表現をそのまま含む文書を検索する。単純にパターンのみを探すのではなく、活用形の変化による同義語のパターンの不一致を解消した検索を行う拡張等が行われる。パターンマッチングの詳細なアルゴリズムについては文字列探索による。

②*ブーリアンモデル*

パターンマッチングに加え、メタデータの属性ごとの絞り込み条件を論理和・論理積などによって組み合わせて併用する方法

情報検索技術の分類

IRI
Ubilabs Innovations

③*ベクトル空間モデル*

キーワード等を各次元とした高次元ベクトル空間を想定し、検索対象データやユーザによるクエリに何らかの加工を行いベクトルを生成。ベクトル空間上に検索対象となるベクトルを配置し、ベクトル化された検索質問とデータのベクトルの相関量（ベクトル間のコサイン、内積、ユークリッド距離など）によって検索の対象のデータと検索質問の関係性の強弱を計算する。

＊参考論文電子情報学会誌

言語表現のベクトル空間モデルにおける最適な計量距離

持橋　大地[†,††]　　菊井玄一郎[†]　　北　　研二[†,†††]

＊ベクトル空間モデルでの単語文書行列

単語文書行列とはメタデータの生成・表現法の一つであり、ベクトル空間モデルによる検索を行う際に非常に頻繁に用いられるメタデータの形式である。一般に単語文書行列は以下に示す構造を持つ。
単語文書行列:

$$\mathcal{M} = \begin{pmatrix} & d_1 & d_2 & d_3 \\ t_1 & 0 & 2 & 1 \\ t_2 & 1 & 1 & 2 \\ t_3 & 0 & 0 & 3 \end{pmatrix}$$

文書*di*に単語*tj*が*n*回出現するとき、*wij*を*n*とし、行列を形成する。単純に出現回数を利用する以外に様々なアルゴリズムによって得た重みを用いる生成方法が多用される。

情報検索技術の分類

④*潜在的意味索引付け*

ベクトル空間モデルの応用として考案された。*高次元ベクトル空間を行列として扱い特異値分解を行い、得られた直交低次元ベクトル空間上で検索*する。

単純なベクトル空間モデルでの検索に比べて、同義語が用いられている文書間の関連を反映し、検索の対象のデータの内容的な偏りに影響を受けにくい検索を行うことが可能。

＊特異値分解：線形代数における、複素数あるいは実数を成分とする行列分解の一手法。行列に対するスペクトル分解定理の一般化で、正方行列に限らず任意形式の行列分解が可能。

検索エンジン（Searching Engine）

インターネット上に存在する情報（ウェブページ、ウェブサイト、ニュースなど）を検索する機能の総称で、主として、Webサーバのソフトウェアとそれを支援するWebブラウザのソフトウで実現される。ロボット型検索エンジン、ディレクトリ型検索エンジン、メタ検索エンジンなどに分類される。インターネットの普及初期には、検索エンジンとしての機能のみを提供していたウェブサイト　そのものを検索エンジンと呼んだが、現在では様々なサービスが加わったポータルサイト化が進んでいる。

①ロボット型検索エンジン

IRI
Ubilabs Innovations

●所定の検索アルゴリズムに従って、Webページ等を検索するサーバ、システムのこと。検索アルゴリズムは、最も単純な場合はキーワードとなる文字列のみであるが、*複数のキーワードにANDやOR等のブーリアン論理式を組み合わせて指定*することが多い。

●*ロボット型検索エンジンの大きな特徴は、クローラ(スパイダー)を用いること*にある。このクローラ機能により、Web上にある多数の情報を効率よく収集可能のため、大規模検索エンジンでは、数十億ページ以上のページからの検索が可能である。

●ページ収集情報は、事前に解析し、索引情報(インデックス)を作成する。英語とは異なり、日本語などの言語では、自然言語処理機能によって生成される索引の性能が決まる。このため、*多言語対応検索エンジンが今後重要*となってくる。

①ロボット型検索エンジン

●検索結果の表示順は、検索エンジンにとって最も重要である。ユーザーが期待したページの検索結果の上位に表示することが重要であるため、*多くの検索エンジンが、表示順を決定するアルゴリズムを非公開にし、その性能が競争状態*となっている。

●検索エンジン最適化業者*（SEO）の存在が、アルゴリズムの非公開要因*である。

●*Googleは、アルゴリズムの一部としてPageRankを公開*しているが、多くの部分が非公開。

● *Webページの更新時刻情報によって新しい情報に限定して検索可能*ものや、検索結果をカテゴリ化して表示するものなど、特長のある機能を搭載しているものもある。

●Google, Yahoo!, infoseek, Technorati,MARSFLAG, Altavista, AlltheWeb, Teoma, WiseNut, Inktomiなどがロボット型検索エンジンを利用している。

②ディレクトリ型検索エンジン

IRI
Ubilabs Innovations

●人手で構築したWebディレクトリ内を検索する検索エンジン。

●人手で構築しているため、質の高いWebサイトを検索可能。

●サイトの概要を人手で記入しており、検索結果から目的のサイトを探し易い。

●人手入力のため、検索対象となるサイト数が少ない。

●WWWの爆発的な拡大から、全Webサイトを即時にディレクトリへ反映させることが困難になり、現在では非主流。

●ディレクトリ型検索エンジンでは、ヒットするサイトがない場合、 ロボット型検索エンジンを用いる併用型が多い。

●Yahoo!, Lycos, Open Directory Project, LookSmartなど。

③メタ検索エンジン

●入力された*キーワードを複数の検索エンジンに送信し、得られた結果を表示する検索エンジン*。

●メタサーチエンジン、横断検索エンジンとも呼ぶ。

●“meta”とはこの場合、“beyond”（横断）の意味。

●検索毎にエンジンを使用するかを選択する「非統合型」と、検索結果を独自のアルゴリズムで総合的に判断して一つの結果として出力する「統合型」とがある。

●統合型では結果表示に広告表示が出来ないため、*Google のようにメタ検索エンジンでの利用を禁止している*場合もある。

検索エンジンの歴史その1

IRI
Ubilabs Innovations

1995 年

アメリカでヤフーとアマゾンが設立。日本のインターネット元年。検索サービスは、登録型のディレクトリサービスが中心で、アメリカで 12 月Altavista が登場、フルテキストサーチが登場。

1996 年

日本でヤフーがサービスを開始した年だがインターネットの検索市場に動きが少ない。検索エンジンはインフォシークが登場。徐々にロボット型の検索エンジンが登場。一般人はヤフー、通は Altavista や infoseek 等の海外のロボット検索エンジンを使用。

1997 年

この年、goo がサービスを開始、同時に楽天もサービスを開始。検索のテクニックを競う「検索の鉄人」が開催。当時、検索結果に思うような結果を出すのは難しかった。鉄人大会の委員長は舛添要一氏。

検索エンジンの歴史その2

IRI Ubilabs Innovations

1998 年
スタンフォード大学の二人（ｻﾘｹﾞｲ・ﾌﾞﾘﾝとﾗﾘｰ・ﾍﾟｰｼﾞ）が Google を設立。日本で、ヤフー、MSN、インフォシーク、goo、エキサイト、ライコス、フレッシュアイが登場、90 年代後半のポータルサイトのメジャープレイヤーへ。ヤフージャパンgooと検索エンジン提携。

1999 年
2000 年問題。i モード開始。検索エンジン企業の買収活発化。

2000 年
Google が台頭。米ヤフーが Google を採用し、日本語サービスも開始。Google の登場でロボット検索の性能が飛躍的に向上。

検索エンジンの歴史その3

2001 年
日本市場でGoogleが台頭 。ヤフージャパンの検索は、goo から Google へ転換し、Google からのトラフィックが急増。ヤフー BB 開始でブロードバンド時代へ。

2002 年
Google の AdWords 広告と Overture の Pay for Performance が日本で開始。

2003 年
goo、infoseek が検索エンジンとして Google を採用。日本の検索エンジン消滅。SEO サービスが本格化。

検索エンジンの歴史その4

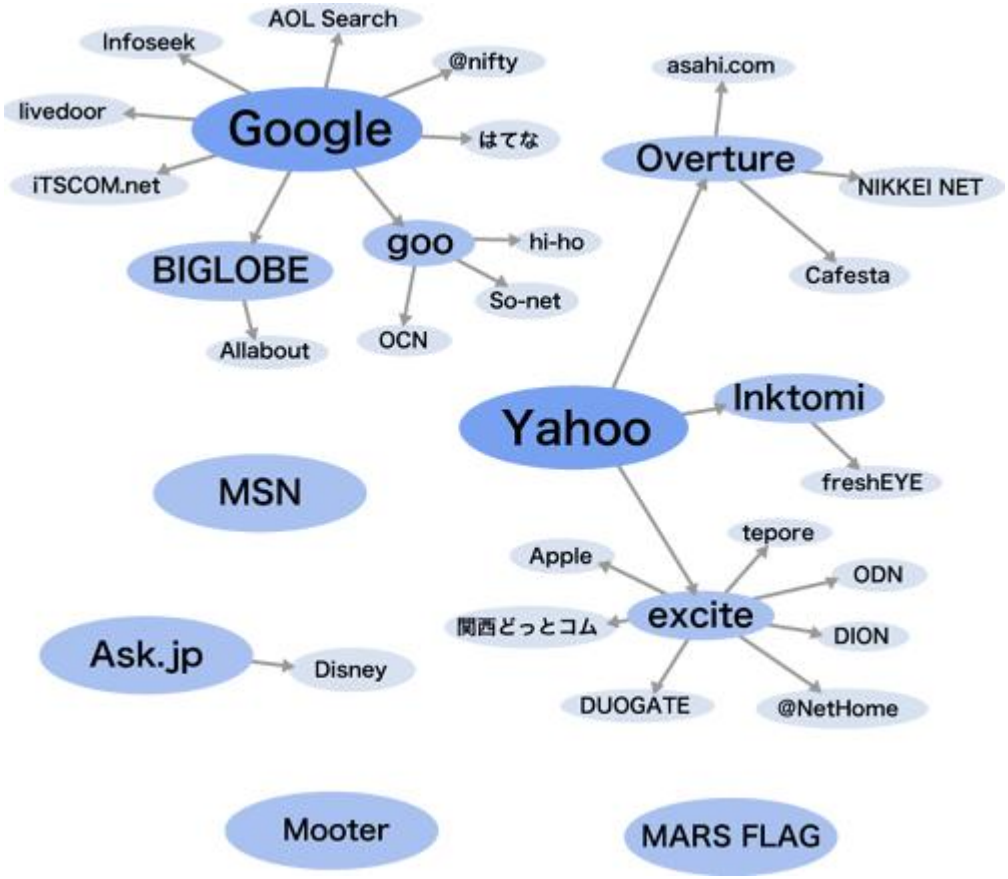
IRI Ubilabs Innovations

2004 年
Google株式市場に上場。マイクロソフトが新検索サービス発表。米国に続き、日本ではヤフーとGoogleの提携が解消。

2005 年
Yahoo!とMSNが独自の検索エンジンを採用。各検索エンジンが、多様なサービスを開始し、地図検索、パーソナライズ、デスクトップ検索などへ拡大。ブログ検索やモバイル検索エンジンなどの新サービス開始。モバイルでのインターネット利用者がパソコン利用を凌駕。

2006 年
不自然な外部リンクに対して、Googleが厳格処置。
携帯の検索サービス登場。NTTドコモ:複数エンジン、au:Google、ソフトバンク:Yahoo!。

日本語のロボット検索エンジン相関図



最終更新日 2006 年 8 月 21 日（調査ECジャパン）

使用言語からみたインターネット人口の割合

Internet Societyによればインターネットで用いられている言語のうち英語が占める割合は85%とされていたが、その後の進歩や各国のインターネットの普及により多言語化が進み、上表に見られるように2000年の年末には英語と非英語の言語人口が逆転し、その傾向は継続。

検索エンジンの新たな課題！

	1998年	1999年	2000年		2001年			2002年		2003年	2004年
	12月	1月	4 - 7月	12月	2月	4 - 6月	7月	1月	6 - 10月	2 - 4月	7月
英語	58%	55%	51.3%	49.6%	47.6%	47.5%	45.0%	43.0%	40.2%	36.5%	35.8%
非英語	42%	45%	48.7%	50.4%	52.4%	52.5%	55%	57.0%	59.8%	63.5%	64.2%

検索エンジンの新たな課題！

IRI
Ubilabs Innovations

●WWW検索エンジンの代表Googleでは100億を超えるWebページが登録。

●検索エンジンの利用者は、容易に検索することが困難に。

●日本語入力機能のないコンピュータを用いて日本語サイト検索は、困難に。

●非英語圏の言語間の検索は中間に翻訳エンジンがないと困難に。

●インターネットの多言語化が今後も増加し、言語間障壁の克服が課題。

⇒これらを解決するのが数理科学の役割！

7. 今後のインターネットにおける数理科学の展望

NGN登場の背景

【1】 IPの登場による固定電話/専用線事業の衰退

【2】 携帯電話事業の成長減速と数年後の衰退懸念

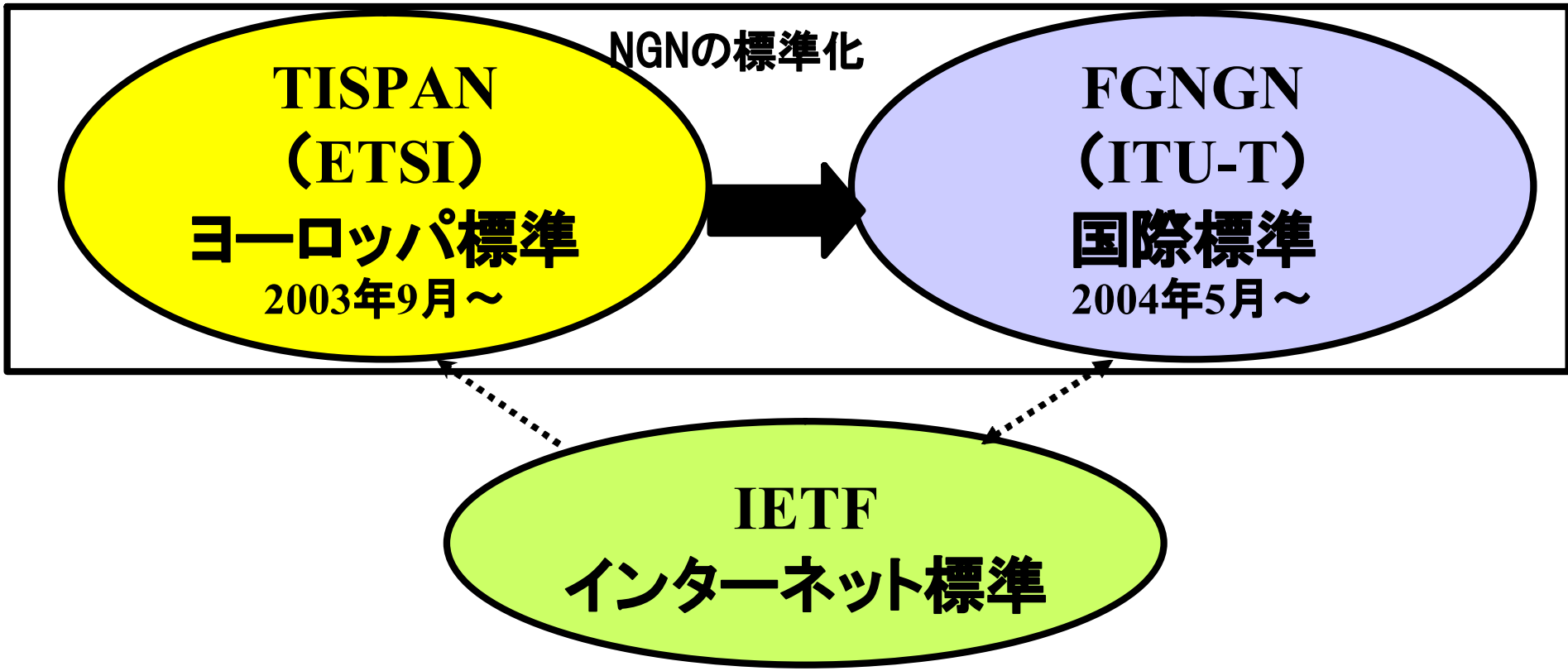
【3】 Skype型電話の爆発的普及

【4】 動画サービスの爆発的普及

【5】「技術革新」・「業界政治」・「国際政治」の複合的背景

⇒重要なのは「技術革新」への対応！

NGNの標準化団体の相関関係



ETSI（European Telecommunications Standards Institute

TISPAN（Telecommunications and Internet converged Services and Protocols For Advanced Networking）

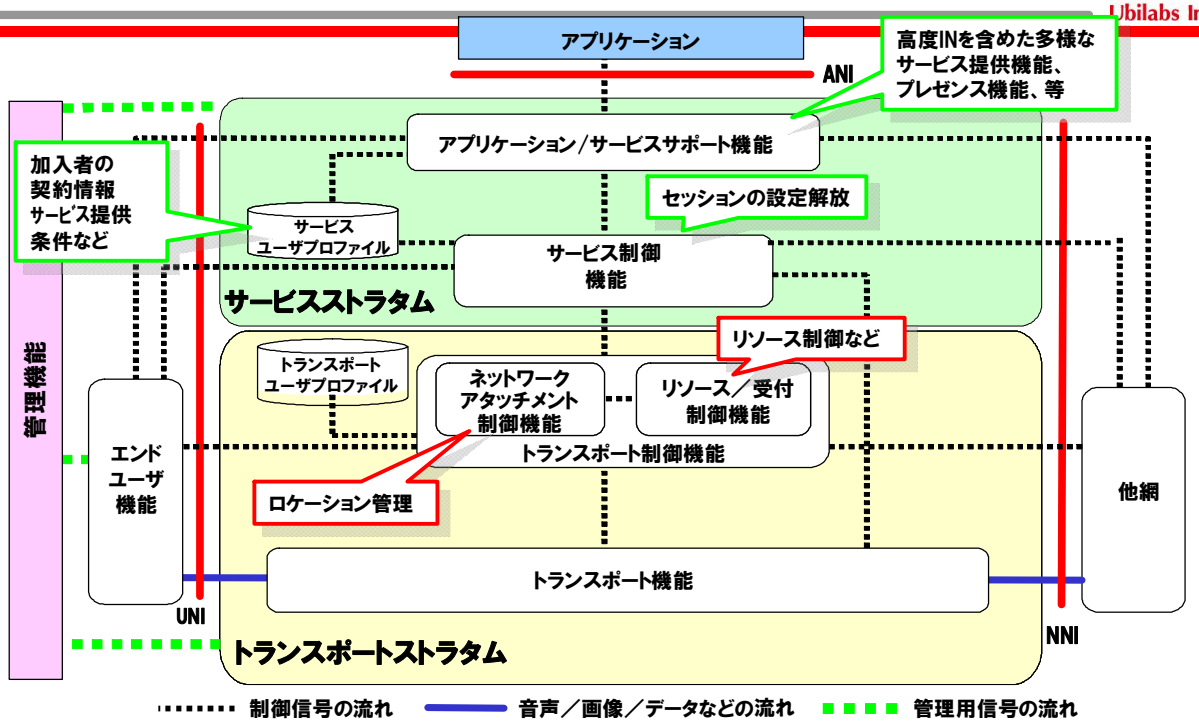
FGNGN（Focus Group Next Generation Network）

ITU-T（International Telecommunication Union-Telecommunication Standardization Sector

IETF（Internet Engineering Task Force ）

通信キャリアによる次世代ネットワーク構築の取り組み（標準化（ITU-T）の動向）

NGNのアーキテクチャ



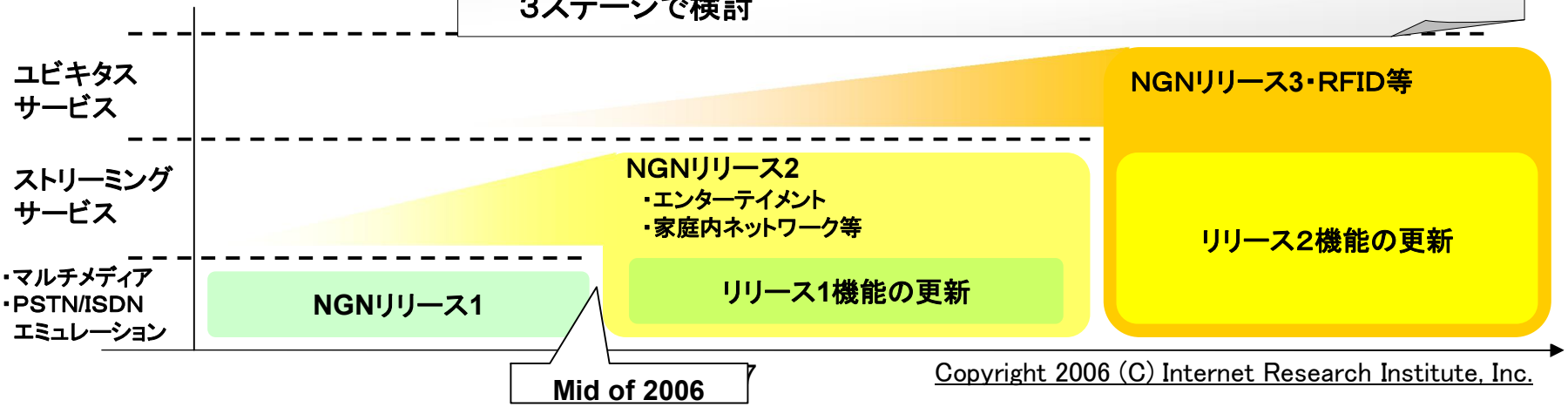
（注）

トランスポート層のトランスポート機能には、アクセス、コア、エッジ機能及びメディアハンドリング機能等が含まれている。

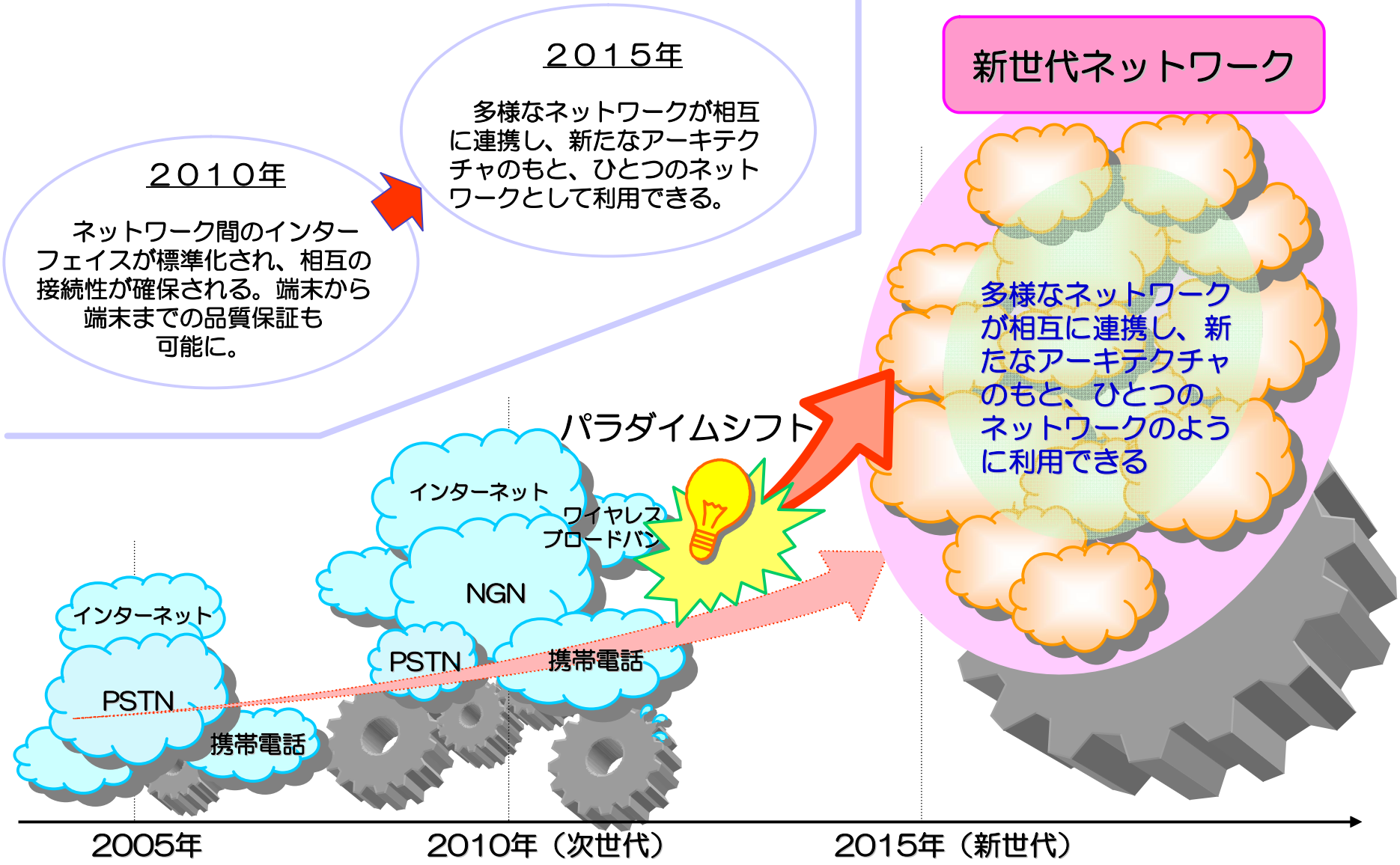
UNI、NNI及びANIはNGN特有の多様なインターフェースを許容する観点からノートが付されている。

NGN標準化のステップ

○ NGNでは特定のサービスと能力を段階的に実現
○ サービスの記述、アーキテクチャとプロトコル要求条件、プロトコルの3ステージで検討



総務省研究会での議論（将来的なネットワークアーキテクチャの進展イメージ）



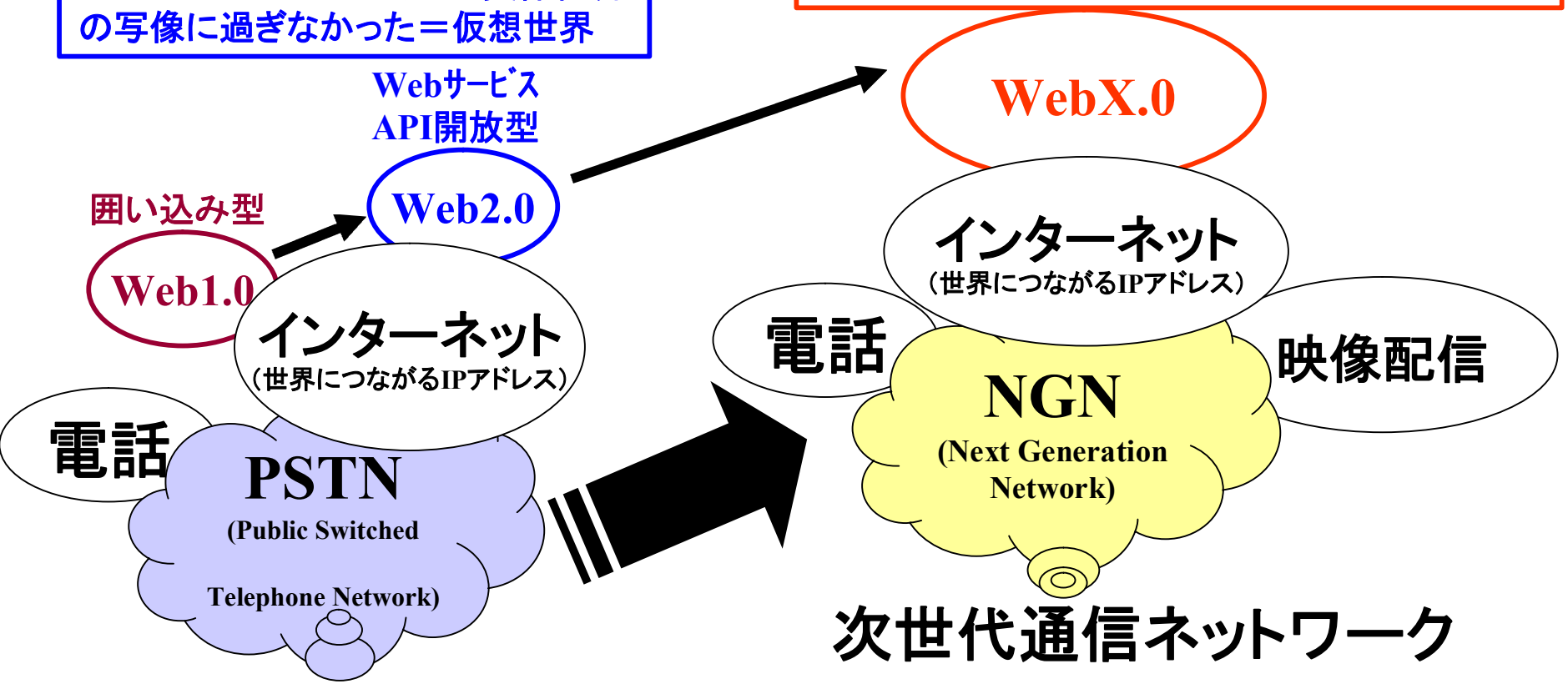
（出所：次世代ネットワークアークテクチャ検討会　資料）

次世代ネットワークとWebサービスの進化が同期

これからのインターネットはネットワークそのものが実体経済へと進化＝実世界

⇒*次なる数理科学の果たす役割！*

これまでのインターネットは実体経済の写像に過ぎなかった＝仮想世界



現在の通信ネットワーク

（通信キャリアが電話番号を管理）

次世代通信ネットワーク

（通信キャリアが独自のIP番号を管理）

ご清聴ありがとうございました