Future of Database Systems

University of California, Berkeley

School of Information

IS 257: Database Management

Lecture Outline

- Future of Database Systems
- Predicting the future…
    - Quotes from Leon Kappelman "The future is ours" CACM, March 2001
- Accomplishments of database research over the past 30 years
- Next-Generation Databases and the Future

- Radio has no future, Heavier-than-air flying machines are impossible. X-rays will prove to be a hoax.
  - William Thompson (Lord Kelvin), 1899

- This "Telephone" has too many shortcomings to be seriously considered as a means of communication. The device is inherently of no value to us.
  – Western Union, Internal Memo, 1876

- I think there is a world market for maybe five computers
  - Thomas Watson, Chair of IBM, 1943

- The problem with television is that the people must sit and keep their eyes glued on the screen; the average American family hasn't time for it.
  - – New York Times, 1949

- Where … the ENIAC is equipped with 18,000 vacuum tubes and weighs 30 tons, computers in the future may have only 1000 vacuum tubes and weigh only 1.5 tons
  – Popular Mechanics, 1949

- There is no reason anyone would want a computer in their home.
  – Ken Olson, president and chair of Digital Equipment Corp., 1977.

- 640K ought to be enough for anybody.
  - Attributed to Bill Gates, 1981

- By the turn of this century, we will live in a paperless society.
  – Roger Smith, Chair of GM, 1986

- I predict the internet… will go spectacularly supernova and in 1996 catastrophically collapse.
  – Bob Metcalfe (3-Com founder and inventor of ethernet), 1995

Lecture Outline

- Review
  - Object-Oriented Database Development
- Future of Database Systems
- Predicting the future…
  - Quotes from Leon Kappelman "The future is ours" CACM, March 2001
- Accomplishments of database research over the past 30 years
- Next-Generation Databases and the Future

Database Research

- Database research community less than 40 years old
- Has been concerned with business type applications that have the following demands:
  - Efficiency in access and modification of very large amounts of data
  - Resilience in surviving hardware and software errors without losing data
  - Access control to support simultaneous access by multiple users and ensure consistency
  - Persistence of the data over long time periods regardless of the programs that access the data
- Research has centered on methods for designing systems with efficiency, resilience, access control, and persistence and on the languages and conceptual tools to help users to access, manipulate and design databases.

Accomplishments of DBMS Research

- DBMS are now used in almost every computing environment to create, organize and maintain large collections of information, and this is largely due to the results of the DBMS research community’s efforts, in particular:
  - Relational DBMS
  - Transaction management
  - Distributed DBMS

Relational DBMS

- The relational data model proposed by E.F. Codd in papers (1970-1972) was a breakthrough for simplicity in the conceptual model of DBMS.
- However, it took *much* research to actually turn RDBMS into realities.

Relational DBMS

- During the 1970's database researchers:
  - Invented high-level relational query languages to ease the use of the DBMS for end users and applications programmers.
  - Developed Theory and algorithms needed to optimize queries into execution plans as efficient and sophisticated as a programmer might have custom designed for an earlier DBMS

Relational DBMS

- Developed Normalization theory to help with database design by eliminating redundancy
- Developed clustering algorithms to improve retrieval efficiency.
- Developed buffer management algorithms to exploit knowledge of access patterns
- Constructed indexing methods for fast access to single records or sets of records by values
- Implemented prototype RDBMS that formed the core of many current commercial RDBMS

Relational DBMS

- The result of this DBMS research was the development of commercial RDBMS in the 1980's
- When Codd first proposed RDBMS it was considered theoretically elegant, but it was assumed only toy RDBMS could ever be implemented due to the problems and complexities involved. Research changed that.

Transaction Management

- Research on transaction management has dealt with the basic problems of maintaining consistency in multi-user high transaction database systems

Transaction Management

- To guarantee that a transaction transforms the database from one consistent state to another requires:
  - The concurrent execution of transactions must be such that they appear to execute in isolation.
  - System failures must not result in inconsistent database states. Recovery is the technique used to provide this.

Distributed Databases

- The ability to have a single “logical database” reside in two or more locations on different computers, yet to keep querying, updates and transactions all working as if it were a single database on a single machine
- How do you manage such a system?
  – It has to become part of the DBMS itself, not left to the application layer

Lecture Outline

- Review
  - – Object-Oriented Database Development
- Future of Database Systems
- Predicting the future…
  - Quotes from Leon Kappelman "The future is ours" CACM, March 2001
- Accomplishments of database research over the past 30 years
- "Next-Generation Databases" and the Future

Next Generation Database Systems

- Where are we going from here?
  - Hardware is getting faster and cheaper
  - DBMS technology continues to improve and change
    - OODBMS
    - ORDBMS
    - NoSQL
    - NewSQL
  - Bigger challenges for DBMS technology
    - Medicine, design, manufacturing, digital libraries, sciences, environment, planning, etc...
    - Sensor networks, streams, etc…
- The Claremont Report on DB Research
  - Sigmod Record, v. 37, no. 3 (Sept 2008)

Examples

- NASA EOSDIS
  – Estimated 10^{16} Bytes (Exabyte)
- Computer-Aided design
- The Human Genome
- Department Store tracking
  – Mining non-transactional data (e.g. Scientific data, text data?)
- Insurance Company
  – Multimedia DBMS support

New Features

- New Data types
- Rule Processing
- New concepts and data models
- Problems of Scale
- Parallelism/Cloud-based DB
- Tertiary Storage vs Very Large-Scale Disk Storage vs Large-Scale semiconductor Storage
- Heterogeneous Databases
- Memory Only DBMS

Coming to a Database Near You…

- Browsibility
- User-defined access methods
- Security
- Steering Long processes
- Federated Databases
- IR capabilities
- XML
- The Semantic Web(?)

The next-generation DBMS

- What can we expect for a next generation of DBMS?
- Look at the DB research community – their research leads to the “new features” in DBMS
- The “Claremont Report” (2008) on DB research was a report of meeting of top researchers and what they thought were the interesting and fruitful research topics for the future

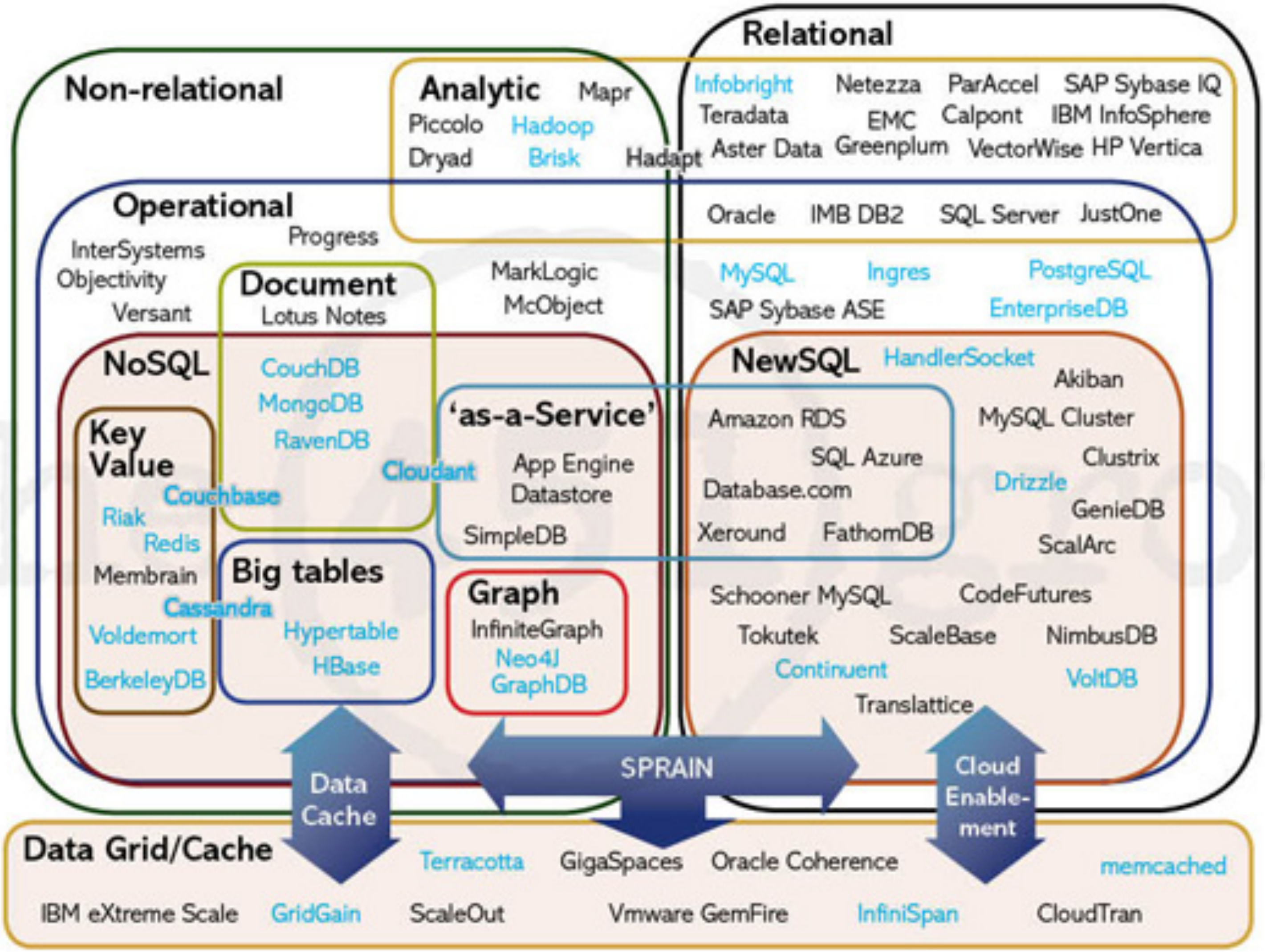
But will it be a RDBMS?

- As we saw last time…
- Recently, Mike Stonebraker (one of the people who helped invent Relational DBMS) has suggested that the "One Size Fits All" model for DBMS is an idea whose time has come – and gone
  – This was also a theme of the Claremont Report
- RDBMS technology, as noted previously, has optimized on transactional business type processing
- But many other applications do not follow that model

Will it be an RDBMS?

- Stonebraker predicted that the DBMS market will fracture into many more specialized database engines
  - Although some may have a shared common frontend
- Examples are Data Warehouses, Stream processing engines, Text and unstructured data processing systems
- We are seeing this with NoSQL, NewSQL and Stream processors

The Database Universe 2013



Will it be an RDBMS?

- Data Warehouses currently use (mostly) conventional DBMS technology
  - But they are **NOT** the type of data those are optimized for
  - Storage usually puts all elements of a row together, but that is an optimization for updating and not searching, summarizing, and reading individual attributes
  - A better solution is to store the data by column instead of by row – vastly more efficient for typical Data Warehouse Applications

Will it be an RDBMS?

- Streaming data, such as Wall St. stock trade information is badly suited to conventional RDBMS (other than as historical data)
  - The data arrives in a continuous real-time stream
  - But, data in old-school RDBMS has to be stored before it can be read and actions taken on it
    - This is too slow for real-time actions on that data
  - Stream processors function by running "queries" on the live data stream instead
    - May be *orders of magnitude* faster
  - Some NewSQL systems also play in this space, like VoltDB that we looked at last time, by reducing RDBMS overhead and running in memory

Will it be an RDBMS?

- Sensor networks provide another massive stream input and analysis problem
- Text Search: No current text search engines use RDBMS, they too need to be optimized for searching, and tend to use inverted file structures instead of RDBMS storage
- Scientific databases are another typical example of streamed data from sensor networks or instruments

XML Databases

- XML data is still not a first-class citizen of RDBMS, and there are reasons to believe that specialized database engines are needed
- Some XML Databases are also billed as NoSQL
  - MarkLogic – which has gotten part of the blame for the implementation fiasco surrounding HealthCare.gov :
    - "Another sore point was the Medicare agency's decision to use database software, from a company called MarkLogic, that managed the data differently from systems by companies like IBM, Microsoft and Oracle. CGI officials argued that it would slow work because it was too unfamiliar. Government officials disagreed, and its configuration remains a serious problem." New York Times, 11/22/2013

Will it be an RDBMS

- RDBMS will still be used for what they are best at – business-type high transaction data
  - But they may be memory-based systems like VoltDB
- But specialized DBMS will be used for many other applications
  - The stream bases, document-oriented dbs, etc.
- Consider Oracle's acquisions of SleepyCat (BerkeleyDB) embedded database engine, and TimesTen main memory database engine
  - specialized database engines for specific applications

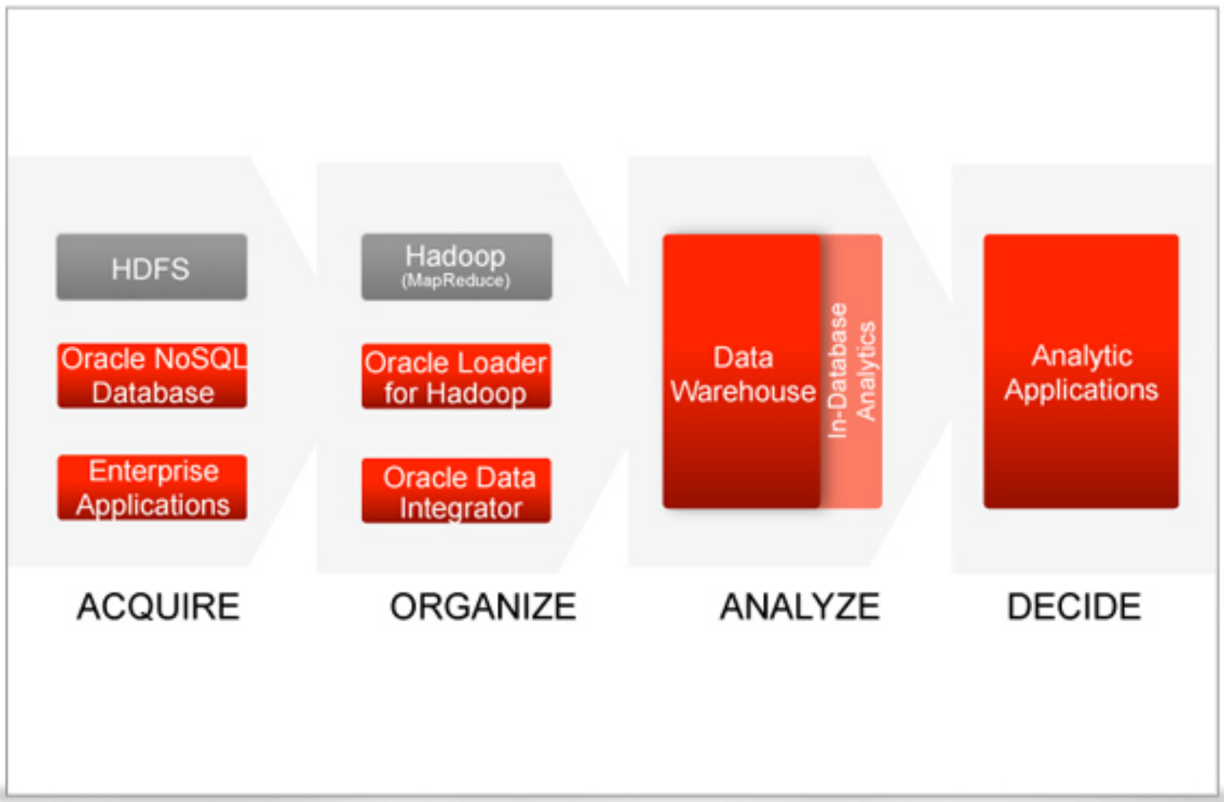
You can buy Big Data…

- Oracle will be happy to sell you systems (hardware and software) to manage your exabytes…

Oracle Exadata Database Machine X3-8

And NoSQL too…

- Oracle “Big Data Appliance”
- With Oracle NoSQL Database (BerkeleyDB)



The Claremont Report 2008

- The group saw a "Turning Point in Database Research"
  - Current Environment
  - Research Opportunities
  - Moving Forward

Current Environment

- "Big Data" is becoming ubiquitous in many fields
  - enterprise applications
  - Web tasks
  - E-Science
  - Digital entertainment
  - Natural Language Processing (esp. for Humanities applications)
  - Social Network analysis
  - Etc.

Current Environment

- Data Analysis as a profit center
  - No longer just a cost – may be the entire business as in Business Intelligence

Current Environment

- Ubiquity of Structured and Unstructured data
  - Text
  - XML
  - Web Data
  - Crawling the Deep Web
- How to extract useful information from "noisy" text and structured corpora?

Current Environment

- Expanded developer demands
  - Wider use means broader requirements, and less interest from developers in the details of traditional DBMS interactions
- Architectural Shifts in Computing
  - The move to parallel architectures both internally (on individual chips)
  - And externally – Cloud Computing/Grid Computing

Research Opportunities

- Revisiting Database Engines
  - Do DBMS need a redesign from the ground up to accommodate the new demands of the current environment?
  - Should the “elephants” be discarded completely or do they have a place?

Research Opportunities-DB engines

- Designing systems for clusters of many-core processors
- Exploiting RAM and Flash as persistent media, rather than relying on magnetic disk
  - Including NVRAM with ram speed and persistent data
- Continuous self-tuning of DBMS systems
- Encryption and Compression
- Supporting non-relation data models
  - instead of “shoe-horning” them into tables

Research Opportunities-DB engines

- Trading off consistency and availability for better performance and scaleout to thousands of machines
- Designing power-aware DBMS that limit energy costs without sacrificing scalability

Research Opportunities-Programming

- Declarative Programming for Emerging Platforms
  - MapReduce
  - Ruby on Rails
  - Workflows

Research Opportunities-Data

- The Interplay of Structured and Unstructured Data
  - Extracting Structure automatically
  - Contextual awareness
  - Combining with IR research and Machine Learning

Research Opportunities - Cloud

- Cloud Data Services
  - New models for “shared data” servers
  - Learning from Grid Computing
    - SRB/IRODS, etc.
  - Hadoop - as mentioned earlier - is open source and freely available software from Apache for running massively parallel computation (and distributed storage)

Research Opportunities - Mobile

- Mobile Applications and Virtual Worlds
  - Need for real-time services combining massive amounts of user-generated data

Moving forward

- Establishing large-scale collaborative projects to address these research opportunities
- What will be the result?

Some things to consider

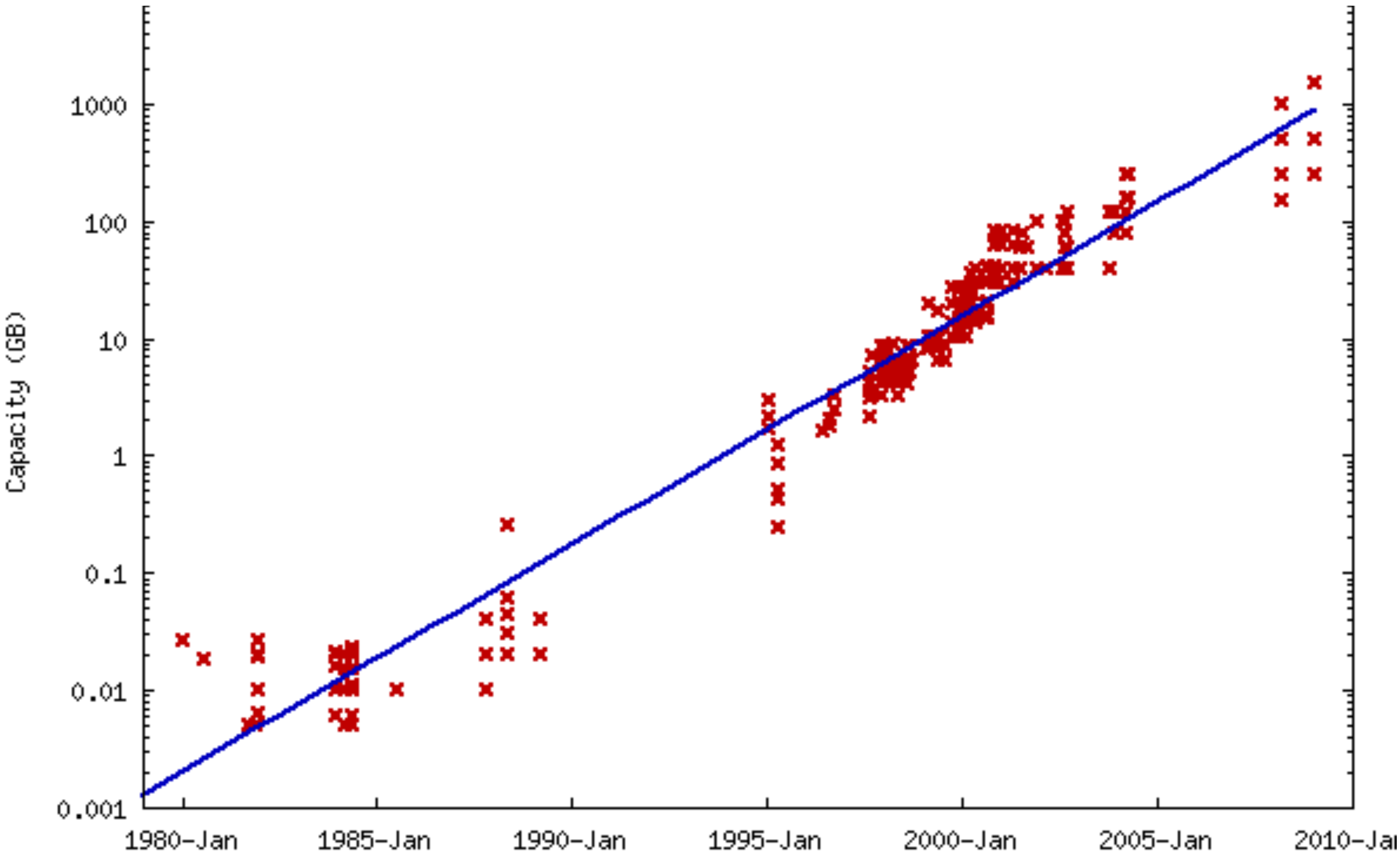
- Bandwidth will keep increasing and getting cheaper (and go wireless) (?)
- Processing power will keep increasing (?)
  - Moore's law: Number of circuits on the most advanced semiconductors doubling every 18 months
  - With multicore chips, all computing is becoming parallel computing
- Memory and Storage will keep getting cheaper (and probably smaller) (?)
  - "Storage law" or Kryder's Law: Worldwide digital data storage capacity has doubled every 9 months for the past couple of decades

- Put it all together and what do you have?
  – *“The ideal database machine would have a single infinitely fast processor with infinite memory with infinite bandwidth – and it would be infinitely cheap (free)” :* David DeWitt and Jim Gray, 1992
- Today it is more likely to be thousands of commodity machines running in parallel (using Hadoop or Spark, for example) with very fast networking between them (but it is definitely not free)
- **But will these assumptions hold into the future?**

Is the future all roses?

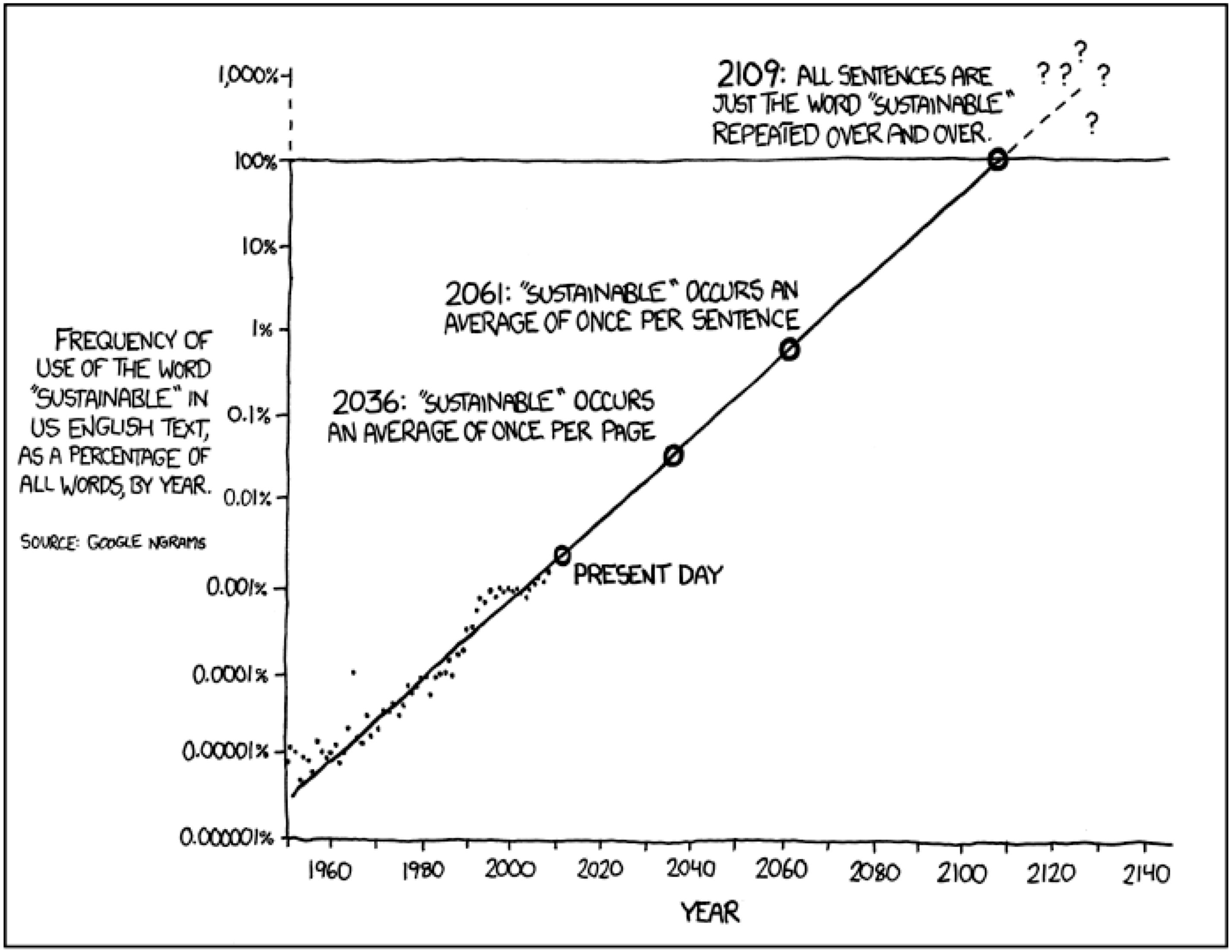
- Most future predictions start with the assumption that:
  - Moores Law will continue double chip complexity or density every 18 months
  - Disk will follow the same path by “Kryder’s Law” which says that the capacity doubles about every two years
- Consider the following analysis, thanks to David Rosenthal of LOCKSS on the issues with archival storage systems (remember Walmart’s Data Warehouse?)

Kryder's "Law"



Source: David Rosenthal - LOCKSS

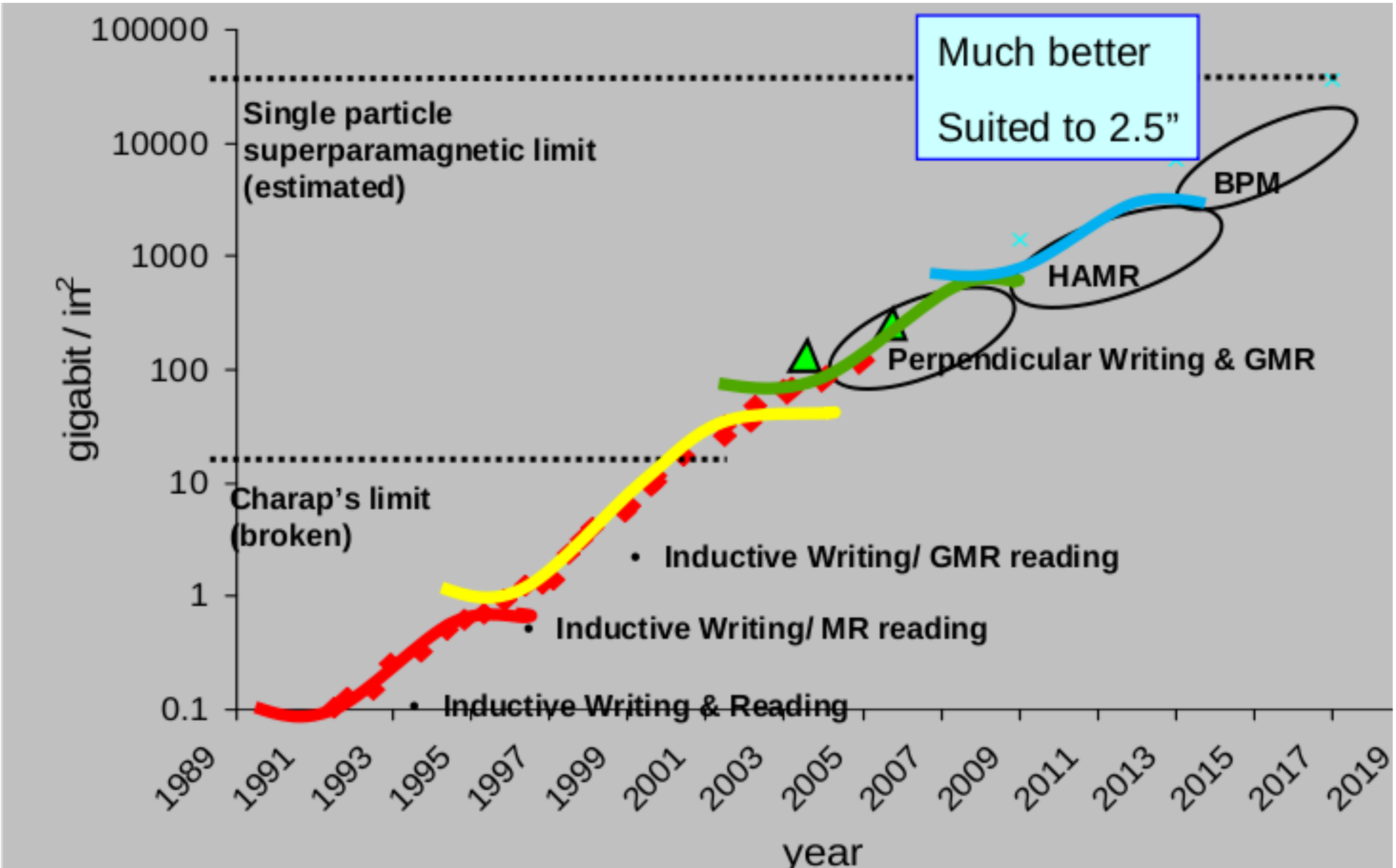
But is it sustainable?



THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.

From XKCD : http://xkcd.com/1007/

Disk Technologies and Scale



Source: Dave Anderson of Seagate - via David Rosenthal - LOCKSS

Kryder's Law?

- It was pretty clear by 2011 that the storage industry had fallen off the Kryder curve
- That was *BEFORE* the 2012 floods in Thailand destroyed 40% of the world's disk manufacturing capacity and doubled disk prices overnight
- Prices in 2013 were still about 60% more than they were before the floods and were not expected to return to pre-flood levels until 2014 (have they?)

Source: David Rosenthal - LOCKSS

The Crisis in three numbers

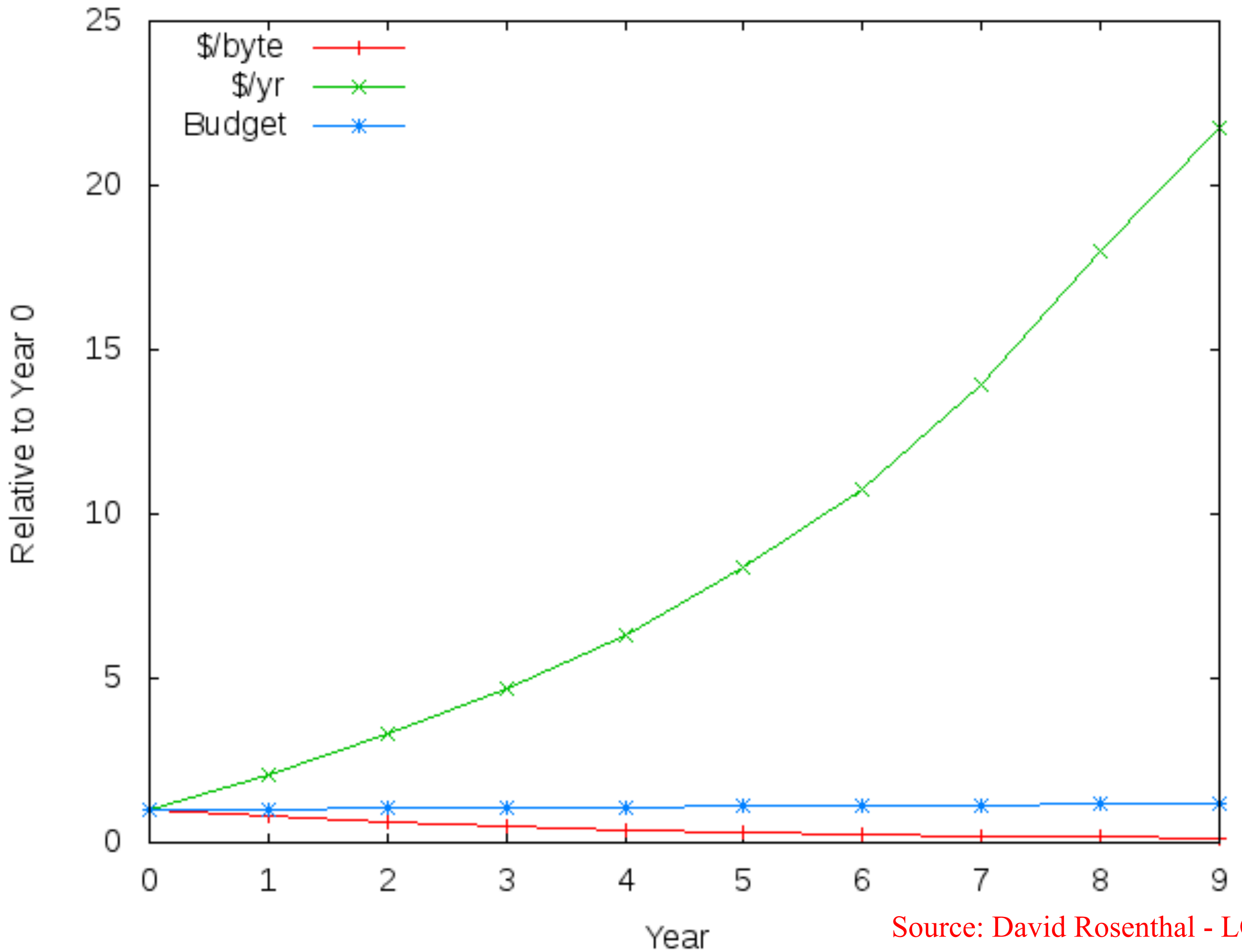
- According to IDC demand for storage grows about 60% per year
- Bit density on disk platters is expected to grow no more than 20% per year for the next 5 years
- Meanwhile IT budgets in recent years have only grown between 0% and 2% per year
- The graph on the next slide projects these numbers for 10 years…

Three Numbers

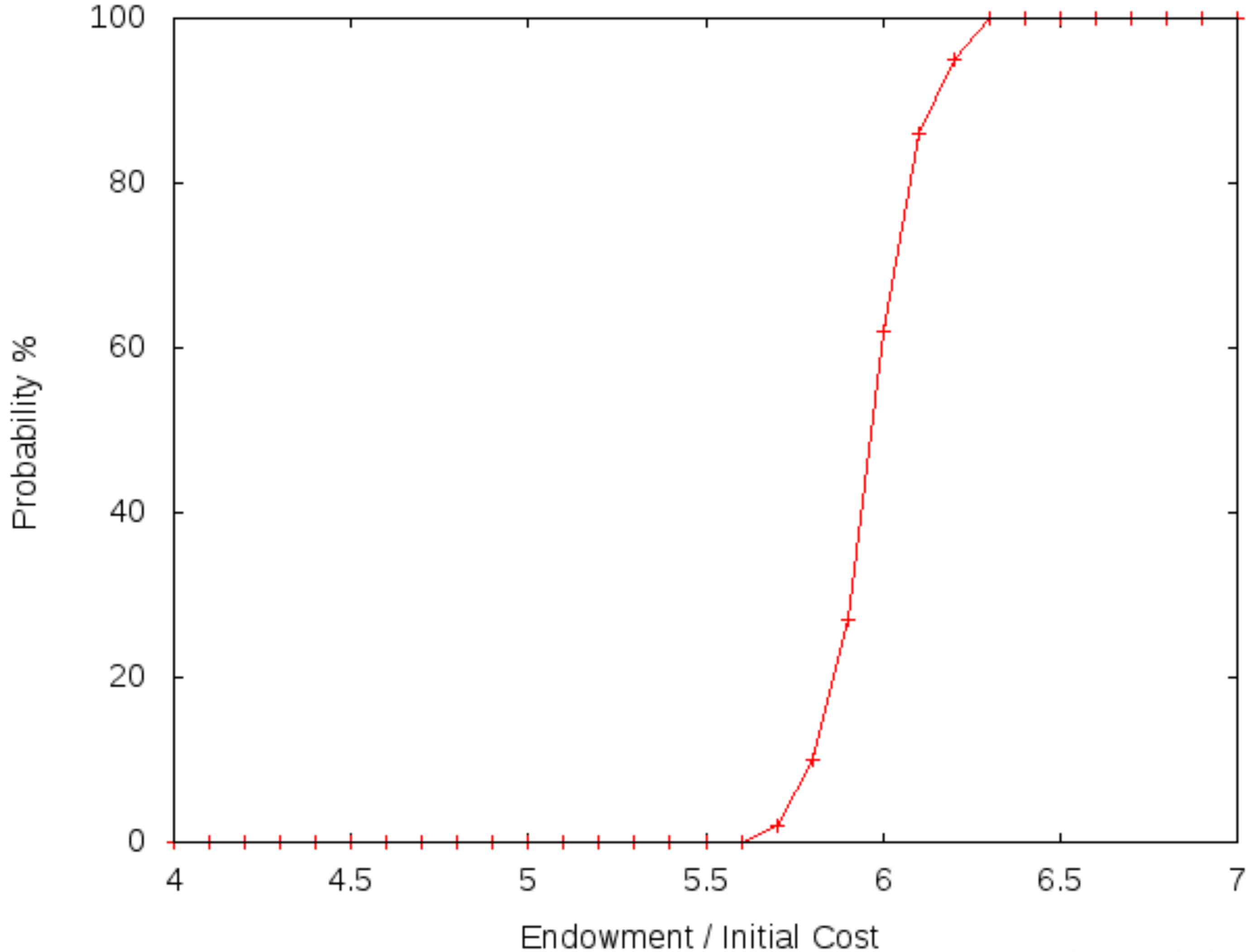
Red = Kryder @ 20%

Blue = IT Budget

Green = Cost of storage with 60% growth



Source: David Rosenthal - LOCKSS



Source: David Rosenthal - LOCKSS

Future DBMS

- What kinds of things can we expect to be incorporated in future DBMS?
- One way to look at it is to see what we are doing now – and just scale it up in size and speed
- Another is to consider what kinds of features would be useful for current "problem areas" or desired applications…

Future trends

- Integration (multiple formats, multiple databases)
- Accessibility (anytime, anywhere)
- Reliability (no data loss - ever)
- Proper display of data (advanced graphics methods dynamically applied)
- Real-time access
- Able to handle the 3 Vs – Volume, Velocity and Variety

Based on slides from Marina L. Gavrilova

Future trends in use

- Real-time decision support
- Information visualization and 3D modeling
- Tracking DB development on-line
- Data exchange between different applications/vendors
- Virtual and augmented reality
- Intelligent Decision support and cognitive methods for data analysis

Based on slides from Marina L. Gavrilova

Future Interfaces

- Scientific visualization tools
- Automated vision tools
- Fourth dimension
- Advanced spatial analysis tools (e.g. S+)
- Beyond the WIMP (windows, icons, menus and pointers) Desktop
  - High Interactivity
  - Multisensory input
  - Multisensory output
  - Touch free displays
  - Virtual and Augmented Reality

Based on slides from Marina L. Gavrilova

Revisiting the Assumptions

- Bandwidth will keep increasing and getting cheaper (and go wireless)
- Processing power will keep increasing
  - Moore’s law: Number of circuits on the most advanced semiconductors doubling every 18 months
  - With multicore chips, all computing is becoming parallel computing
- Memory and Storage will keep getting cheaper (and probably smaller)
  - “Kryder’s law”: Worldwide digital data storage capacity has doubled every 18-24 months for the past decade
  - E.g. Resistive Random Access Memory "RRAM” – 20 times faster and 10 x lower power – but persistent like flash
- **So what will the database environment of the future look like?**

?