Relative Trend of Technologies

Year

1
10
100
1000
10000

Improvement Factor

Processor 1983 @ 0.33 MIPS
Storage 1983 @ 10MB
Communications (home) 1983 @ 1200bps

Source: IBM Research
What Will We Carry?

- Carry Computer
  - ThinkPad

- Carry Storage
  - Microdrive

- Carry Card
  - SmartCard

- Carry Nothing
  - Biometrics
HDD Areal Density Perspective

45 Years of Technology Progress

<table>
<thead>
<tr>
<th>Production Year</th>
<th>Areal Density Megabits/in²</th>
</tr>
</thead>
<tbody>
<tr>
<td>1960</td>
<td>1E-3</td>
</tr>
<tr>
<td>1970</td>
<td>1E-2</td>
</tr>
<tr>
<td>1980</td>
<td>1E-1</td>
</tr>
<tr>
<td>1990</td>
<td>1E+0</td>
</tr>
<tr>
<td>2000</td>
<td>1E+1</td>
</tr>
<tr>
<td>2010</td>
<td>1E+2</td>
</tr>
<tr>
<td>120GXP</td>
<td>1E+3</td>
</tr>
</tbody>
</table>

IBM RAMAC (First Hard Disk Drive)

IBM Disk Drive Products

Industry Lab Demos

1st AFC Media
1st GMR Head
1st MR Head
1st Thin Film Head
Microdrive II
Travelstar 40GN
Ultrastar 73LZX
100% CGR

~17 Million X Increase
25% CGR
60% CGR

~17 Million X Increase

Advanced Storage Roadmap

Atom Surface Density Limit

Nanotechnology Probe
Contact Area Limit

Superparamagnetic Effect

Enhanced Magnetic HDD

Holography Probe-like Storage

Atom Level Storage

Industry Lab Demos
3.5 Inch FF
2.5 Inch/1.0 Inch FF
>10 Inch FF

Areal Density, Gbits/in²

Availability Year
Magnetic Recording Basics

- Direction of Disk Motion
- Inductive Write Element
- GMR Read Sensor
- Magnetic Transition
- Grain Structure
- Track of Recording Media
- W, B, t
Areal Density and Media Grain Size to Maintain ~1000 grains/bit

- 800 nm, 80 nm: 12 Gbits/in$^2$, bpi/tpi = 10
- 380 nm, 50 nm: 35 Gbits/in$^2$, bpi/tpi = 8
- 40 nm, 160 nm: 100 Gbits/in$^2$, bpi/tpi = 4
Grain Structures in Magnetic Media

Areal density ~10 Gbits/in²

Areal density ~25 Gbits/in²

Magnification = 1 million
Particle energy $E_{\text{particle}} \propto \text{volume of grain}$

Thermal stability requires that $E_{\text{particle}} > 55k_B T$ to store information for $>10$ years

Superparamagnetic effect
Antiferromagnetically Coupled (AFC) Media Structure

- CoPtCrB Top Magnetic Film
- Ruthenium Film 6 Å
- CoPtCrB Bottom Magnetic Film
AFC Media Stability

Amplitude loss after 10 years

Magnetic thickness \( \text{Mrt (memu/cm}^2) \)

Single layer media

AFC media

"Pixie Dust"

Self Assembly for the Future

Ultra High Density Magnetic Recording

Single magnetic domain per bit
- Perpendicular media

Challenge
- Regular array over large area
- Adequate magnetic field to write particles

100 Gbit/in²
- ~135 particles
  4:1

1 Tbit/in²
- ~13 particles
  1:1

13 Tbit/in²
- ~1 particle
  1:1

Lithography Challenges

SEMICONDUCTOR AND THIN FILM HEAD FEATURE SIZES

Increases in areal density are achieved with smaller head feature size. Areal density and lithography projections imply a cross over in head and IC minimum features.

Magnetic Tunnel Junction

- First ferromagnetic electrode acts as spin filter
- Second FM layer acts as spin detector
Microactuator Technology for Track Following and Servoing

Dual Stage Actuator using MEMS Technology

Actuator Motion

Actuator Motor Coil

Suspension

Slider

Wiring Traces

Tracks From The Recording Media

Microactuator

GMR Head

Tape has Great Headroom for Growth

- Tape uses same basic magnetic recording technology as HDD
- Tape areal density is much lower than HDD

<table>
<thead>
<tr>
<th></th>
<th>HDD</th>
<th>Tape</th>
<th>Ratio (HDD/Tape)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bits per inch</td>
<td>530,000</td>
<td>130,000</td>
<td>4</td>
</tr>
<tr>
<td>Track per inch</td>
<td>64,000</td>
<td>900</td>
<td>70</td>
</tr>
<tr>
<td>Areal Density (Gb/in²)</td>
<td>34</td>
<td>0.1</td>
<td>280</td>
</tr>
</tbody>
</table>
Storage Infrastructure

Cost:
- dollars
- pennies
- pennies
- fractions

Performance:
- nanoseconds
- milliseconds
- seconds
- minutes

Production Site:
- Memory
- Disk
- Virtual Tape
- Physical Tape

Backup Site:
- Memory
- Disk
- Virtual Tape
- Physical Tape

Fabric
Estimated Relative Price Trends

- JBOD prices do not include RAID
- JBOD Unix Disk
- JBOD Windows Disk
- Open Tape Library

Source: Various IBM and Industry Studies
An Approach to Data Preservation


Joint study with the Koninklijke Bibliotheek (Dutch National Library)
Preserving Programs as Well as Data


For example: the cost to manage storage is typically twice the cost of the actual storage system.


1984

$2 million System

$1 million Storage Administration

2000

$1 million System

$2 million Storage Administration

Causes of Unplanned Application Downtime

- Operator Errors: 20%
- Application Failures: 40%
- Technology Failures: 40%

**Making the Front Page**

- **eBay**
  - Outage: 22 hours 12 June 1999
  - Operating System Failure
  - Cost: $3 million to $5 million revenue hit and 26% decline in stock price

- **AT&T**
  - 13 April 1998 outage: Six to 26 hours
  - Software Upgrade
  - Cost: $40 million in rebates
  - Forced to file SLAs with the FCC (frame relay)

- **America Online**
  - 6 August 1996 outage: 24 hours
  - Maintenance/Human Error
  - Cost: $3 million in rebates
  - Investment: ???

- **E*Trade**
  - 3 February 1999 through 3 March 1999: Four outages of at least five hours
  - System Upgrades
  - Cost: ???
  - 22 percent stock price hit on 5 February 1999

- **Dev. Bank of Singapore**
  - 1 July 1999 to August 1999: Processing Errors
  - Incorrect debiting of POS due to a system overload
  - Cost: Embarrassment/loss of integrity; interest charges

- **Charles Schwab & Co.**
  - 24 February 1999 through 21 April 1999: Four outages of at least four hours
  - System Upgrades/Operator Errors
  - Cost: ???; Announced that it had made $70 million in new infrastructure investment.

- **NYSE**
  - June 8, 2001
  - >1700 stocks stopped trading for 90 minutes
  - Software Upgrade
  - Cost: ???

Various sources including Gartner Group
(R)evolution of User Experience
Autonomic Computing

http://www.research.ibm.com/autonomic
Self-defining: A system’s understanding of its make-up, parameters and connections with other systems.
Self-defining

**Self-configuring and Self-optimizing:** The system’s ability to adjust to its configuration and resource allocation to achieve predetermined goals.
Autonomic Computing

Self-defining

Self-configuring and Self-optimizing

Self-healing and Self-protecting: The system’s ability to anticipate and respond to attacks and failures by reallocating workflow or shifting specific functions to achieve stability.
Autonomic Computing

Self-defining

Self-configuring and Self-optimizing

Self-healing and Self-protecting

Contextually Aware in a Heterogeneous Environment:
The system’s ability to work seamlessly with other systems and adjust its actions based on context.
Autonomic Computing

Self-defining

Self-configuring and Self-optimizing

Self-healing and Self-protecting

Contextually Aware in a Heterogeneous Environment:

Anticipatory: The system’s ability to anticipate workflow challenges and optimize the system for a user's immediate needs.
Levels of Autonomic Computing Sophistication

Serving the World (people, business processes)

Heterogeneous Components Interacting

Homogeneous Components Interacting

Components

Well known examples

SMS

SNMP

Adaptive network routing, Network congestion control

High availability clustering

ESS

RAID

DB Optimizer

Virus Management

Current Research and Product Directions

Storage Tank

Collective Intelligence Storage Bricks

Oceano

Regatta self-healing, LPAR

SMART/LEO

Software Rejuvenation

Future Goal

New packaging concepts for storage

Subscription computing

More of the same and better

Future Innovations
Storage Systems

Key Trends:

- Virtualization
- Self-Management
- Modularity
- Fail-in-place
- Policy Management

*Mandated by:* TCO, Availability and Ease of Use
*Enabled by:* increases in processor speed and disk areal density
The Move Towards Modularity

Monolithic
- Scaling is extremely coarse
- High Management costs
- High entry cost
- Very robust components
- Failure disruptions can be major
- Failed components repaired

Modular
- Scaling is fine grain
- Low Management costs
- Low entry cost
- Moderately robust components
- Failure disruptions are small
- Failed components not repaired
Fail-in-place

Goal is to reduce cost by increasing availability
- Can service actions be minimized or even eliminated?
  - Many service actions result from previous service actions

- Unavailability = \( \frac{MTTR}{MTTF} \)

- To achieve better availability:
  - \( \uparrow MTTF \) or \( \downarrow MTTR \)

Collective Intelligent Storage Bricks
- Overprovision the system
  - Seal the bricks
- Reliability Increases by...
  - Improved sparing
  - High levels of redundancy
  
  [Link: http://www.almaden.ibm.com/cs/storagesystems/CiB]
Self-Management

Eliminate Hot Spots

Add a disk, move data and balance

Proactive copies for hot spot elimination

Automatic Data Recovery

- Traditional RAID functions (parity, mirror, etc. ...)
- Copies can be used for higher levels of redundancy
Fail-in-place

Allows New Packaging Geometries

Various "Ice Cube" shapes

Brick, Cube, Node

IceCube Assembly

Ice Cube Prototype Brick
- SpecInt2000: 633
- Watt: 200
- Size: 20 cm = 7.87"

700 brick capacity

Slot for ‘cold rail’ at ground potential

Bi-directional ‘Coupler’ @ 10Gb/s

No wires, fibers, connectors, fans....

Collective Intelligent Bricks
IceCube ⇒ 1-PB

- 32 Racks
- 640 CIB
- 8 240-GB 3.5" Disks per CIB
- 275 W per CIB
- 5.5 kW per Rack

Total power = 250 kW
>75 dB air noise

31 kW

Water Chiller

http://www.almaden.ibm.com/cs/storagesystems/IceCube

- 640 CIB
- 8 240-GB 3.5" Disks per CIB
- 275 W per CIB

Total power = 220 kW
Quiet! (64 dB)

25 kW

Water Chiller
Bandwidth and Storage Virtualization

Example 10x10x10 IceCube ....
- Few Petabyte capacity
- Bisectional Bandwidth 6000 Gbits/s in each dimension
- External Bandwidth 4000 Gbits/s
  - Ports on four vertical walls
- Latency 130 nanoseconds per hop (only!)

Huge Bandwidth: Storage Virtualization (SV) becomes very practical
- Data can be distributed in cube nearly without regard to location
- Software for SV much easier to develop
- Most tasks of storage administrator go away
  - “just add more bricks” when SV software tells him/her

Result:
- 1 storage system administrator per 5 TB (today)

⇒ 1 administrator per Petabyte (in the future)
### Levels of Autonomic Computing Sophistication

#### Serving the World (people, business processes)
- SMS

#### Heterogeneous Components Interacting
- SNMP
  - Adaptive network routing, Network congestion control
  - High availability clustering

#### Homogeneous Components Interacting
- ESS
- RAID
- DB Optimizer
- Virus Management

### Future Innovations
- Oceano
  - Regatta self-healing, LPAR
  - SMART/LEO
  - Software Rejuvenation

### Current Research and Product Directions
- Storage Tank

### Future Goal
- New packaging concepts for storage
- Subscription computing
- More of the same and better
Policy Managed Storage:
Storage Tank

- Automated, policy-based storage and data management
- High performance, multi-platform file sharing

See Work-in-Progress talk this evening for more information on Storage Tank

www.almaden.ibm.com/cs/storagesystems/stortank
Storage Devices and Systems: Key Drivers of the IT Industry

Storage Devices Challenges

- Fundamental limits to be overcome
- Prospects for new technology, materials, etc...

Storage Systems Challenges

- Software for implementing policies automatically
- Intelligent modular hardware
- TCO
- Reliability
- User Experience

Serving business and societal needs
Thank you!