Our Expanding Digital World
Can we Contain it? Can we Manage it?

Dave Reinsel
Group Vice President, Storage and Semiconductor Research
May 2008
Agenda

The Expanding Digital Universe
- Its growth and characteristics
- How do we contain it?
- How do we manage it?
- The dangers that lie ahead

Key Takeaways
IDC’s Report on the Digital Universe

- Measure the total amount of digital information created in the world
- Forecast the digital universe by region – NA, WE, AP, ROW – to 2011
- Analyze the characteristics of the information in the digital universe based on information source and type
- Measure and forecast the available storage capacity through 2011
- Derive implications for society, enterprises, and governments
- Draw conclusions for enterprise IT organizations and vendors
# Number of Model Elements

## Image Capture/Creation
- High End Cameras
- Digital Cameras
- Camcorders
- Camera phones
- Webcams
- Surveillance
- Scanners
- Multifunction Peripherals
- OCR
- Bar Code Readers
- Medical Imaging
- Digital TV
- Digitized Movies & Video
- Special Effects
- Graphics Workstations

## Data Creation
- **PC applications**
- **Data base**
- **Office Applications**
- **Email**
- **Video/teleconference**
- **IM**
- **Other**

## Data Storage
- **HDD**
- **Optical**
- **Tape**
- **NV Flash Memory**
- **Memory**

## Digital Voice Capture
- Landline Telephony
- Voice over IP
- Mobile Phones

## Server Workloads
- **Business Processing**
- **Decision Support**
- **Collaborative**
- **Application Development**
- **IT Infrastructure**
- **Web Infrastructure**
- **Technical**
- **Other**

## Other
- Terminals, ATMs, Kiosks, Specialized Computers
- Industrial machines/cars/toys
- RFID
- Sensors
- Smart Cards
- Video games
- MP3 players
- SMS
- GPS
Key Findings

The digital universe will grow over six-fold, from 281 exabytes in 2007 to 1,773 exabytes in 2011.
Key Findings

The digital universe will grow over six-fold, from 281 exabytes in 2007 to 1,773 exabytes in 2011.

Roughly 8.9TB/s

About 43GB/person in the world

If we put all of this information in typical books, we could have 19 separate stacks that each could reach the sun in 2011?

56TB/s

3 stacks of books from the sun to Pluto (in 2010, it was just 2 stacks!)

When do we surpass a Yottabyte?
Internet and Broadband. Today, over 1.3 billion on-line users WW, and over 50% with BB – continuously increasing.

About 15 TRILLION emails sent in 2007 – 60% are spam! 1 email with a 1MB attachment sent to 4 people, can translate into over 50MB with backups and network overhead.

IDC estimates that 370 billion pictures were captured by digital still cameras and camera phones in 2007 (500 billion by 2011).
And it’s not ‘just’ digital TV – it’s also HIGH DEF

High Definition:

- Increases storage requirements
- 1 hour of HD video is roughly 10GBs of storage (depends on encoding rates)
- Requires more bandwidth and special equipment
- But is it worth it….?
Fueling the Digital Universe: DTV

HIGH DEF

Std. DEF
**Fueling the Digital Universe:**

The Digital Shadow

Only $\frac{1}{2}$ of the information related to you is information you create yourself.

The rest is simply information *'about'* you that is captured and stored...

... this is your [Digital Shadow](#)

... and it’s only going to grow
Key Findings

- The digital universe will grow from 281 exabytes in 2007 to 1,773 exabytes in 2011

- > 90% of the information in the digital universe is unstructured and absolute
  # of files growing faster than the TBs
Technology Options for Managing Pain Associated with File Overload

Increasingly moving toward Clustered File solutions

Clamped File Storage
(File Systems & NAS)

File-Virtualization

Purpose-Built NAS

Local File System (GPS)

File-Level

Pain

Clumped

Standalone
Key Findings

- The digital universe will grow from 281 exabytes in 2007 to 1,773 exabytes in 2011.
- > 90% of the information in the digital universe is unstructured and absolute.
  # of files growing faster than the TBs.
- By 2010 ~ 70% of the digital universe will be created by individuals.
“Edge” Devices: Mobile, Connected, and Creating Files

Secure

Smart

24x7

Protected

Enterprise Computing
Enterprise Storage vs. Personal Storage

From 2006 to 2012, more HDD terabytes will be shipped into personal storage than into enterprise storage.

Personal storage: USB-attached external HDDs

Enterprise storage: storage systems, servers, or content depots that reside in organizations

Prompting an increase in online-storage providers
The Content Disruption 2.0

Content and the advertising surrounding it are the major if not only sources of revenue

- Reliable content access and protection are mission critical

Manage large, fast growing pools of content

- 1PB to >40PB and growing at >100% a year
Key Findings

- The digital universe will grow from 281 exabytes in 2007 to 1,773 exabytes in 2011
- > 90% of the information in the digital universe is unstructured and absolute
  # of files growing faster than the TBs
- By 2010 ~ 70% of the digital universe will be created by individuals
- At the same time organizations will have responsibility or liability for 85% of the information
User Creation; Enterprise Worries

**Handle With Care**

- User* Generated Content
  - 1,234 Exabytes WW

- Organizational Touch** Content
  - 1,530 Exabytes WW

* Consumers and Workers Creating, Capturing, or Replicating Personal Information

** Transported, Hosted, Managed, or Secured

2011
1,773 Exabytes
Does <Delete> really mean Delete?

We spend billions of dollars to ensure data persistency, but we give little thought to data eradication!
What Will We Keep Online/Offline Archives By Far are the Majority

Archived data dwarfs active data in most organizations
- Significant duplication
- Limited organization
- Limited management capability

Impact on Companies
- Reducing media consumption becomes high priority (VTL)
- Data de-duplication becomes important feature (even compression)
- Proactive data management and organization become critical
Content Creation VS Duplication

We are ‘Pacmen’ and ‘Pacwomen’

- “might need it some day” mentality
- “Digital” enables multiple copies
- MP3s
- Movies, movies movies
- Our digital shadow
- Can’t possibly store everything, can we?

75% of the digital universe
is a copy
a replicate
a duplicate
a backup
a mirror image
is redundant

25% of digital universe is created or generated
Key Findings

- The digital universe will grow from 281 exabytes in 2007 to 1,773 exabytes in 2011.
- > 90% of the information in the digital universe is unstructured and absolute.
- # of files growing faster than the TBs.
- By 2010 ~ 70% of the digital universe will be created by individuals.
- At the same time organizations will have responsibility or liability for 85% of the information.
- In 2007, the information created and replicated (the DU) exceeded the available storage capacity.

Can we Containerize it….ALL?
Information And Storage

- **Information**
- **Available Storage**

 transient information, unused streams, temporary data

<table>
<thead>
<tr>
<th>Petabytes</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100K</td>
<td>200K</td>
<td>300K</td>
<td>400K</td>
<td>500K</td>
<td>600K</td>
<td>700K</td>
</tr>
<tr>
<td>800K</td>
<td>900K</td>
<td>1M</td>
<td>1.1M</td>
<td>1.2M</td>
<td>1.3M</td>
<td>1.4M</td>
<td>1.5M</td>
</tr>
<tr>
<td>1.6M</td>
<td>1.7M</td>
<td>1.8M</td>
<td>1.9M</td>
<td>2M</td>
<td>2.1M</td>
<td>2.2M</td>
<td>2.3M</td>
</tr>
</tbody>
</table>

- 2005 2006 2007 2008 2009 2010 2011

© 2008 IDC
“Available” Storage Worldwide not Enough to Save all Data Created

“…every it – every particle, every field of force, even the space time continuum itself – derives its function, its meaning, its very existence entirely – even if in some contexts indirectly – from the apparatus-elicited answers to yes-or-no questions, binary choices, *bits.*”

1A John Wheeler quote in an essay titled, “Information, Physics, Quantum: The Search for Links, “ in *Complexity, Entropy, and the Physics of Information*
Reliability is Becoming More Important

Use of S/ATA HDDs into enterprise storage is increasing rapidly

RAID technology provides data protection from HDD failures…but not all RAID strategies can protect sufficiently from BER*

*Bit Error Rate
Enterprise drives are spec’d at (1) irrecoverable read error for every $10^{15}$ bits read – some ATA drives are at $10^{14}$ – effectively increasing the expected failure during a RAID rebuild event by an order of magnitude.

The use of very large drives causes concern with large RAID arrays.

The use of a TB drive may be discouraged until acceptable protection schemes (such as Dual Parity RAID or more advanced ECC by HDD manufacturers) are implemented.

Assume a BER of $10^{15}$

<table>
<thead>
<tr>
<th>Drive Size</th>
<th>RAID 7+1 (SP*)</th>
<th>RAID 15+1 (SP*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>300GB</td>
<td>2%</td>
<td>4%</td>
</tr>
<tr>
<td>500GB</td>
<td>6%</td>
<td>10%</td>
</tr>
<tr>
<td>1,000GB</td>
<td>10%</td>
<td>14%</td>
</tr>
</tbody>
</table>

*Single Parity
How Would you like a Terabyte Disk Drive?

You may have just won….

WINNER!...?

Provide to: Storage OEMs WW

In a single HDD: 1,000,000 Megabytes!*

*For purposes of measuring storage capacity, a megabyte (MB) is equal to 1,000,000 bytes. Total operating capacity varies depending on operating environment.
HDD Areal Density Growth

- 25% agr
- 60%
- 100%

- MR Nano Heads
- GMR Heads
- TuMR/PMR sub-Pico Heads
- AFC Media
- DTM/BPM Media
- 35 - 40%
HDD Areal Density Growth

The chart illustrates the growth of HDD areal density from 1990 to 2015, comparing various storage technologies like HDDs, Magnetic Linear Tape, DVD/MO, Holographic, and Flash/Solid State Devices. The y-axis represents areal density in Gbits/in², and the x-axis shows availability years from 1990 to 2015.

Key observations:
- HDDs have consistently shown the highest areal density growth.
- Flash/Solid State Devices are projected to have the highest areal density by 2015.
- Magnetic Linear Tape and DVD/MO have lower areal density growth.
- Holographic and MRAM technologies fall in between these two extremes.

Source: IDC 2008
CAUTION

Information ‘management’ overload. Dangers that need to be addressed. Opportunity knocks!
Datacenter storage requirements continue to increase 50-55% year after year.
Annual Growth Curves Reveal Problem

- Enterprise HDDs
- Enterprise Storage Systems
Bottom line: HDD AD is NOT keeping up!
The cost of energy continues to rise and companies do not have unlimited cooling capacity.

Companies do not have unlimited space and power.
A Caveat to Metcalfe’s Law?
- In general, the value of a network increases as the number of active nodes increase.
- IDC research consistently reveals that datacenter storage requirements are increasing 50–60% per year for the next several years.
- Google has leveraged this network quite successfully!

Can too much of a good thing turn sour?
- Telephone and telemarketers (caller ID)
- Email and spam (filters and blockers)
- Data growth and data irrelevance
Some Critical Keys to Information Management

The actual data

The ability to find the data

The ability to read the data

The accuracy and security of the data
The Danger of Expiration

“I guess we never worried much about data preservation.”
The Danger of Expiration: Conflicting Curves

# Years

Media Life

Amount of Data Capture

Stored Capacity

Something must be done
Not a new issue:

“The content and historical value of many governmental, organizational … and personal documents may be irretrievably lost to future generations if we do not take steps to protect them,” Jeff Rothenberg, 1999.

*Jeff Rothenberg, a preservation specialist at Rand Corp.*
# Storage Technology Futures to 2015

<table>
<thead>
<tr>
<th>Cross-Point Arrays</th>
<th>Magnetic</th>
<th>Optical</th>
<th>Mechanical</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAND Flash</td>
<td>HDD</td>
<td>CD/DVD</td>
<td>Probe Storage</td>
</tr>
<tr>
<td>NOR Flash</td>
<td>Tape</td>
<td>Magneto-Optical</td>
<td>Carbon Nanotube</td>
</tr>
<tr>
<td>NROM Flash</td>
<td></td>
<td>Holographic</td>
<td></td>
</tr>
<tr>
<td>FRAM</td>
<td></td>
<td>Near-Field</td>
<td></td>
</tr>
<tr>
<td>PFRAM</td>
<td></td>
<td>Biological Macromolecules</td>
<td></td>
</tr>
<tr>
<td>MRAM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OVRAM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organic Molecules</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Future:**
- Magnetic Race Track
- Spintronics

Promising technologies include:
- Magnetic Race Track
- Spintronics
Precise and Efficient Data Migration
Extending Media Life Not Enough

- Efficient
- Precise
- Secure

New Storage Technology
Can We Find the Data? The ’Ole “Needle in a Haystack”

How long do we have?

- Regulations
- Probes

Can we actually find the data?

- Indexing, metadata
- Filtering, querying

Solutions

- Active archiving
- Automatic discovery

Easier for text, but what about pictures, video, sound?

Time is of the essence!

Search is a killer app for storage!
Neil Armstrong was the first man to walk on the moon, but Stan Lebar was the man that was going to prove it.

← What the world saw broadcast

← What the world could’ve seen

*Photo Source: Wired Magazine, January 2007*
“Lost in Space” — Apollo 11 Story

But, the world never will … until they find the original tapes!

**NOT Henry Winkler**

*Photo Source: Wired Magazine, January 2007*
Data Capture Transitions

- **Written, Picture, Language**
- **Analog**
- **Digital (1s and 0s)**

<table>
<thead>
<tr>
<th># Years</th>
<th>Stored Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Analog</td>
</tr>
<tr>
<td>1</td>
<td>Analog</td>
</tr>
<tr>
<td>2</td>
<td>Analog</td>
</tr>
<tr>
<td>3</td>
<td>Analog</td>
</tr>
<tr>
<td>4</td>
<td>Analog</td>
</tr>
<tr>
<td>5</td>
<td>Analog</td>
</tr>
<tr>
<td>6</td>
<td>Analog</td>
</tr>
<tr>
<td>7</td>
<td>Analog</td>
</tr>
<tr>
<td>8</td>
<td>Analog</td>
</tr>
<tr>
<td>9</td>
<td>Analog</td>
</tr>
<tr>
<td>10</td>
<td>Digital (1s and 0s)</td>
</tr>
<tr>
<td>11</td>
<td>Digital (1s and 0s)</td>
</tr>
<tr>
<td>12</td>
<td>Digital (1s and 0s)</td>
</tr>
<tr>
<td>13</td>
<td>Digital (1s and 0s)</td>
</tr>
<tr>
<td>14</td>
<td>Digital (1s and 0s)</td>
</tr>
<tr>
<td>15</td>
<td>Digital (1s and 0s)</td>
</tr>
<tr>
<td>16</td>
<td>Digital (1s and 0s)</td>
</tr>
<tr>
<td>17</td>
<td>Digital (1s and 0s)</td>
</tr>
<tr>
<td>18</td>
<td>Digital (1s and 0s)</td>
</tr>
<tr>
<td>19</td>
<td>Digital (1s and 0s)</td>
</tr>
<tr>
<td>20</td>
<td>Digital (1s and 0s)</td>
</tr>
</tbody>
</table>
Can We Read the Data?
Interpreting the 1s and 0s
Can We Read the Data? Interpreting the 1s and 0s

**100 YEAR ARCHIVE TASK FORCE**

**Document** “best practices for long-term digital information retention”

**Influence** ILM as a core management and automation practice for long-term archive

Industry standard interfaces between applications and storage systems

**PDF/A** is a constrained form of Adobe PDF version 1.4 intended to be suitable for long-term preservation of page-oriented documents

**Extensible Markup Language (XML)** (e.g., ODF, Open XML)
Are we interpreting the data correctly? Has anything been lost in the translation?

“I didn’t murder him.”

A. I didn’t murder him.
B. I didn’t *murder* him.
C. I didn’t murder *him*.
D. I DIDN’T MURDER HIM.
Data security is increasing in priority

- WORM
- Access/replication/modify rights
- Encryption

Data security can also mean “complete” data erasure

2010 WW security spending more than doubles the $32B spent in 2005
Key Takeaways

The Digital Universe is Growing 50-55% Annually
- Vast majority is unstructured and duplicate information
- Majority of content creation driven by consumers
- Results in Creators and Custodians

The Creators will be increasingly enabled and mobile

The Custodians will be increasingly challenged
- What to keep and How to keep it (media life, efficiency, power)
- How to keep data relevant (metadata, response times, availability, the ‘right’ data)
- How to keep data safe from datajacking (security, encryption)

Hardware still important, but software is increasingly vital
The industry needs to standardize on common global standards — XAM and PDF/A are worthy initiatives, but there could be others.

Power issues are only going to get worse and technologies should be introduced to increase storage efficiency (hardware and software)
Thank you!

Please email me at dreinsel@idc.com