DATA STORAGE TECHNOLOGIES & NETWORKS

(CS C446, CS F446 & IS C446)

LECTURE 04 – STORAGE

Components of a Storage System Environment

- Host
  - CPU, internal memory and disk storage and I/O devices
- Connectivity
  - Interconnection between hosts or between a host and peripheral device
  - Physical components of connectivity (hardware) are bus, port and cable
  - Logical components of connectivity are protocols (PCI, IDE/ATA, SCSI)
- Storage

Data Explosion –Data Generation

- Data Generation –e.g. Mass General Ortho
- Massachusetts General Hospital, State of Massachusetts, U.S.A -Orthopaedic Department)
  - Serves 100 (new) patients a day // information on the Mass General website
  - 3 X-Ray images per patient // guesswork :-)
  - Images are stored in digitized form
  - Average X-ray image size is 40cm x 25cm
  - Resolution of images: 20 pixels per mm. // Fuji Film data

Data Explosion –Data Generation

- Each pixel has 8 grey levels (i.e. 3 bits per pixel)
  - // guesswork again :-)
- HIPAA (U.S. government) regulations -digital records must be stored for 7 years.
  - // Google for HIPAA
- Required size: approx. 1.15 Tera Bytes
  - Recalculate for changes
    - Color images (technicolor)? 3-D images?
    - Increasing resolution

Data Explosion –Data Generation

- Data Generation –e.g. Genome Database (simplistic)
- Simplest organism (bacterium) -approx. 600,000 base pairs
- Human or Mouse Genome -3 billion base-pairs
- Each base pair requires 2x2 bits
  - (two bits each for identifying A,G,C, or T and the counterpart)
- A database of genomes of 2000 complex organisms (say at least 1 billion base pairs)
- Total size: 1 Terabytes

Data Explosion –Data Replication

- Data Replication: Examples
  - Case Google Search: #web pages
    - 20 billion of (>25 billion) in 2008
    - 30 KB to 100KB each (say average, 50KB)
    - Total Space: Approx. 1 PetaBytes
    - Number of websites/pages is ever increasing
    - Google Library (Books)
    - Google Earth (Geo/Carto graphic Images)
  - Case Google Earth
    - Emerged (land) area of earth: 149 million square km
      - i.e. 149 x 10^{12} m^2
    - Resolution of GE: 15m per pixel (circa 2007)
    - # pixels: (150 x 10^{12})/15x15 = 0.67 x 10^{12}
    - Size of pixel: depth, color, texture –a few bytes
    - Total space: in Terabytes

Data Explosion

- Case Cuil –the latest search engine in 2009
  - 121 billion web pages in 2008
  - approx. 6 Peta Bytes
- Examples of data collection*:
  - Birth certificates by hospitals in the U.S.
    - 1983: 280 bytes, 1996: 1864 bytes
  - Grocery store purchase entry
    - 1983: 32 bytes, 1996: 1272 bytes
  - Reference: L. Sweeney, Information Explosion. *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, L. Zayatz, P. Doyle, J. Theeuwes and J. Lane (eds), Urban Institute, Washington, DC, 2001.
- More generally,
  - Businesses collecting more information 24x365
  - New (automated) technologies for data collection
    - e.g. RFID, http request mining and cookies, mining of emails

Data Explosion –Data Replication

- Websites are replicated for distribution:
  - Consider Yahoo!, Time Warner, Google or MSN
    - Sites with the largest number of visitors
    - 100+ million per month (for each site)
  - Problem: Not all users are in the same (geo) location.
    - Solution: Replicate and Distribute –
      - Referred to as “Akamaization” after Akamai Corp.
    - Copies of frequently referred sites are cached close to the user
  - Data Explosion: Billions of pages replicated in dozens!

Data Explosion – Other Issues

- Data quality (resolution)
  - Refer to example data collections
- Data Availability and Access
  - Copies/Replicas for simultaneous access or low latency access
    - All major websites are “Akamaized”
  - Copies/Replicas for fault tolerance / disaster recovery
    - Companies were up and running within a few hours after WTC collapse
- Data Provenance
  - Tracking of origin and lifecycle (i.e history) of data

Data Explosion – Other Issues

- Data Permanence
  - Regulatory Compliance
    - HIPAA (U.S. Govt.) requires 7 years of data to be stored by hospitals
    - Sarbanes-Oaxley (U.S. Govt.) requires documentation on corporate governance
      - i.e. all corporate decisions / deliberations must be recorded i.e. all emails must be saved (forever??)
- Low cost storage
  - Leads to Increased access to storage
  - Cost per byte on hard disk << Cost per byte on paper
    - 2TB off-the-shelf hard disk costs around Rs. 5,000
    - 1 MB per 0.5 paise or 1GB per Rs 5