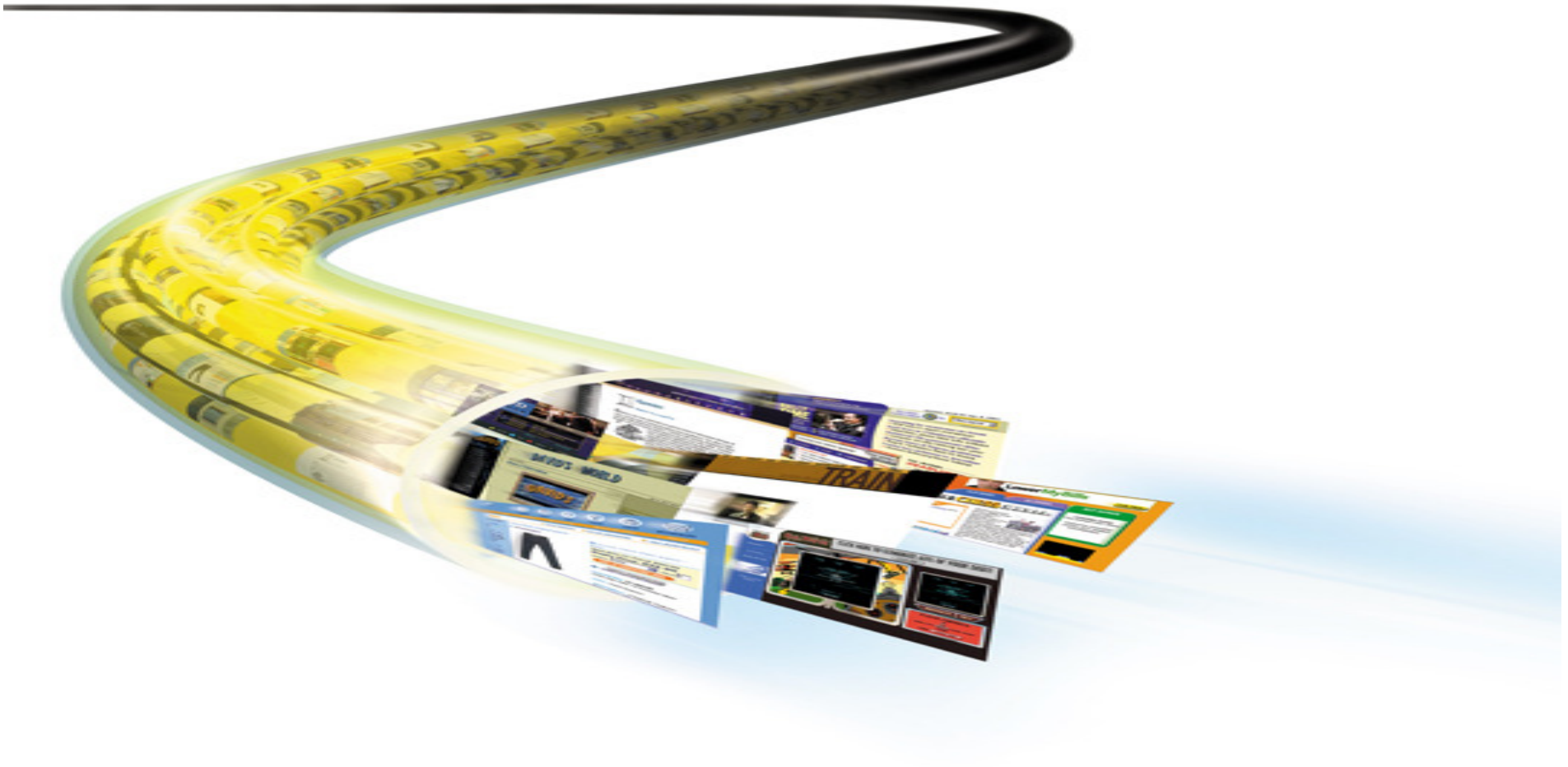
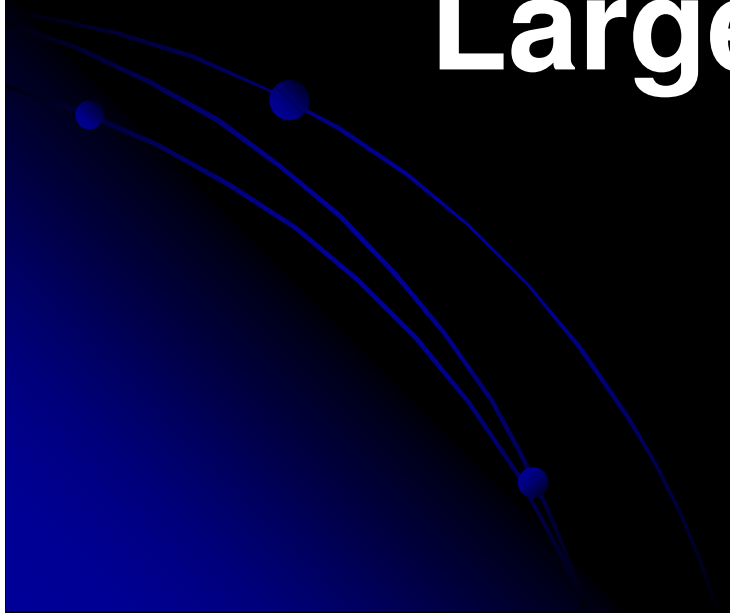

TRAIN

Did you know…

Large Databases?



AT&T has 11 exabytes (10^7 TB) of wireline, wireless, Internet data

That's

2+ trillion calls

= 1 million Libraries of Congress

In 2004

National Energy Research Scientific Computing Center (US)

2.8 petabytes

Google's BigTable (US)

1-2 petabytes

Wal-Mart (US)

500 TB

10^7 transactions / day

in 2004

World Data Center For Climate (D)

300 TB

Sprint (US)

2.85 trillion rows

ChoicePoint (US)

250 TB

Did you know…

Enterprise Database Growth?

OLTP Databases

2X every 5 years

2 years (core databases only)

OLTP Workloads

4X every 3 years

OLAP Databases

3X every 3 years

OLAP Workloads

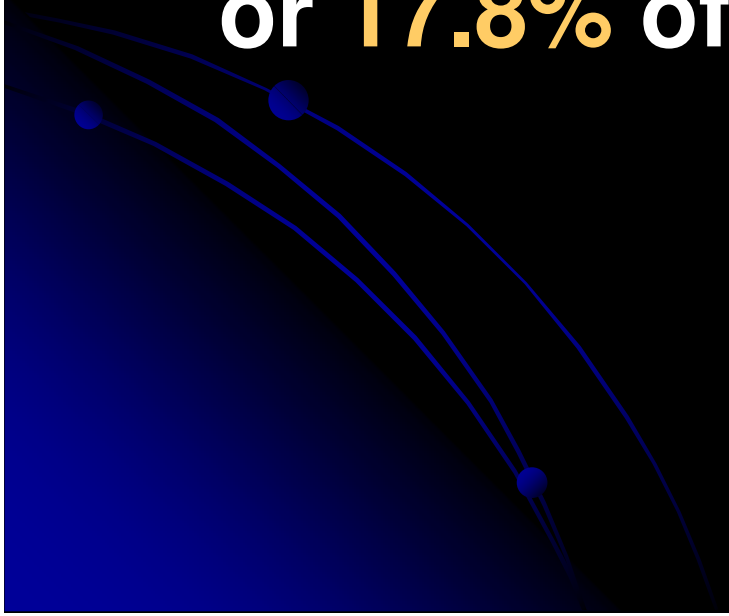
2X every 3 years

Did you know…

Our Digital World?

The Internet has …

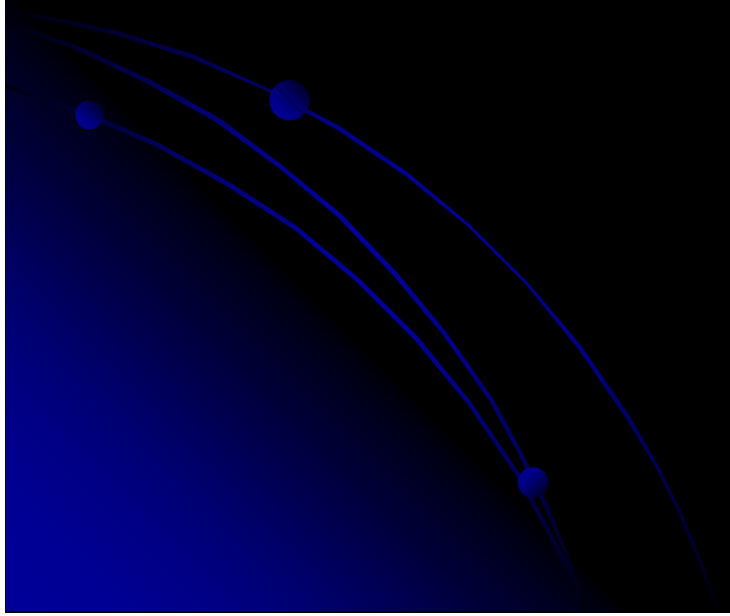
- 1/2 billion hosts (IP addresses)
- 1.17 billion users
- or 17.8% of the world's population



The Web has …

- 109 million distinct web sites
- 29.7 billion web pages
- ~5 pages for every man, woman, and child on the planet

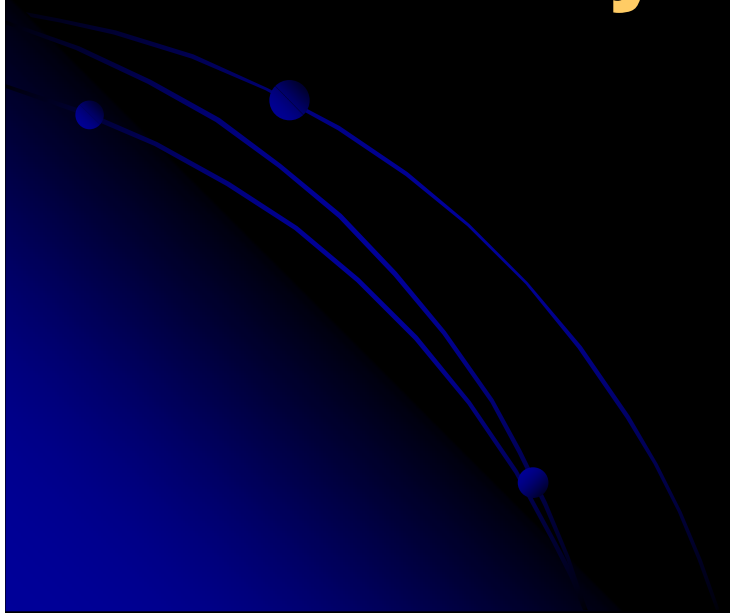
7.2 billion Web searches/month (3.9 billion by Google) far exceed the world population



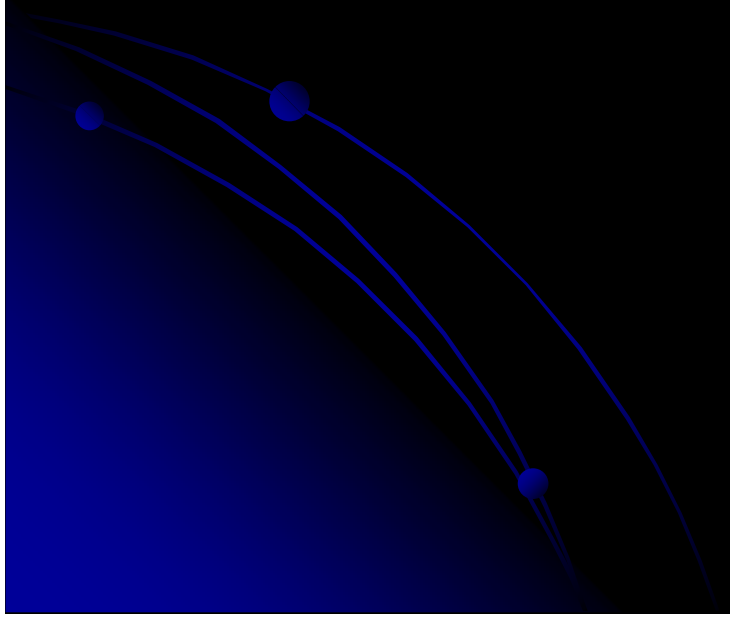
161 exabytes (10^8 TB) of information was created or replicated worldwide in 2006.

That’s more than in the previous 5,000 years.

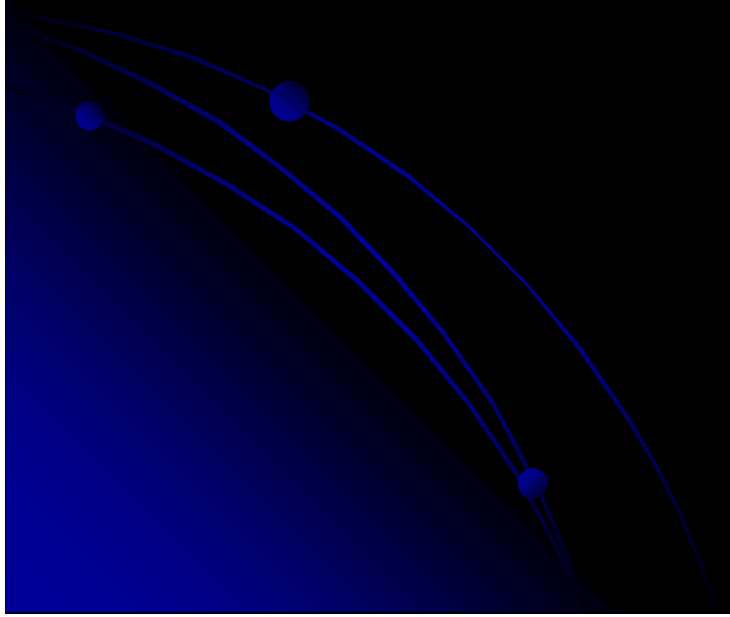
IDC estimates
6X growth by 2010 to
988 exabytes (a zetabyte) / year



New technical information doubles every 2 years.



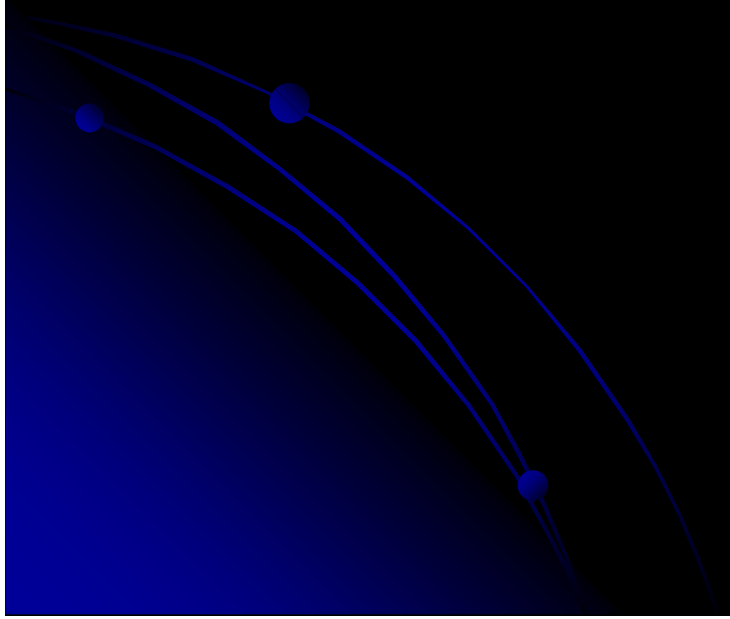
… every 72 hours
by 2010.



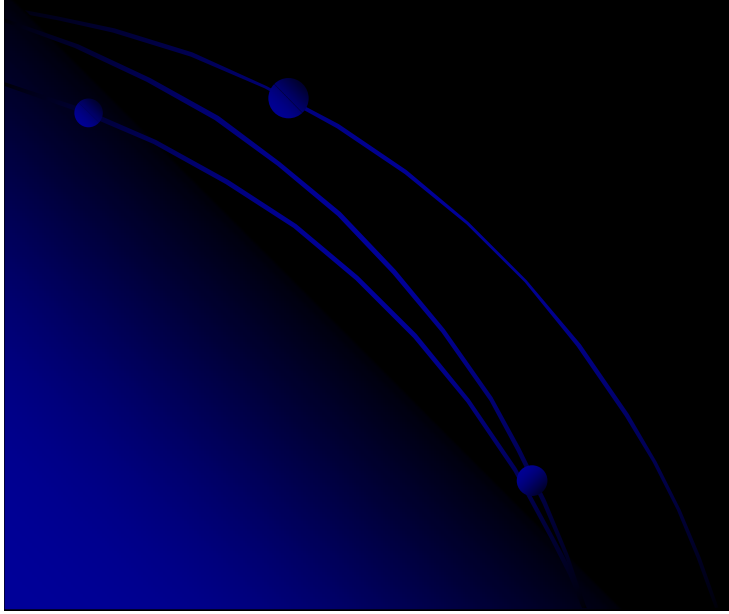
Did you know…

the largest source of data?

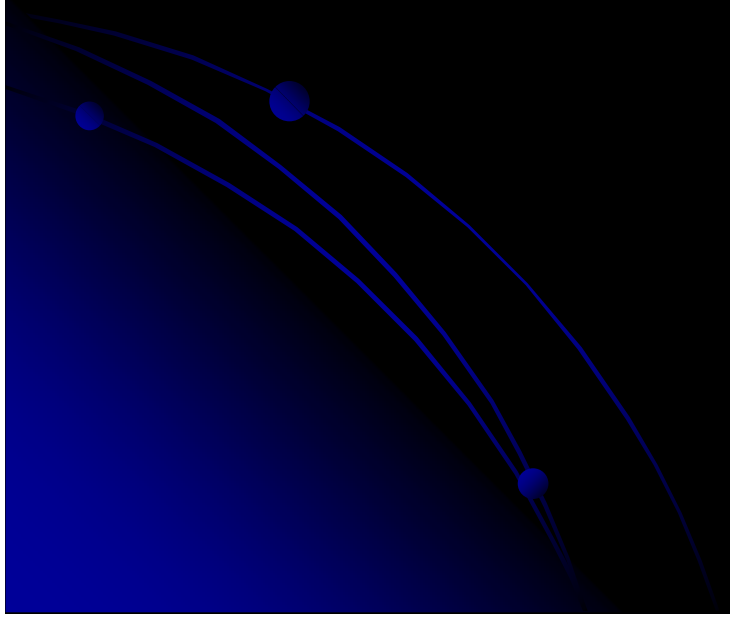
User Generated Content
[UGC]



With 185+ million registered users (April 2007) MySpace …



would be the 6th largest country (between Brasil and Pakistan).



The average MySpace page is visited 30 times a day, to access …

Images

- 1+ billion
- Millions uploaded / day
- 150,000 requests / second

Songs

- 25 million
- 250,000 concurrent streams

Videos

- 60 TB
- 60,000 uploaded / day
- 15,000 concurrent streams

Servers

- 6,000 web
- 650 ad
- 250 database

Facebook has …

- 1.8 billion photos
- 31 million active users
- 10^5 new users / day
- 1,800 applications

YouTube Videos

1.7 billion served / month

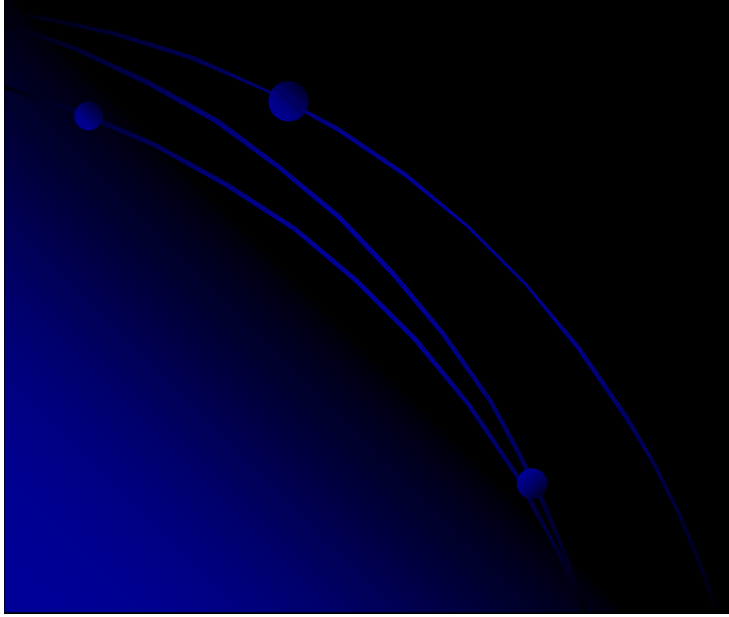
1 million streams / day

= 75 billion e-mails

Did you know…

devices that produce UGC?

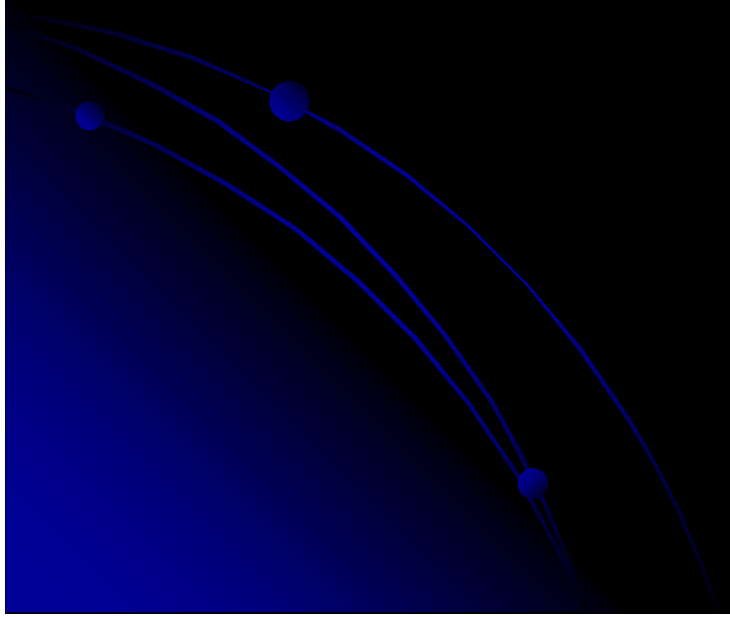
The world's 4+ billion devices
- cameras, phones, PCs, CCTVs, -
will increase 50% by 2010.



Did you know…

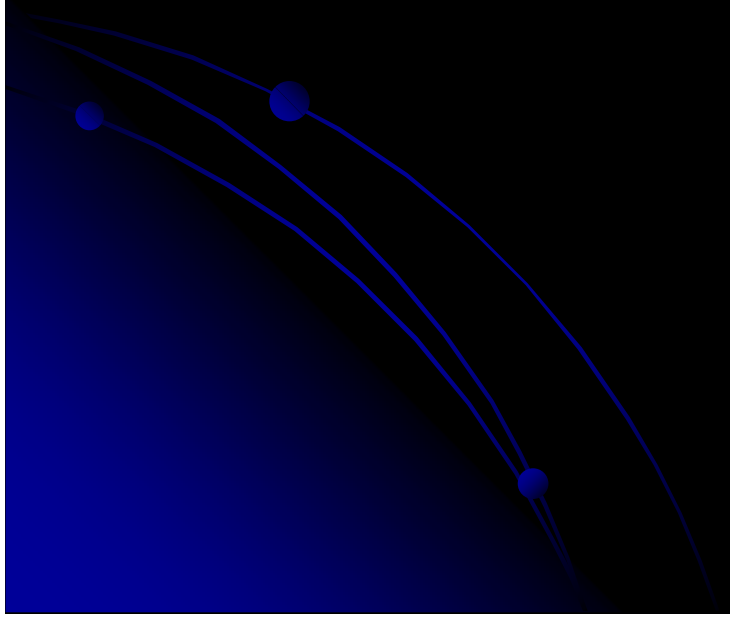
Internet traffic?

Third-generation fiber optics

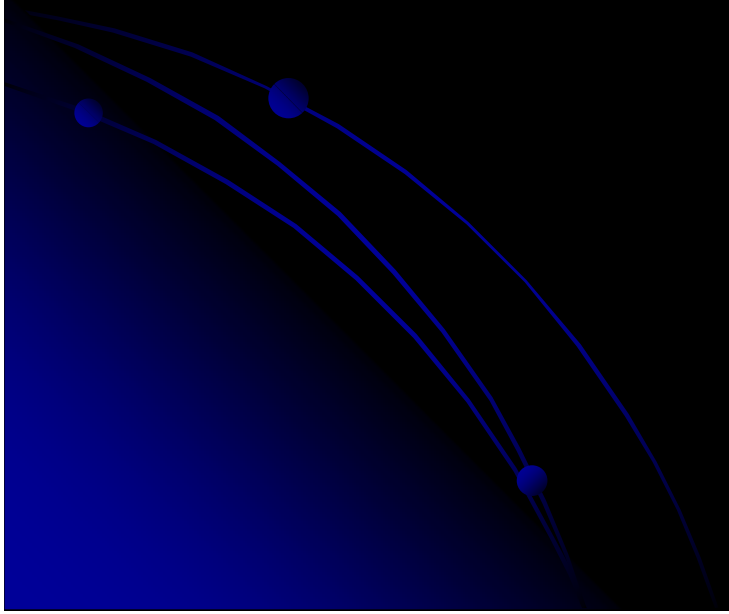


… transmits 10 trillion
bits / second.

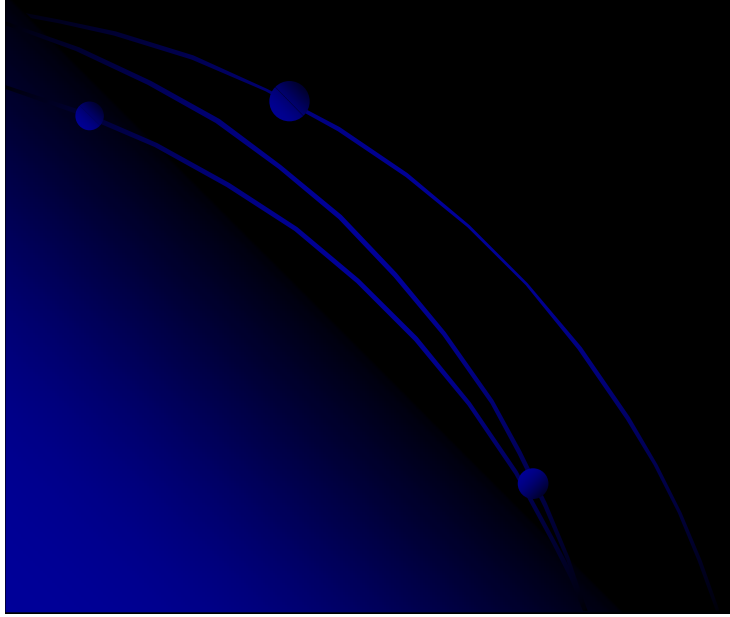
Equal to 1,900 CDs or
150 million simultaneous phone
calls every second.



… predicted to triple
every 6 months
for at least 20 years.

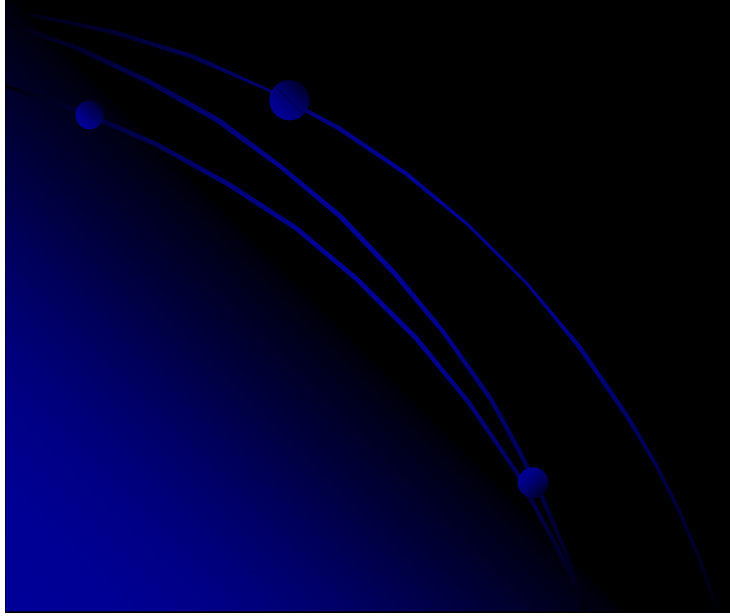


Cisco predicts that IP traffic will quintuple 2006 - 2011



… to 11 exabytes/month.

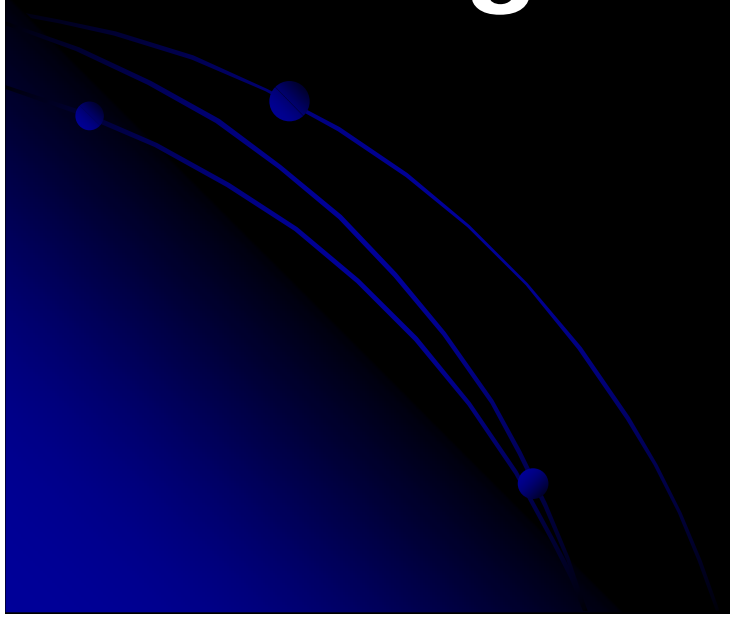
… and consumer IP traffic will surpass business in 2009.



Did you know…

Communications?

3+ billion calls per day
- wireline,wireless, and VoIP-
are growing at 50% CAGR



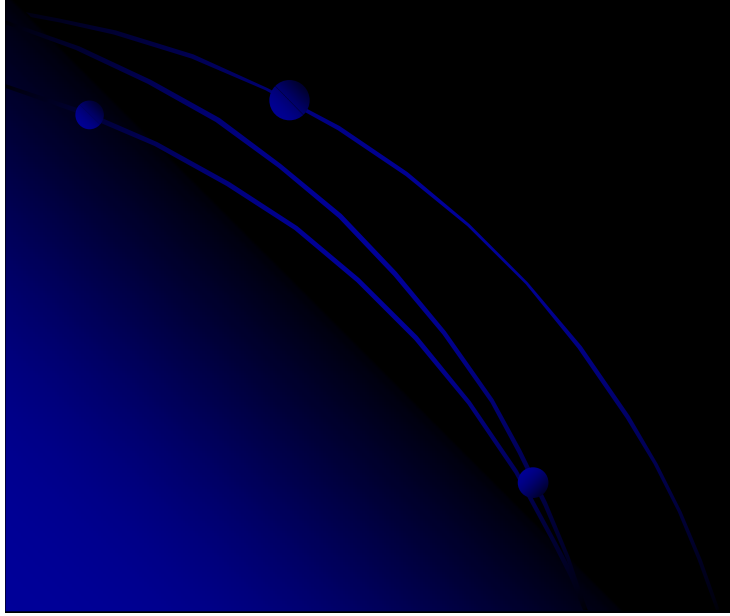
The number of text messages
sent daily
exceeds the world population

Did you know…

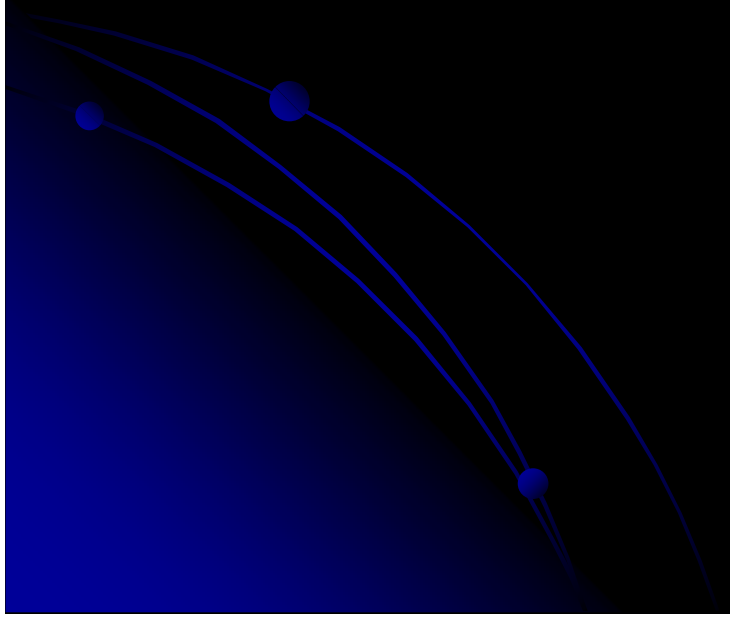
that this data is …

- 25% original; 75% replicated
- 25% from the workplace; 75% not
- 95% unstructured and growing
- While enterprise data is 10-15% structured and decreasing

What does this all mean?

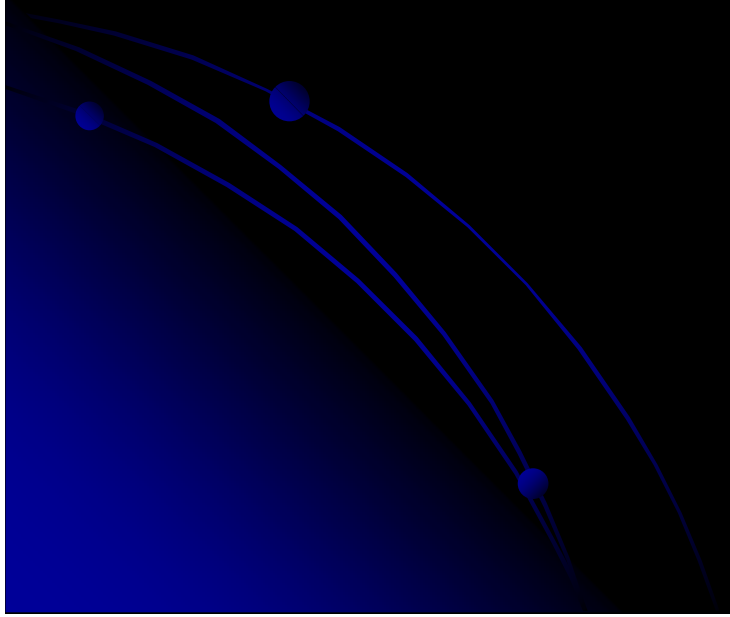


Shift happens.



We are entering

...



Computer Science 2.0

A New World of Data Management

Dr. Michael L. Brodie
Chief Scientist

Outline

- Databases - Bedrock of Modern Business
- Computer Science 2.0
- A New World of Data Management
- Semantic Technologies in CS 2.0
- Conclusion

© verizon 2007

Databases: Bedrock of Modern Business

- Robust, prototypically successful software market
  - US$15 billion /year
  - Continued growth despite growing gap: 10% @ 2X vs. 6X

- Databases in Large Enterprises (Fortune 50) & Ecosystems
  - Online data: petabyte - exabyte
  - Systems: 5,000 - 10,000; 10% mission critical
  - Databases: 3 - 5 landscapes / system
  - Overlaps: M&A, re-organization, …
  - DBMSs: 3 major; 20+ other
  - DBAs: 1,500+
  - Database lesson: minor DBA errors cost … plenty

© verizon 2007

Fundamental Logical Requirements

- **Communications Ecosystem**
  - Marketing
  - Ordering
  - Provisioning
  - Network operations
  - Billing
  - Care
  - Infrastructure

- **History**
  - In the beginning: stand-alone functions & product lines
  - 1980s: manual linkage
  - 1990s: long-distance & local, wireline & wireless, Internet, …
  - 2000s: business processes; Web - customer portals

- **Logical Requirements**
  - Massive Integration across the ecosystem: processes, applications, databases
  - Complex processing
    - Precise - finance, billing, operations, …
    - Approximate - trend analysis, BI, What if, …

© verizon 2007

Database Challenges

- **Barriers**
  - … complexity
  - … cost
  - … architectural
  - … operational (QoS)

- **Operational Requirements**
  - Agility
  - Self-management
  - Cost

© verizon 2007

Computer Science 2.0 Requirements

Physical

- Massive scale, growth, & complexity
  - Data
  - Update
  - Analysis
- Application specific processing
  - Protein folding, time series, stock trading, social computing, …
- Stringent operational requirements
  - Periodically disconnected
  - Fast recovery / failover
  - Security
  - Etc.
- Distributed across
  - Data stores
  - Platforms: Internet / wireline / wireless / devices
  - An enterprise ecosystem
  - Web
- Unstructured 95%
- Replicated 75%
- Heterogeneous
- Meta-data

Logical

- Ecosystem (Integration)
  - Processes
  - Aplications
  - Databases
- Complex Processing
  - Queries, transactions, business processes, and analysis
  - Precise and approximate

Operational

- Agility
- Self-management
- Cost

© verizon 2007

… in a Nutshell

- Scale
- Growth
- Diversity
- Trust

© verizon 2007

First Wave CS 2.0 Data Stores

Alternative RDBMSs

- Parallel processing appliances
  - Netezza
  - ParAccel
  - SAP BI Accelerator
  - Greenplumb
  - DATAllegro
- Stream processing
  - StreamBase
- XML
  - Tamino
  - X-Hive
  - TigerLogic
- OLAP Column Stores
  - Vertica
  - Sybase IQ Analytic Server
  - SAND Technology
  - ParAccel
  - MySQL

Web-Scale Physical Stores

- Google's BigTable - analytical workloads
- Yahoo!'s PNUTS - transactional workloads
- Yahoo!'s UDS, UPES, YDHT - Web Apps
- OceanStore: global store on unreliable servers
- Triple stores
  - Garlik: Person information on British citizens
  - YARS: web search & inference
  - Vertica: web search & inference
  - RDFLIb: meta-data store
  - RDF: meta-data store
  - SIMILIE
  - Piggy Bank / Semantic Bank
- User Generated Content stores
  - YouTube.com: videos
  - Joost: Internet-based TV
  - Flickr: photos
  - MySpace: social networking
  - Smugmug: photos
  - SecondLife: social networking
- Peer-to-Peer Data Stores: support P2P
- GRID: Scientific Data Services OGSA-DAI

© verizon 2007

First Wave CS 2.0 Targets

Physical

- Massive Scale, growth, & complexity
  - Data
  - Update
  - Analysis
- Application specific processing
  - Protein folding, time series, stock trading, social computing, …
- Stringent operational requirements
  - Periodically disconnected
  - Fast recovery / failover
  - Security
  - Etc.
- Distributed across
  - Data stores
  - Platforms: Internet / wireline / wireless / devices
  - An enterprise ecosystem
  - Web
- Unstructured 95%
- Replicated 75%
- Heterogeneous
- Meta-data

Logical

- Ecosystem (Integration)
  - Processes
  - Applications
  - Databases
- Complex Processing
  - Queries, transactions, business processes, and analysis
  - Precise and approximate

Operational

- Agility
- Self-management
- Cost

© verizon 2007

Computer Science 2.0

End of the Computing Era

Imagine a World in Which …

© verizon 2007

The Problem Solving Era

Computer science in the 20th century was about perfect solutions in closed domains and applications. Computer science in the 21st century will be about approximate solutions and frameworks that capture the relationships of partial solutions and requirements ... Dieter Fensel, IEEE Intelligent Systems, November/December 2007.

The old computing is about what computers could do; The new computing is about what people can do… Ben Shneiderman, Leonardo's Laptop: Human Needs and the New Computing Technologies, MIT Press, 2002

Problem Solving Worlds

- **Services Science (IBM)**
- **Immersive, Interactive Worlds**
  - **Information / High Performance Workplace**
  - **ERP**
  - **MathWorks**
  - **Second Life / Social Netwoking**
  - **Warcraft**
  - **Arden**
  - **Google Earth**

© verizon 2007

In Computer Science 2.0

Plumbing should
... disappear
and
manage itself

Cloud Computing

Computer Science 2.0 Plumbing

Agility

- **Service-Oriented Business Platforms (Cloud)**
  - Business Models: services science
  - Process Models: business processes
- **Service-Oriented Software Platforms (Cloud)**
  - Computational Model: remote invocation
  - Application Model: composite applications
  - Service-Life-Cycle: service composition and re-use
  - Development: independent and collaborative (pub / sub)
  - Data Model: (composite) data services
- **Service-Oriented Hardware Platforms (Cloud)**
  - Hardware elements: devices, networks, CPUs, blades, …
  - Virtual platforms: virtualization, P2P, GRID, SANS, …
    - Lateral and horizontal scaling

© verizon 2007

Problem Solving in the Clouds

- **Many moving parts**
  - Self-describing (meta-data)
  - Self-managing

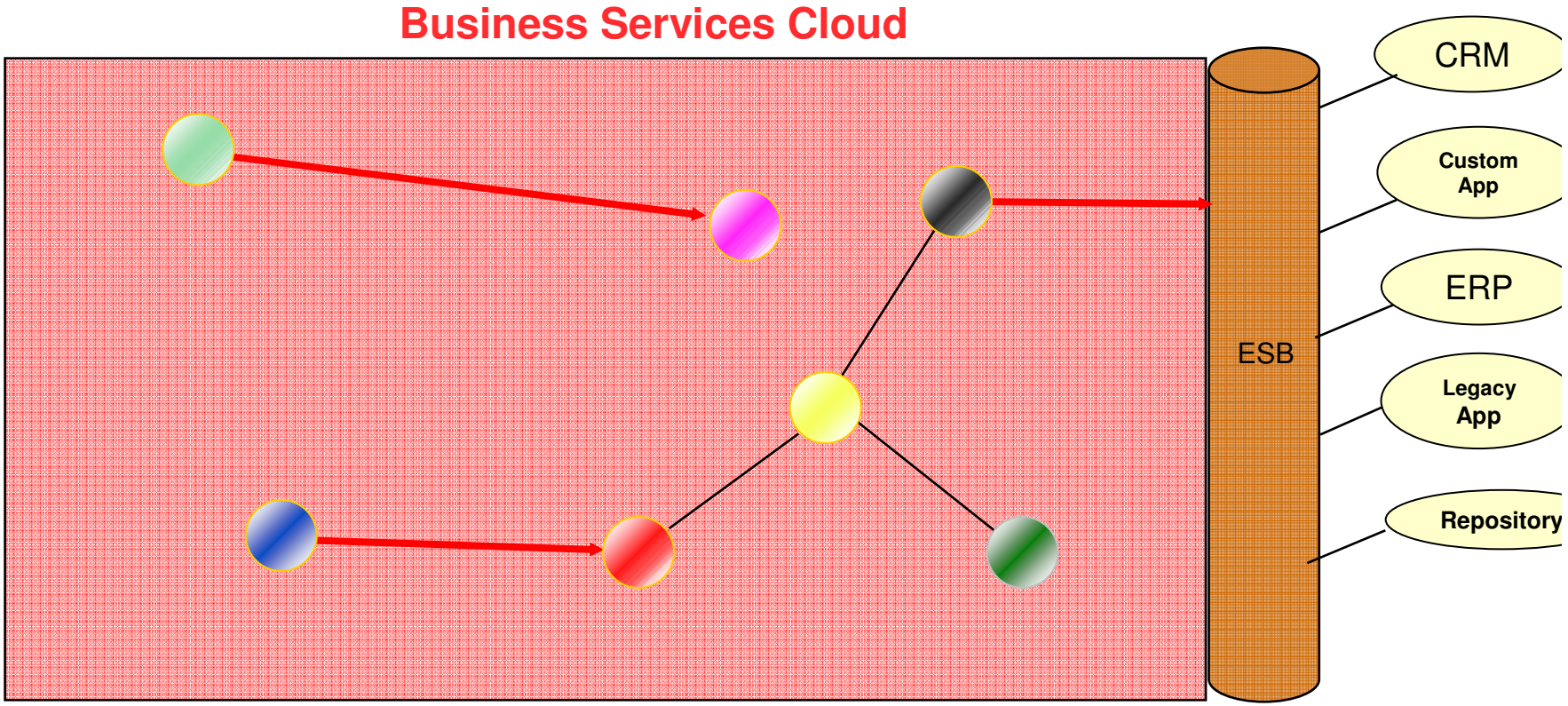
- **X-Driven**
  - Policies
  - Rules
  - Meta-data
  - Models
  - Context, e.g., provenance

Information that drives your business

should drive your digital business

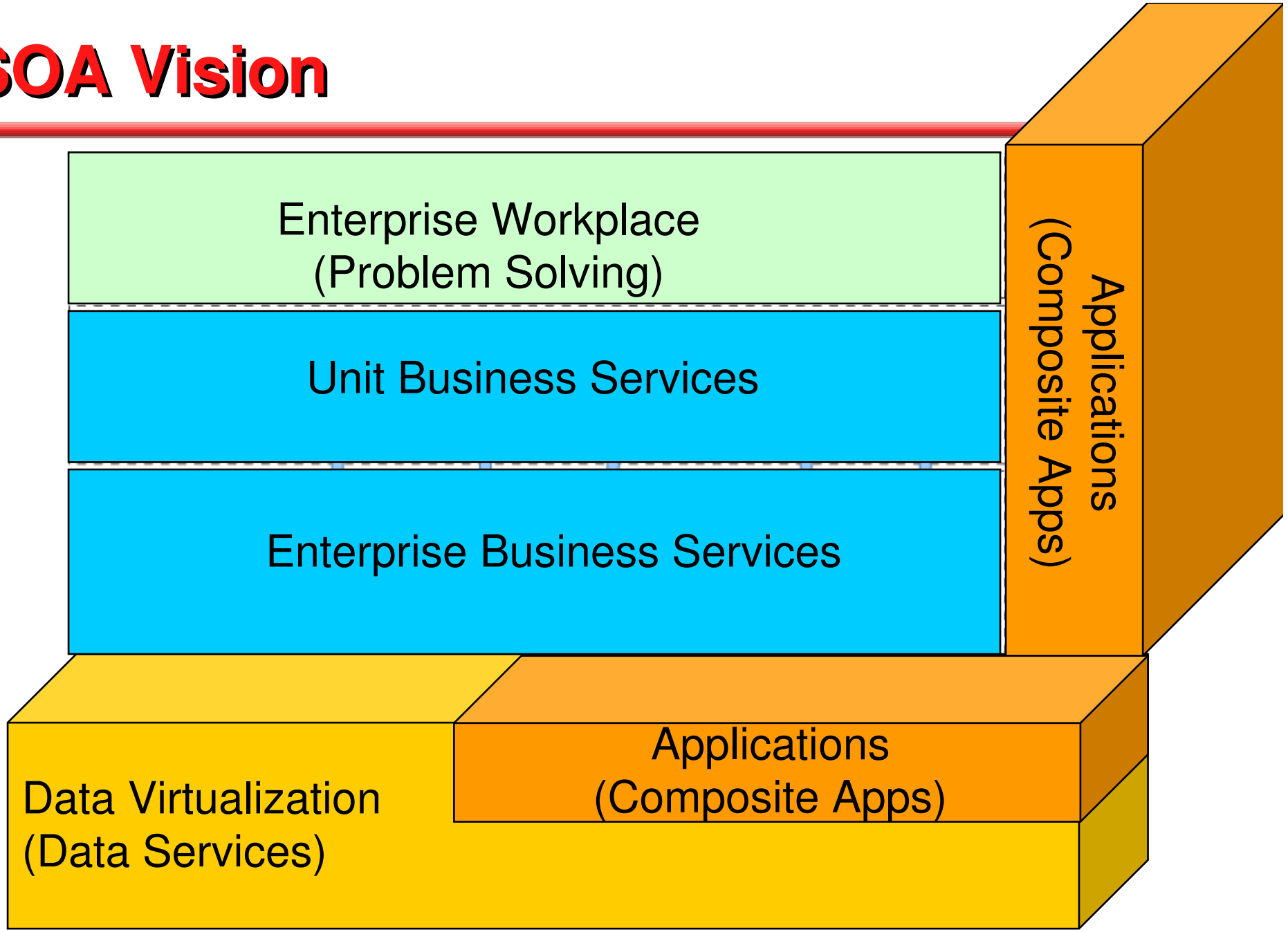
© verizon 2007

SOA Vision



- **Service Life Cycle**
  - Discover
  - Match
  - Select
  - Negotiate
  - Adapt
  - Compose
  - Invoke
  - Mediate
  - Monitor

SOA Vision


Enterprise Workplace
(Problem Solving)
Unit Business Services
Enterprise Business Services
Applications
(Composite Apps)
Data Virtualization
(Data Services)
Applications
(Composite Apps)

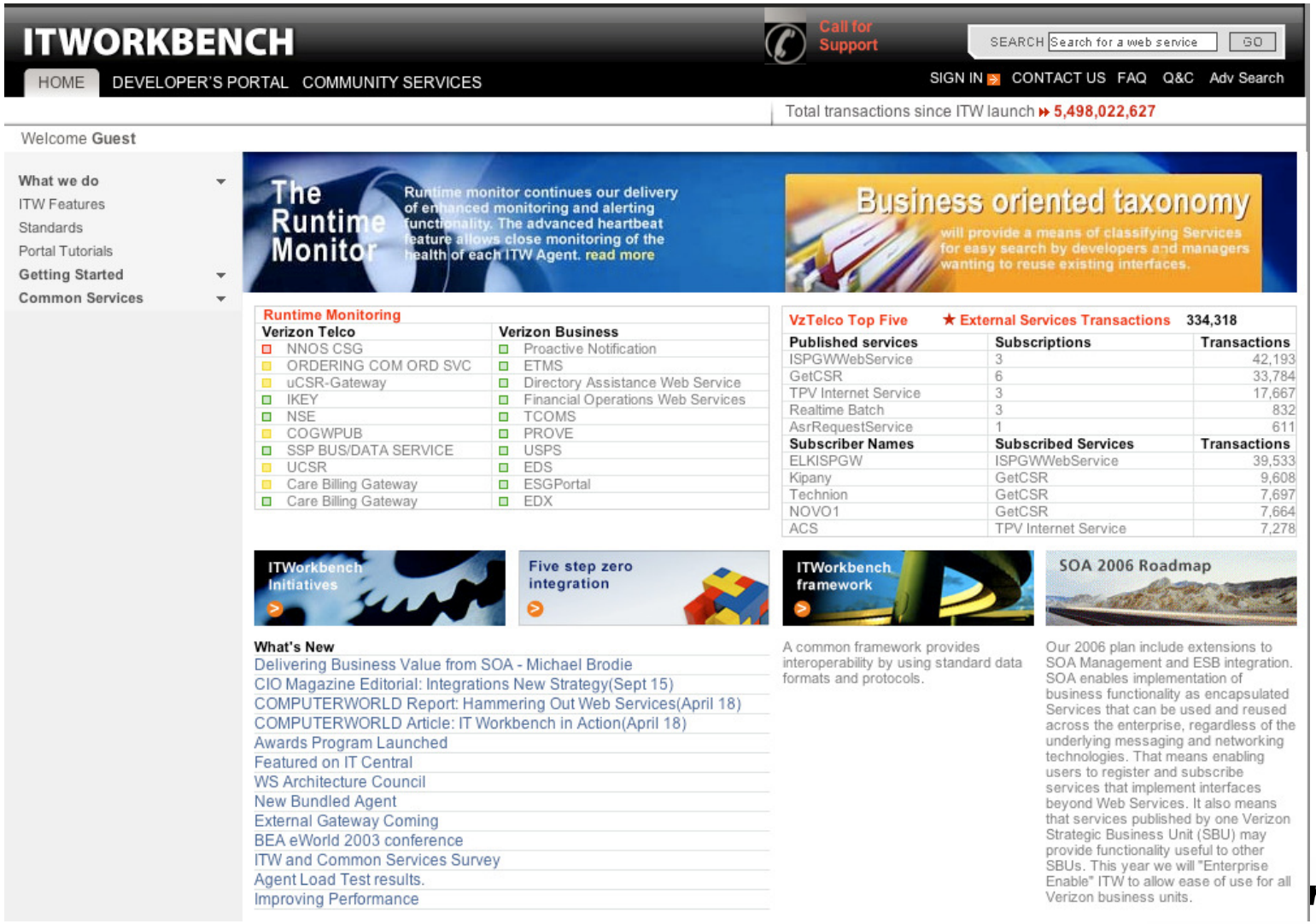
SOA Reality

- **Vendors Committed**
  - SOA will underlie the Next Generation of Computing
- **SOA Evolution**
  - Fundamentally new - misunderstood, complex, many moving parts
  - 2003-2004: Enterprise class solutions claimed
  - 2007: No complete SOA solutions
  - Evolving
    - 2006: Registry & repositories - core of SOA - introduced
    - Framework, Methods, Best Practices, Standards, …
- **SOA Adoption Rates in Enterprises**
  - 60-75% using SOA in some form
  - 20% selectively use SOA w/o strategy
  - 20% enterprise SOA strategy

© verizon 2007

IT Workbench: Verizon's SOA

- Developers Workbench
- SOA Execution Environment



SOA Scale

Verizon ITW	Gartner SOA Adoption Stage 4: Plateau
❑ Published services • Total: ~1,000 o Internal: 900+ o B2B: 100+ • Growth: 50+ / month	> 500
❑ Subscribers • Total: ~900 o Internal: 825+ o B2B: 70+ • Growth: 35 / month	> 50
❑ Service calls/day • Total: ~22 million o Internal: 21 million o B2B: 1 million • Since inception: 5+ billion	> 1 million

© verizon 2007

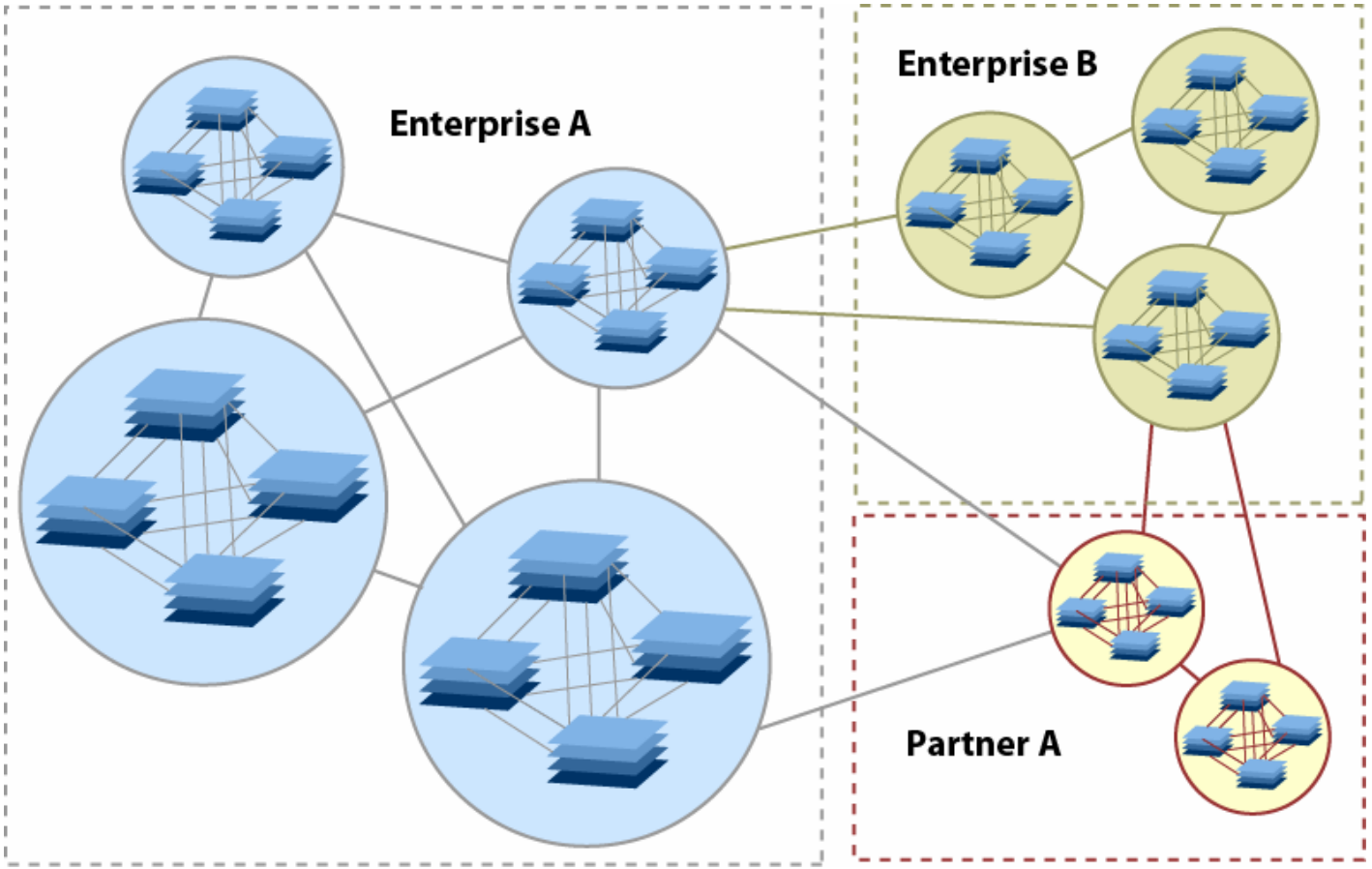
SOA Challenges & Maturity

- **Challenges**
  - Bigger Picture (CS 2.0)
  - No “Architecture” in SOA
  - No “integration solution”, yet every operation requires “integration”
  - Governance
  - Federation across heterogeneous SOA backplanes, e.g.,
    - Transparent development - What is composition (SCA)
    - Transparent QoS: performance, latency, and transaction
  - Dynamic SOA

- **Fundamentally new technologies take 20-30 years to mature**
  - Experiment: now
  - A robust, scalable platform
    - COTS product ecosystems: 2010-2012
    - Enterprise ecosystems: 2012-2015
  - Federated SOA: 2012+
    - Compose and execute a significant service-oriented application across an ecosystem transparently of multiple SOA backplanes including specifying and monitoring QoS
  - Dynamic SOA: trivial (2010), moderate (2015), significant (?)

© verizon 2007

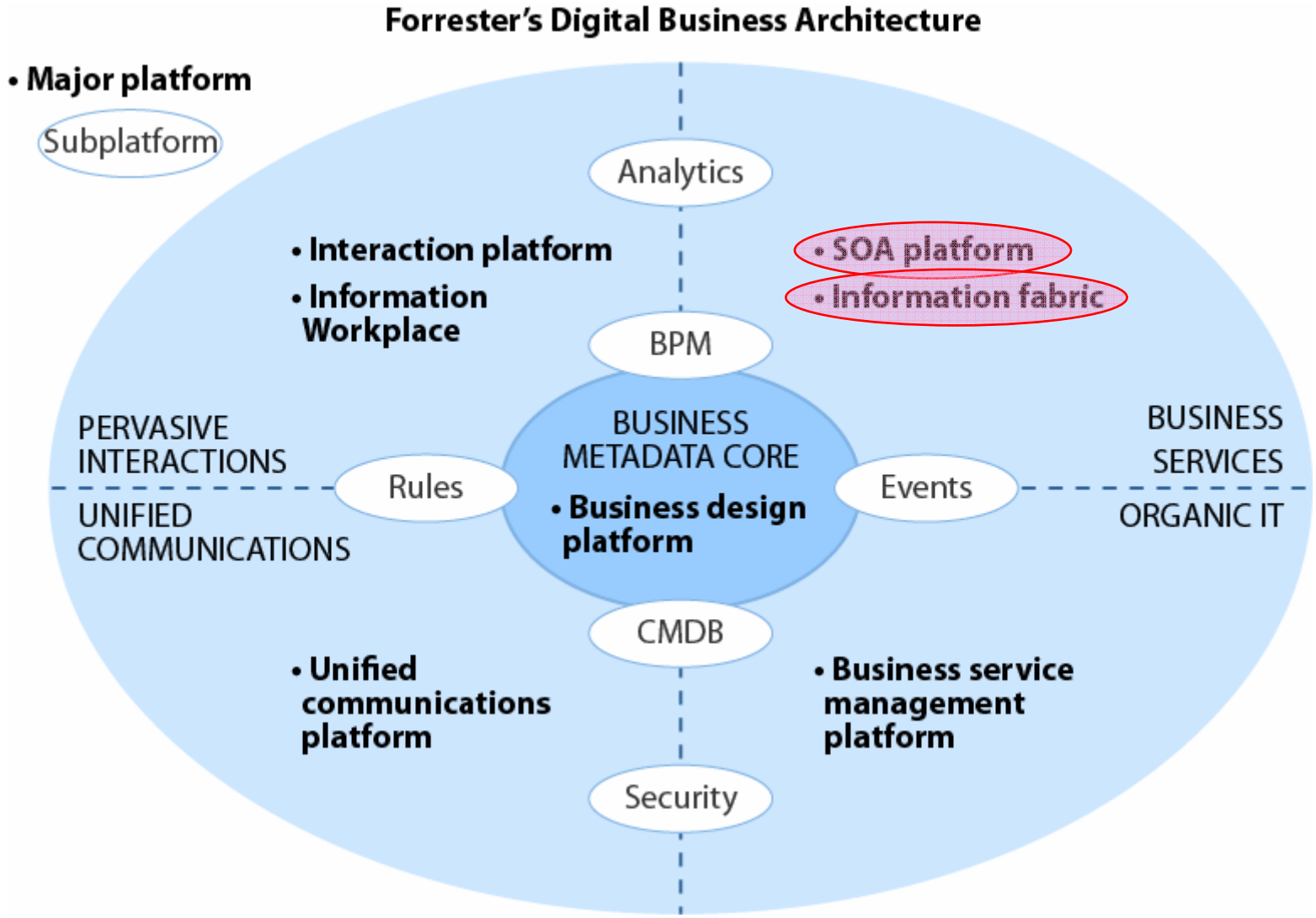
Multiple-Fabric Integration



© 2007, Forrester Research, Inc. Reproduction Prohibited

July 2007 **"A Taxonomy Of Platforms For Your Digital Business"**

Major Platforms Within The World Of Digital Business Architecture



© 2007, Forrester Research, Inc. Reproduction Prohibited

Digital Business Architecture

- **Seven strategic platforms**
  - SOA platform
  - Information fabric
  - Interaction platform
  - Information Workplace
  - Unified communications platform
  - Business service management platform
  - Business design platform

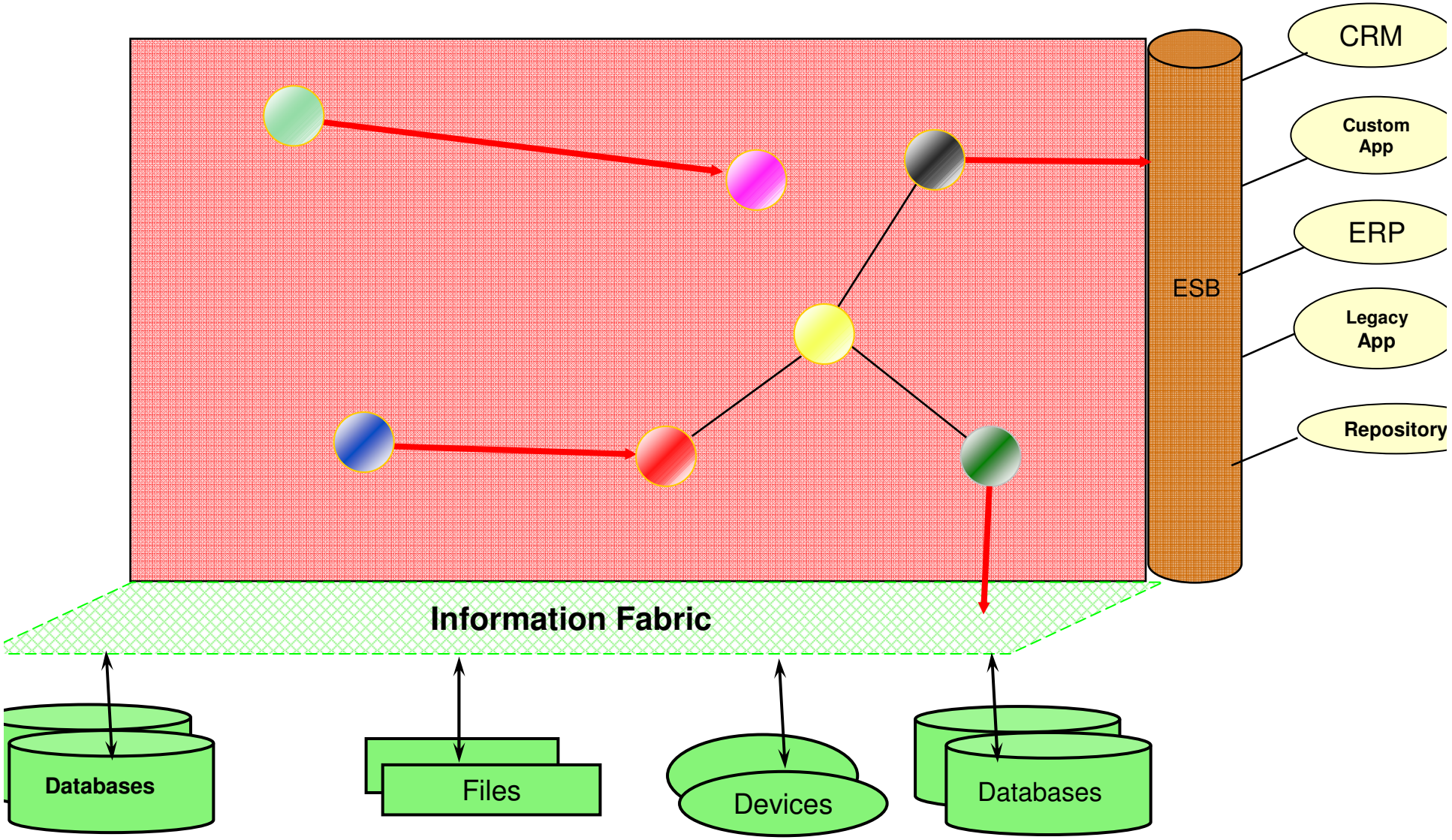
- **Six subplatforms**
  - Event management
  - Business process management
  - Business rules
  - Analytics
  - Configuration management
  - Security

© verizon 2007

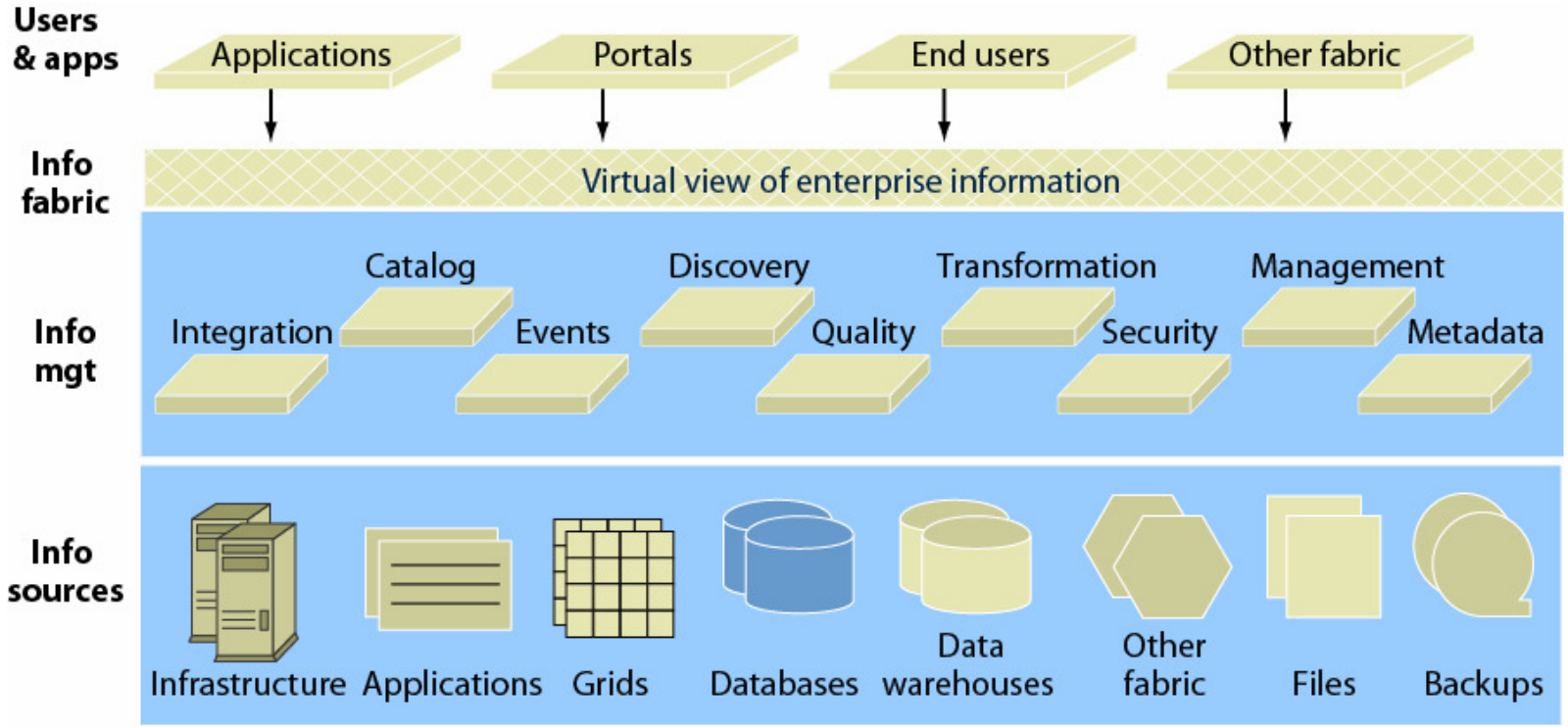
A New World of Data Management

Data Management in SOA

SOA Vision


CRM
Custom App
ERP
Legacy App
Repository
ESB
Information Fabric
Databases
Files
Devices
Databases

Logical Design



© 2007, Forrester Research, Inc. Reproduction Prohibited

Information Fabric Components

Information Fabric Components

- Service registry and discovery
- Integrated enterprise search
- Design-time metadata repository
- Distributed transaction management
- Distributed runtime metadata repository
- Distributed information access
- Information transformation and assembling
- Information quality
- Information security
- Distributed information cache

Is the Information Fabric the DBMS of Computer Science 2.0?

© verizon 2007

Information Fabric Adoption

- **Strategies and Standards**
  - Data Services (SDO standard)
  - Data Virtualization
  - Information - as -a - Service (IaaS)

- **Products: Service Oriented Access to data**
  - BEA AquaLogic Data Services Platform
  - MetaMatrix
  - Composite Software
  - [Fedora Digital Repository System]

- **Product Potential**
  - IBM
  - Oracle
  - Informatica
  - GemStone Systems
  - Microsoft
  - Progress Software
  - Sybase
  - Tangosol
  - Terracotta
  - Xcalia
  - Ipedo

© verizon 2007

SOA Scale & Re-Use

- **SOA Scale: in the future**
  - **Large SOA CRM application ~2,000**
    - **User interactions**
    - **Business services**
      - **Fine grained: < 2K lines of C#**
      - **SOA Re-write of 2 applications: 60 + 36**
      - **Converge to one: reduce to 30**
    - **Infrastructure services**
    - **Data services, e.g., 1,000-1,200 major, 600 minor**
  - **SOA Enterprise**
    - **Internal: ~1 million (e.g., above X ~5,000)**
    - **B2B: thousands**
  - **SOA Web: ~ billions [10,000 WSDL-based Sept 2007]**
- **Reuse: Now**
  - **Find 1 in 1,000**
- **Semantic Web Services**
  - **Enrich Web Service descriptions: SAWSDL**

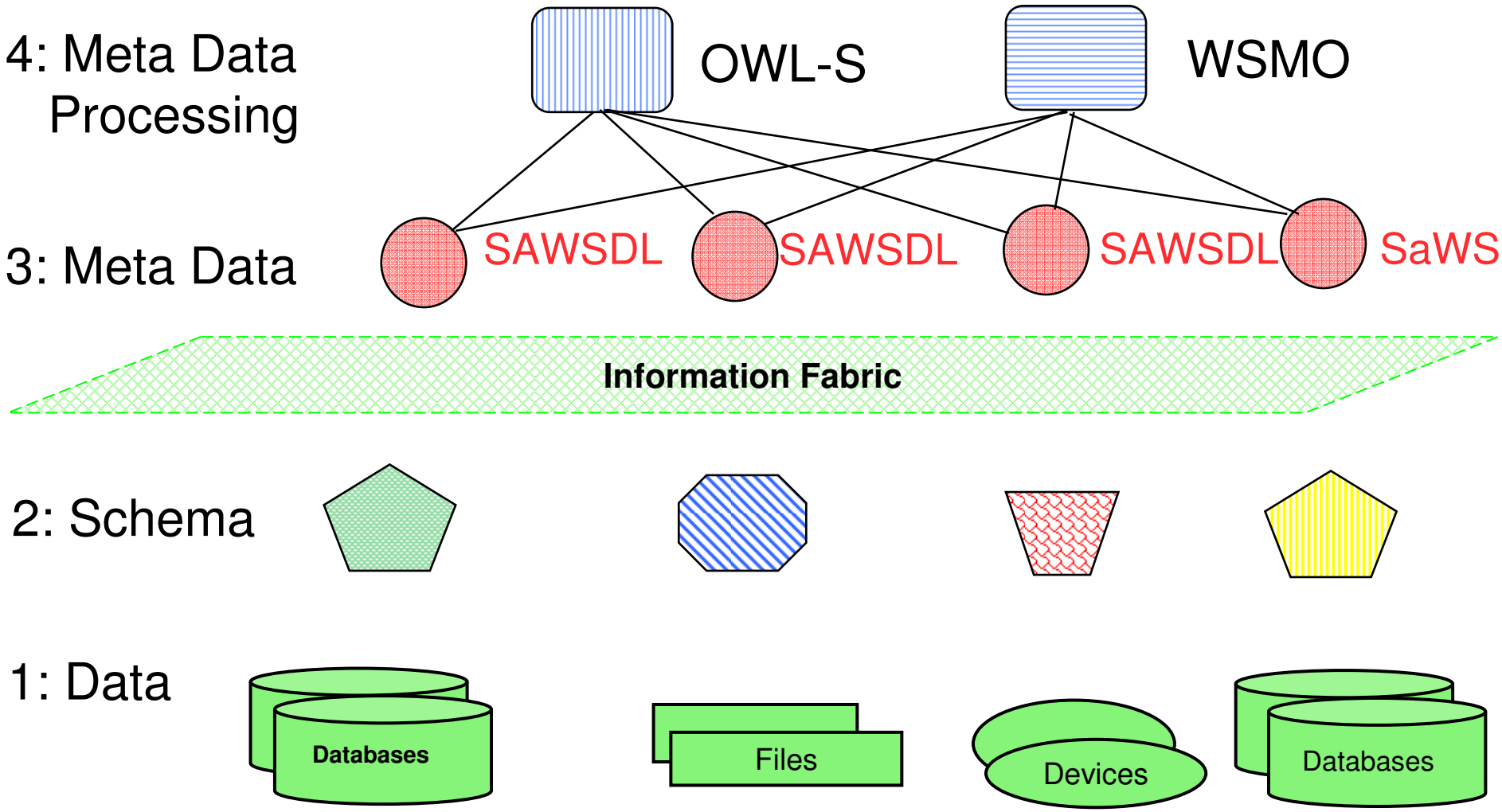
© verizon 2007

Semantic Technologies in Computer Science 2.0

Semantic Web Services

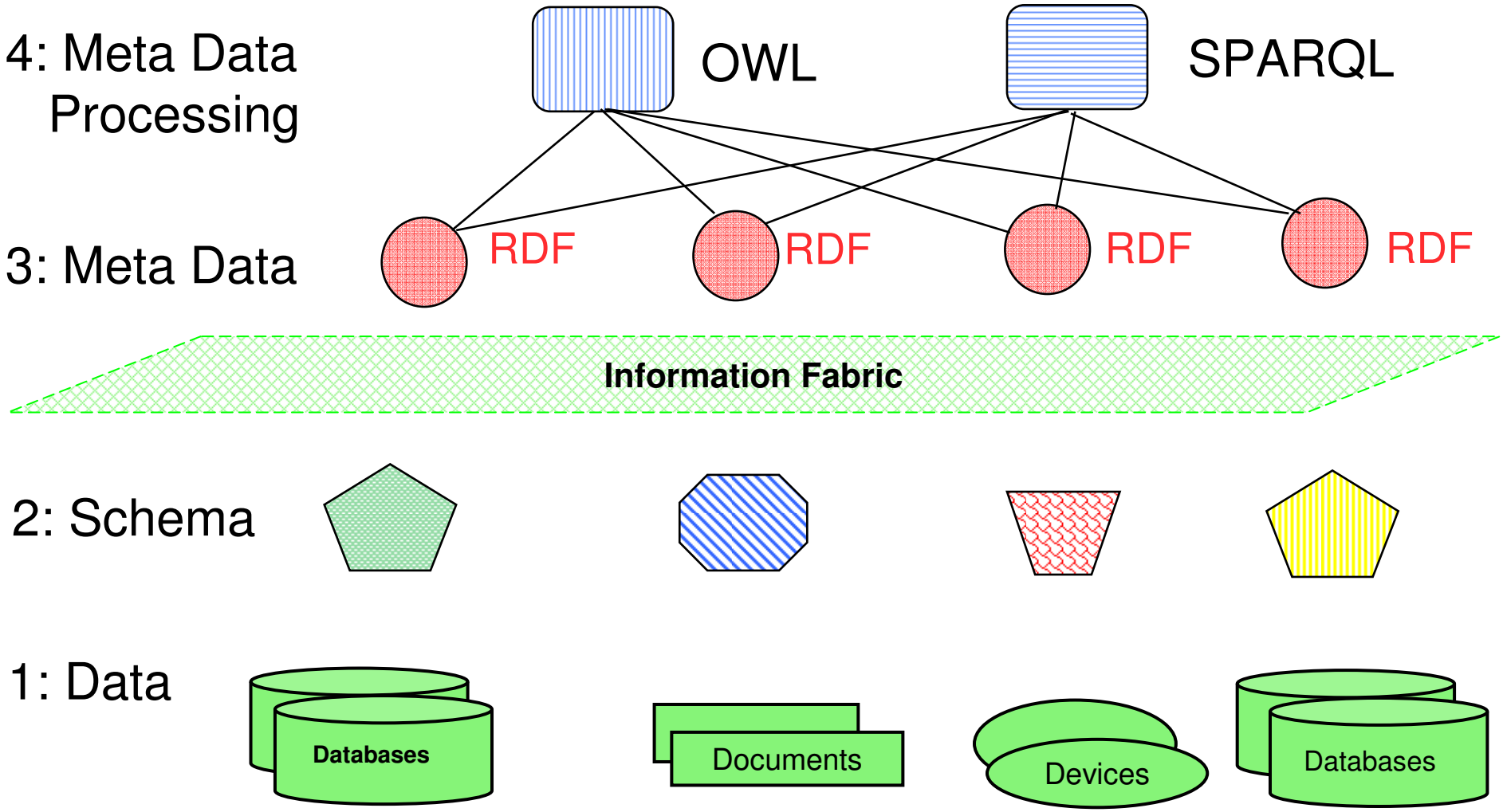
Semantic web

Web Scale SOA Data Integration (SWS)



© verizon 2007

Web Scale Data Integration (SW)



© verizon 2007

Semantic Technologies

Meta-data

- Process Languages
  - Business Process Execution Languages (BPEL)
  - BPLE For Web Services (BPEL4WS)
- Web / Web Service Description Languages
  - Resource Description Framework (RDF)
  - RDF schema (RDFS)
  - Web Ontology Language (OWL)
  - Web Ontology Language for Services (OWL-S)
  - Web Services Modeling Ontology (WSMO)
  - Web Services Description Language (WSDL)
  - Semantic enhancement of WSDL (WSDL-S)
  - Semantic Annotations for WSDL (SAWSDL)
- Ontology stores (RDF, OWL)
  - Triple stores
- Ontology tools, management, validation
  - RDF generators
  - Integrated Ontology Development Toolkit (IBM)
  - Unicorn (IBM) Semantic metadata discovery and management
  - Ontobroker, Ontobroker OWL (Ontoprise)
  - IE tools, mapping tools, ...
  - **Protégé**
  - **Solvent**
  - More @ Simile.mit.edu

Processing & Architectures

- Query / Reasoning / Inference
  - SPARQL: Distributed query over heterogeneous, distributed web resources
- Execution Environments
  - BPEL4WS (service / process models)
  - WSMX (Web Service Modelling eXecution environment) WSMO reference implementation
- Semantic application development environments
  - Top Quadrant: TopBraid
- Domain Specific Ontologies

© verizon 2007

Semantic Technology Adoption

- **1990+** Domain-specific ontologies
  - For large scale, complex information requirements: life sciences, healthcare, library, defense, government, energy, financial services, …

- **2007: Industrial scale use**
  - Search, information access, data quality and assessment, e-discovery and compliance, and content evaluation
  - Web resource descriptions: 2.3 million RDF web pages, 466 million triples (Sept 07)
  - Social network tagging - FoaF, del.icio.us, flickr, MySpace, Slashdot, Wikipedia, Yahoo! Answers, YouTube, Zillow.com.

- **2017: Semantic Data and Semantic Web Applications**
  - Ontologies (RDF, OWL): scale, information integration, and data quality

- **2027**: Semantic Web and Semantic Environments

© verizon 2007

Source: Enterprise Semantic Web in Practice: Jeff Pollock, Oracle, SemTech 2007

Vendor Implementations: Survey of Tools

Triple Stores

- @Semantics RDFStore
- Franz Allegrograph
- IBM Boca
- Intellidimension RDF Gateway
- Northrop Grumman Tucana Suite
- Ontotext OWLIM
- OpenLink Virtuoso
- Oracle Database 10.2

- **Enterprise Search and Collaboration**
- OpenLink Semantic Web Data Spaces
- Radar Networks
- Siderean Seamark Navigator

Reasoners

- RacerPro

Middleware

- IBM WSSR
- Microsoft Connected Services Framework
- Ontology Works
- Ontoprise
- Oracle Fusion Middleware 11g
- Profium Semantic Information Router
- Software AG (webMethods)
- Thetus Publisher

Metadata Tagging

- Adobe XMP & Aduna Metadata Server

Development Environments

- Altova SemanticWorks
- HP Jena
- TopQuadrant TopBraid Composer

Semantic Technology Adoption

- SOA vendors proceeding cautiously
  - Modest RDF, OWL support
    - Registry, repository
    - More?
  - Experimenting with semantics
    - Customer demand, e.g., life sciences

- Goals
  - Solutions
  - Standardization
  - Enterprise-class

© verizon 2007

If it Works for SOA then …

- **Data Management**
  - DBMS - semantic heterogeneity
  - Data Warehouse
- **Meta-Data Management**
  - Metadata Management
  - Master Data Management
- **Data / Information Integration**
  - Extract, Translate and Load (ETL)
  - Enterprise Information Integration (EII)
  - Enterprise Information Management (EIM)
- **Content Management**
  - Web content management
  - Record management
  - Document management
  - Portals
- **Governance**
  - Policy management
  - Rules management
  - SLA management

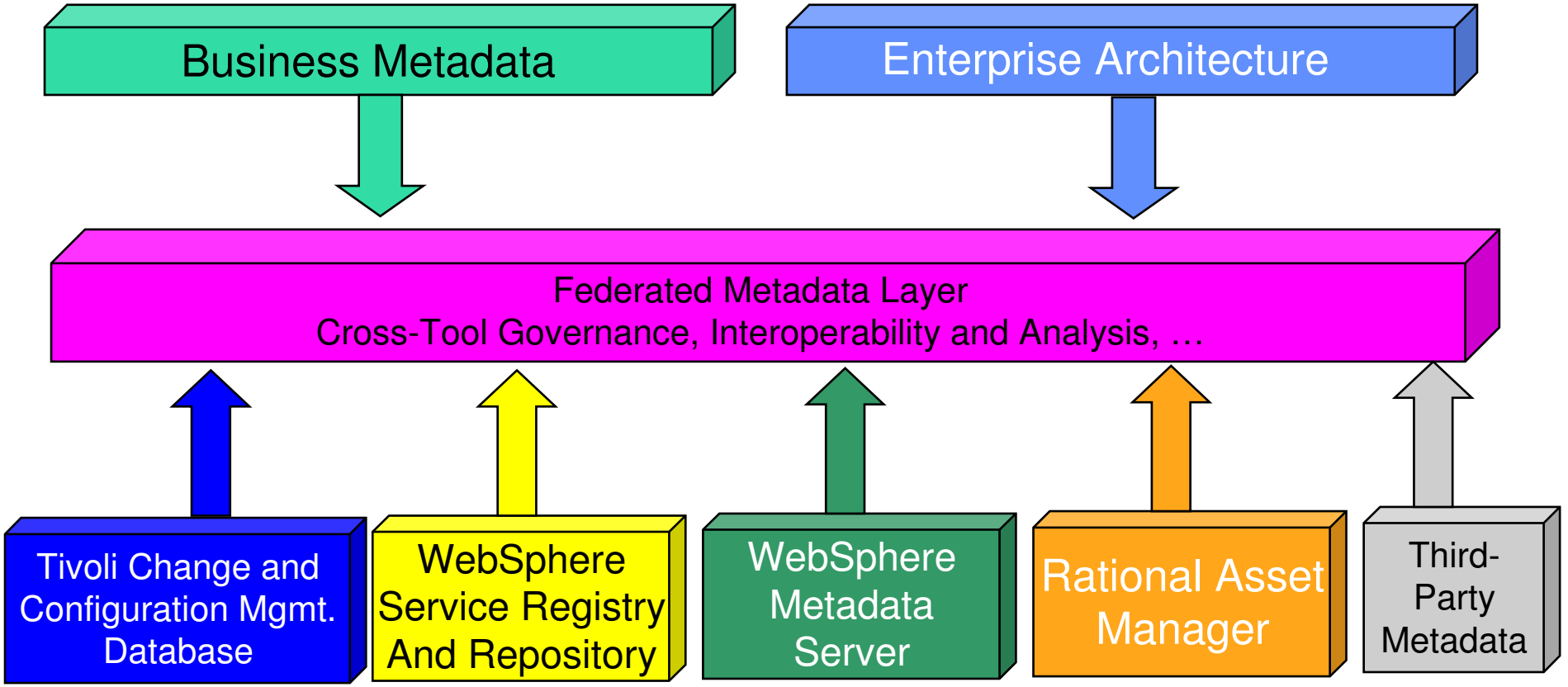
© verizon 2007

And …

- **Application Integration**
  - Enterprise Application Integration (EAI)
  - B2B Gateway
  - Enterprise Service Bus (ESB)
- **Process Integration**
  - Business Process Management (BPM)
  - Business-to-Business Integration (B2Bi)
- **Processes, Applications, and Information Integration**
  - Application Servers
  - Application Server Platforms
  - Application Platform Suites
- **Enterprise Workplace**
  - Enterprise Search
  - Collaboration
  - Communication: E-mail, IM, …
  - Role-based Workspace
  - Identity management

© verizon 2007

Federating Metadata …



Source: IBM

Conclusion

- **Wave of automation**
  - Scale
  - Growth
  - Complexity
  - Diversity
  - Trust
- **Computer Science 2.0**
  - Cloud computing
  - Service-orientation
  - Semantic Technologies
- **New World of Data Management**
  - Physical: wave of CS 2.0 data stores
  - Logical: Information Fabric
- **Database - Semantic Technology Collaboration for Integration**
  - Database: scale, data, information,
  - Semantic Technology: semantics

© verizon 2007