Revolution in Storage

James Hughes

Fellow, VP, Chief Architect

Massive Storage

HUAWEI

Agenda

- **Economics**
- **Technology Shifts**
- **Open Questions**
- **Predictions**

Agenda

- **Economics**
- **Technology Shifts**
- **Open Questions**
- **Predictions**

Prediction is very difficult, especially if it's about the future.

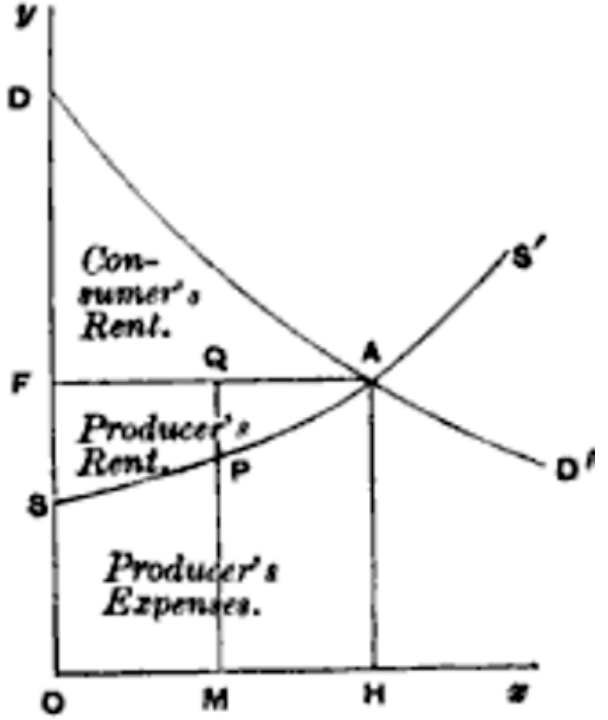
Niels Bohr

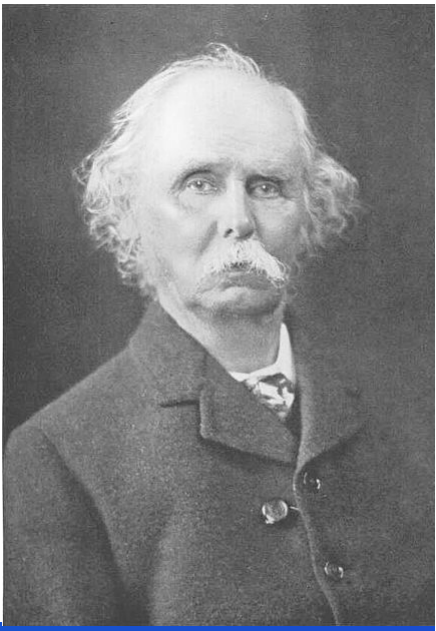
Demand for Storage

- **The demand to store more data is not slowing down**
  - Enabling new applications
    - Recording internet traffic
    - All CCTV surveillance for years
  - All human experience of 7B people is 1,000 EB
- **Recording less valuable information "just in case"**
  - The future value of information is not known
- **All predictions that demand for computing or storage will be satisfied have all failed over the years**

Storage is a Price Elastic Market

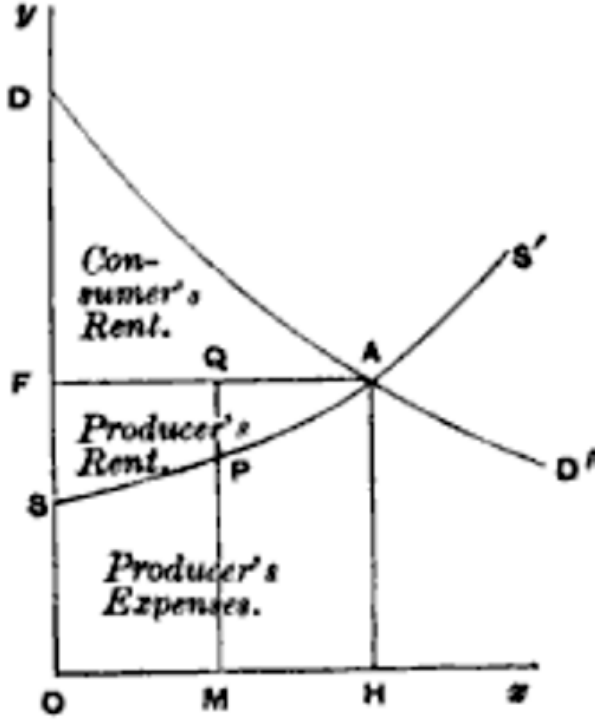
- **Price elasticity of demand**
  - Alfred Marshall (1890)

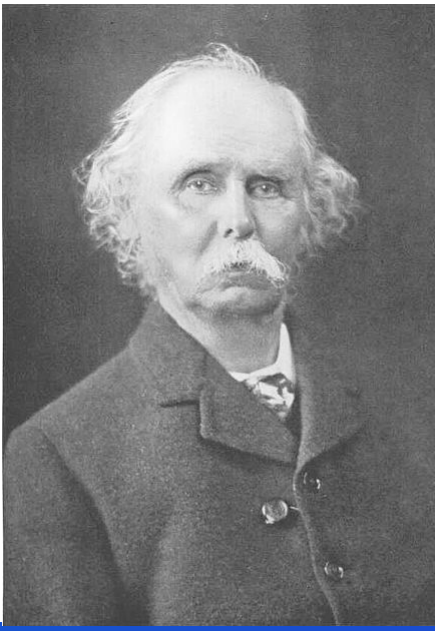




Storage is a Price Elastic Market

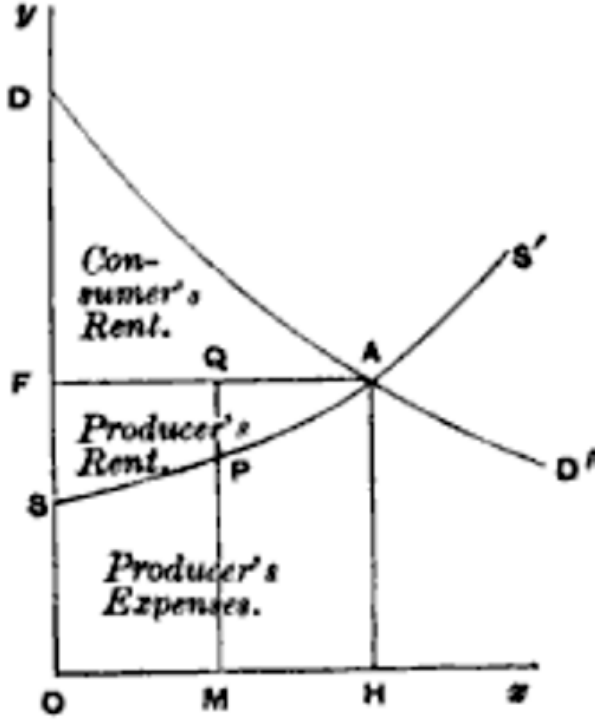
- **Price elasticity of demand**
  - Alfred Marshall (1890)
- **As the price of Storage approaches $0**
  - Demands for storage will approach infinity

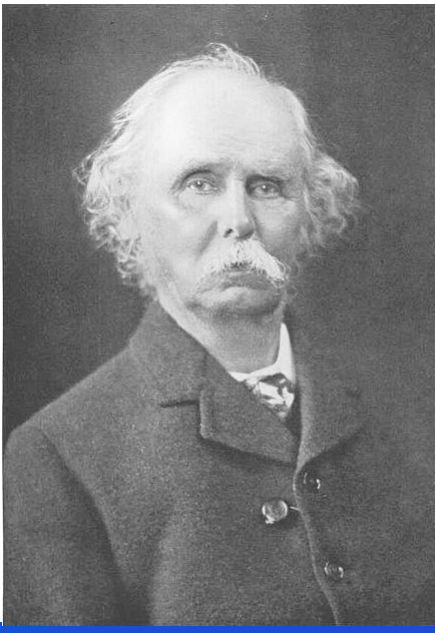




Storage is a Price Elastic Market

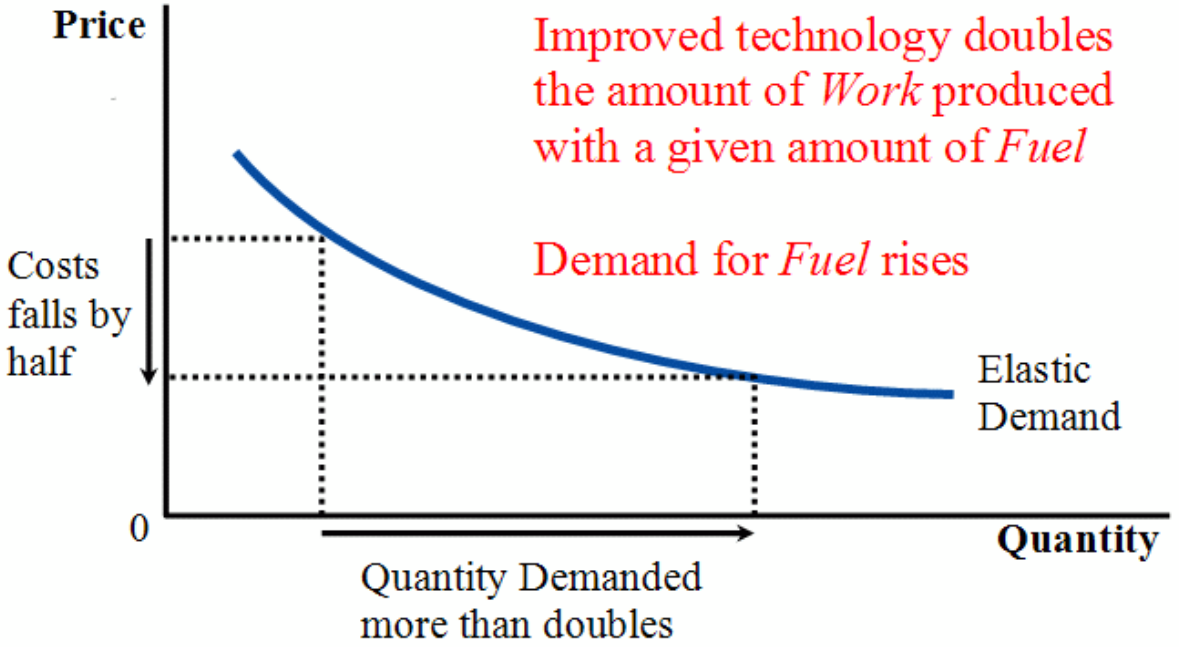
- **Price elasticity of demand**
  - Alfred Marshall (1890)
- **As the price of Storage approaches $0**
  - Demands for storage will approach infinity
- **If the price of a Cisco router approaches $0**
  - Demands for routers will ***not*** approach infinity





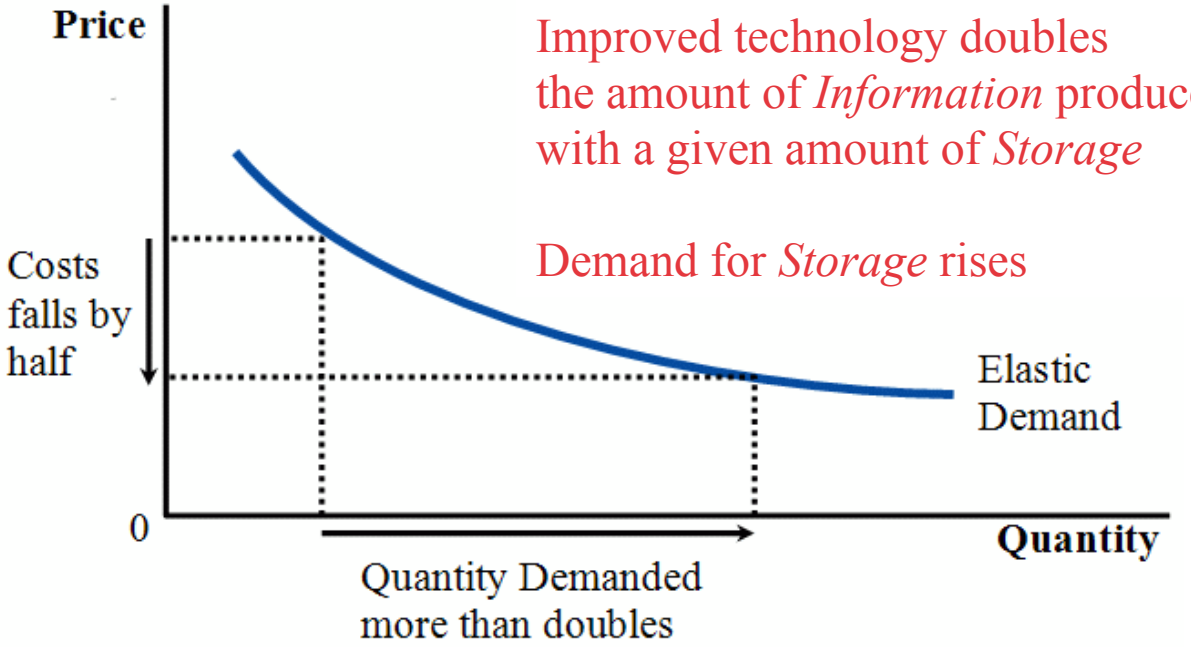
Cloud Computing will increase this trend

- **Jevons Paradox**
  - Cloud Computing increases the efficiency of computing....



Cloud Computing will increase this trend

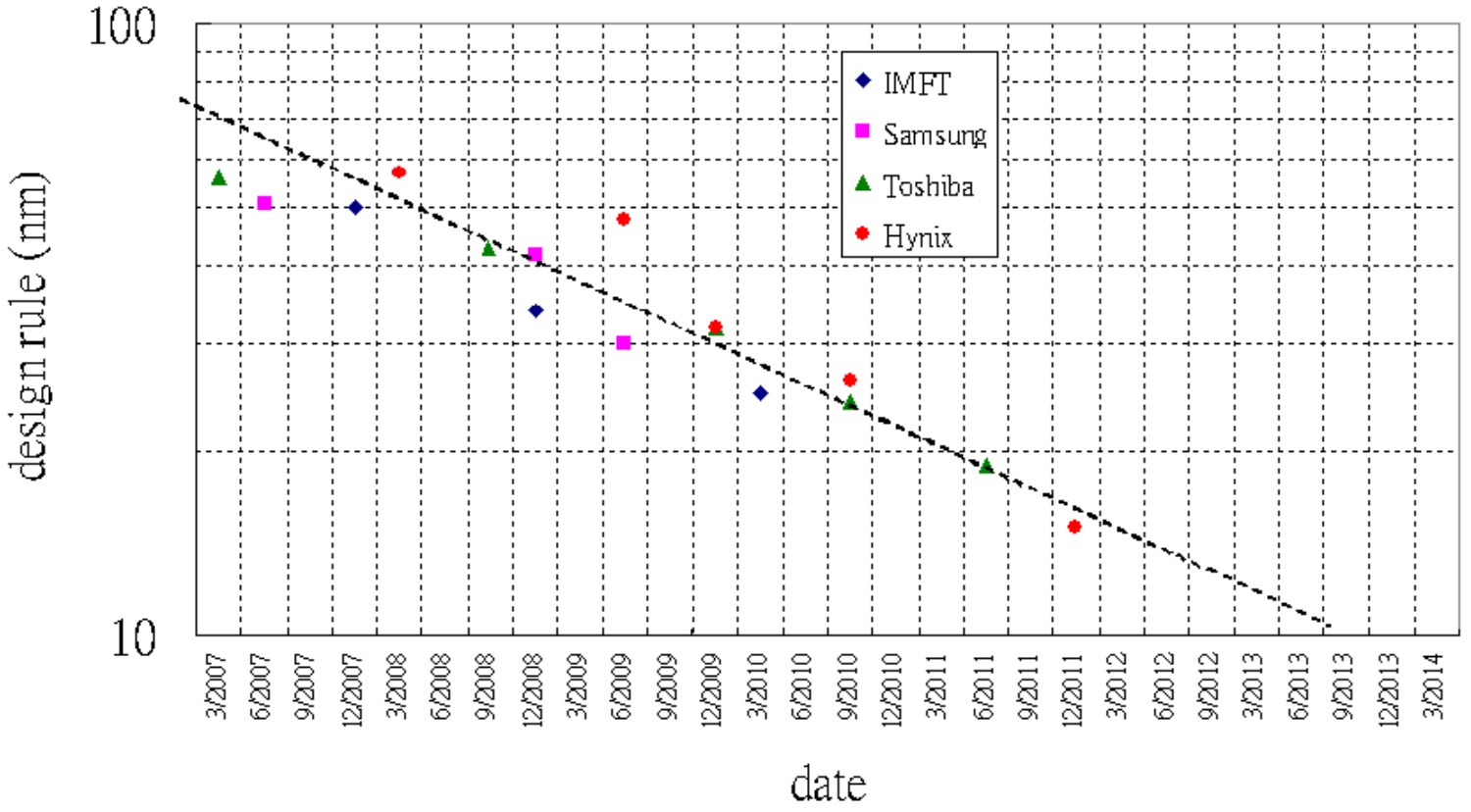
- **Jevons Paradox**
  - Cloud Computing increases the efficiency of computing....



Storage Technology

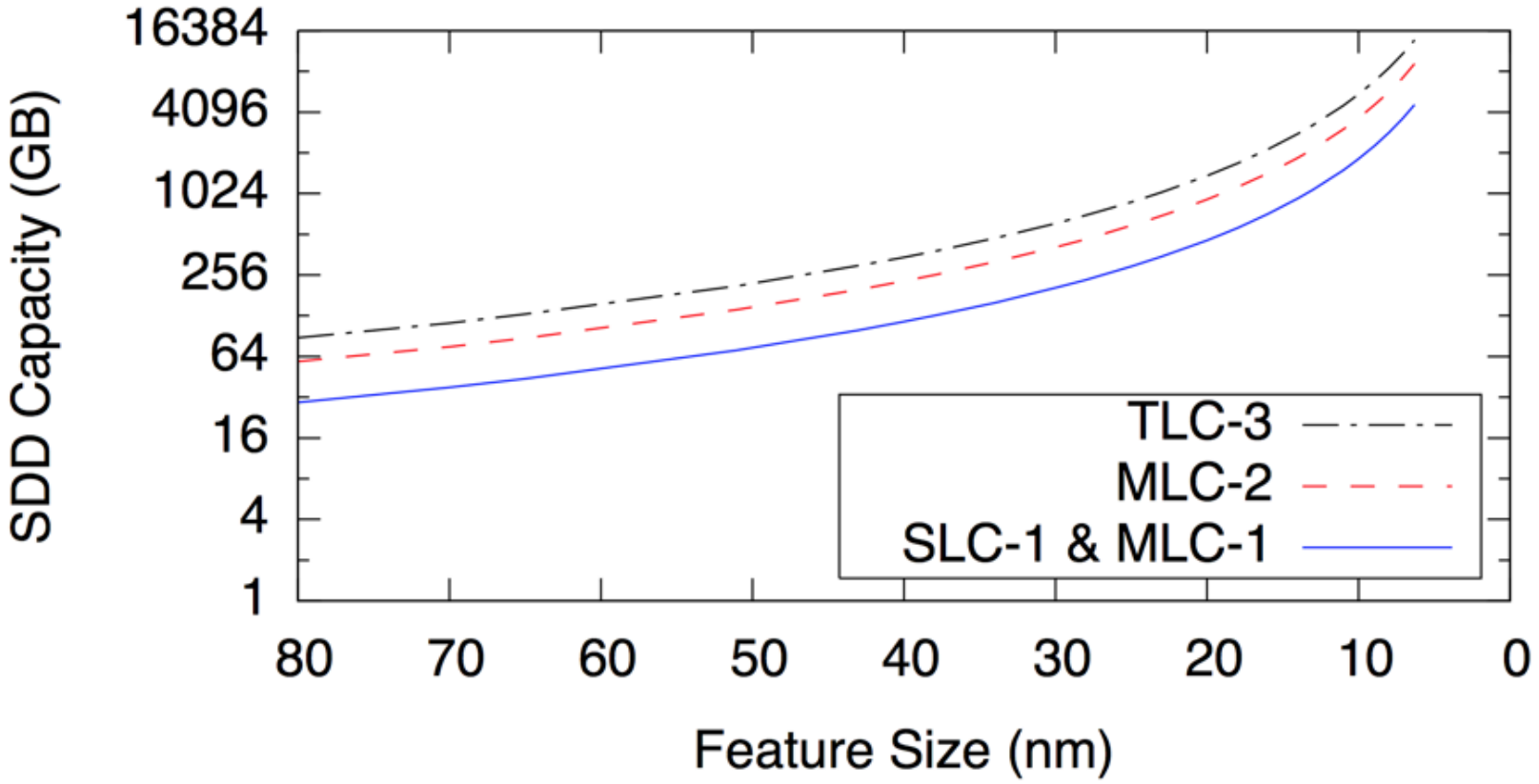
- **Flash Devices**
- **Shingled Disks**
- **Log Structure**
- **Distributed Hash Tables**
- **Metadata Servers (not)**
- **Object Storage**

Moore's Law for Flash Scaling



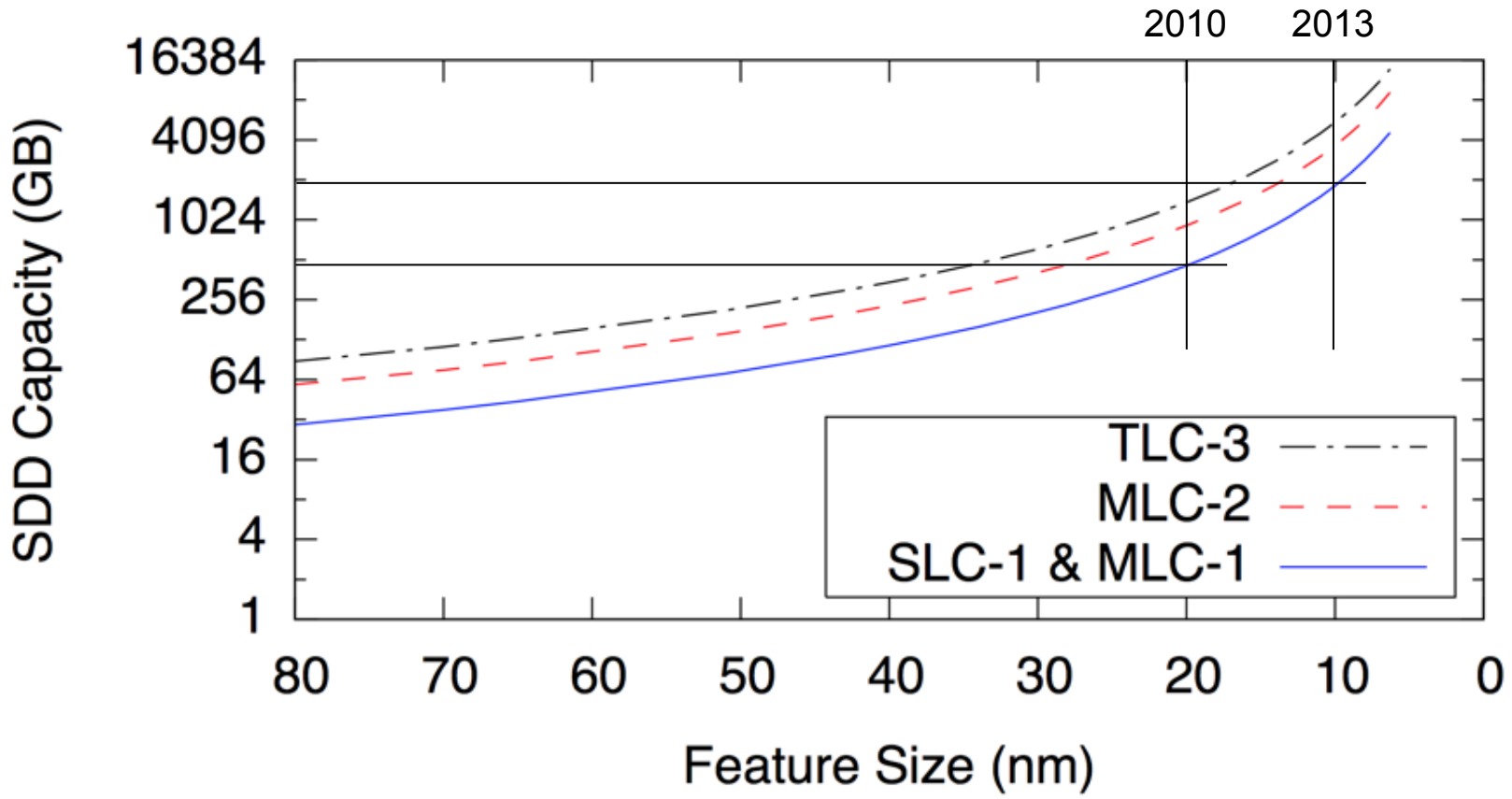
http://upload.wikimedia.org/wikipedia/commons/6/64/NAND_scaling_timeline.png

Flash Drive Density Forecast



http://cseweb.ucsd.edu/users/swanson/papers/FAST2012BleakFlash.pdf

Flash Drive Density Forecast

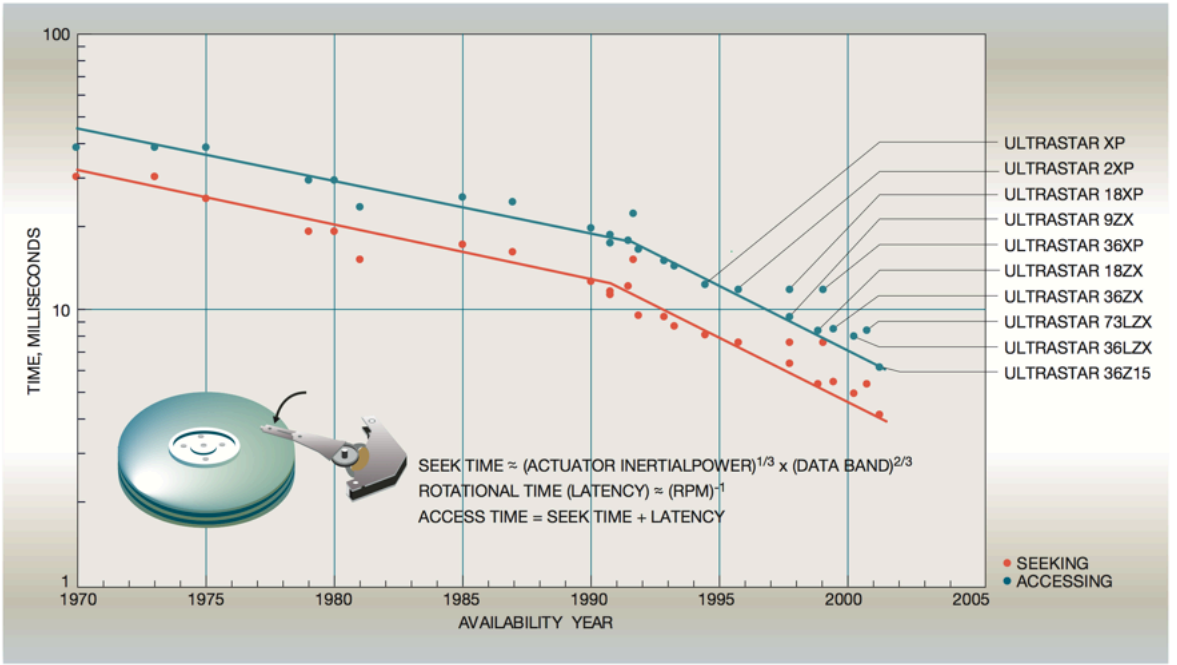


http://cseweb.ucsd.edu/users/swanson/papers/FAST2012BleakFlash.pdf

Disk Performance

- **Factor of 10x performance in 30 years**
  - Processors are 1,000,000x in 30 years

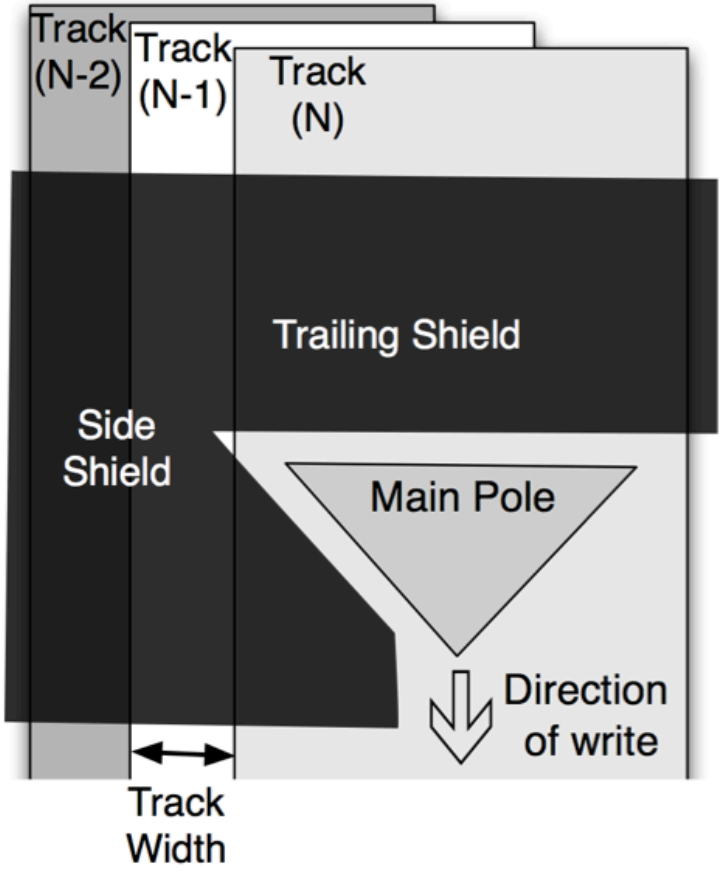
Figure 11 Disk drive access/seek times



http://www.cs.princeton.edu/courses/archive/spr05/cos598E/bib/grochowski.pdf

Shingled Disks

- **Write head larger than read head**
  - Turns Disk into a sequential media
- **All updates to data and metadata are written sequentially to a continuous stream, called a log**
- **Disk API of sectors is no longer "natural"**
  - One read may require several seeks



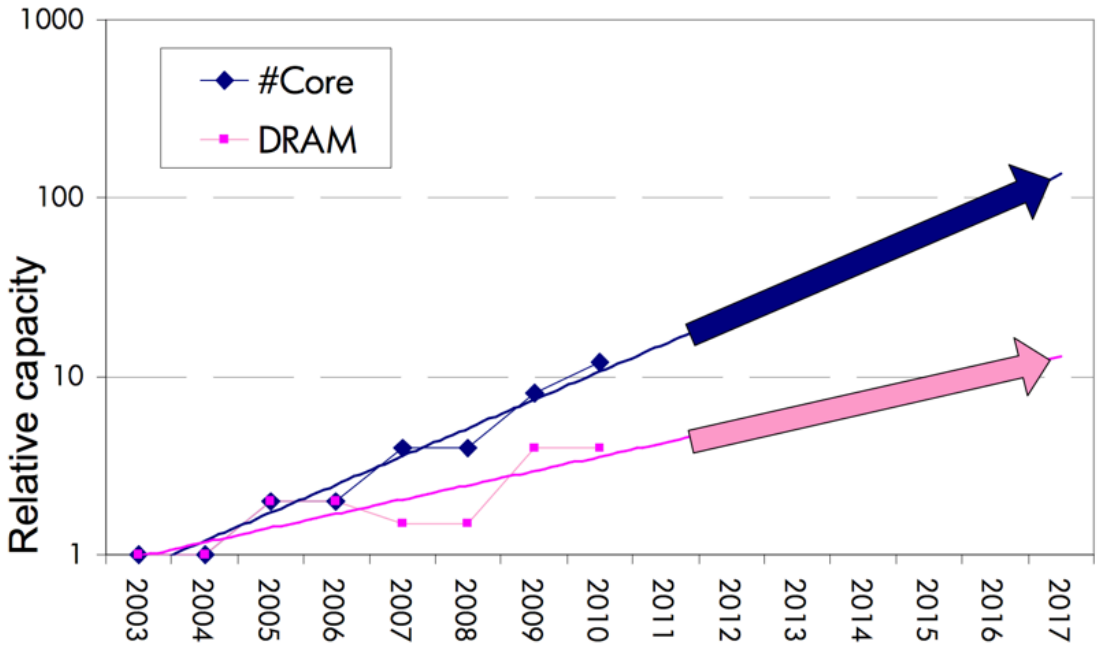
http://www.ssrc.ucsc.edu/Papers/amer-ieeetm11.pdf

Log Structured Storage

- **How much is erased on a reposition?**
  - Tape - the remainder of the tape
  - Singled disk - the remainder of the track group
  - Flash - the entire page
- **All persistent Storage systems do/will implement log structure**
  - e.g. “NoSQL Database of sectors”
- **Does it make sense to layer a database on top of a database?**
  - Could we use the log structure of the media to provide a more natural storage systems, not mimicking an antique paradigm?

Single System Performance Trend

- **Leading to disaggregation of servers**



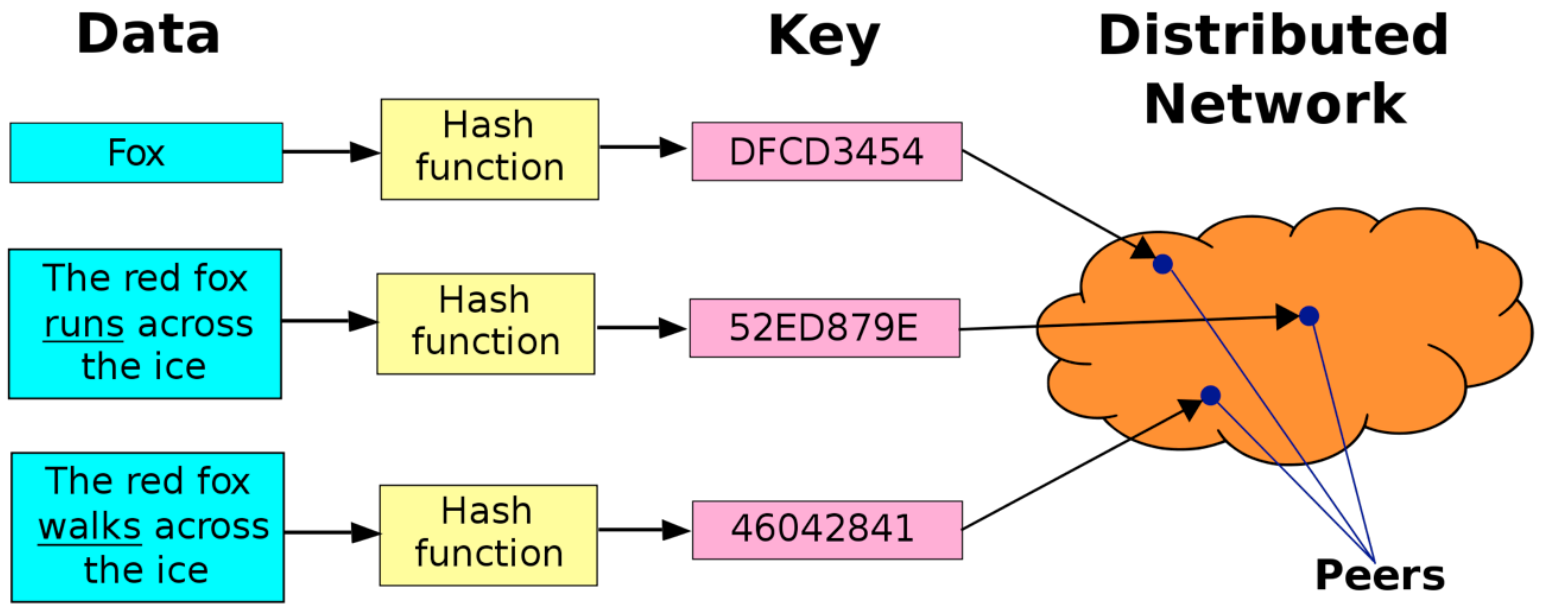
(a) Trends leading toward the memory capacity wall

http://web.eecs.umich.edu/~twenisch/papers/isca09-disaggregate.pdf

Scaling Storage

- **Distributed Hash Table**
  - Key/Value Store

RAM	Memcached
Flash	Voldemort
Disk	Cassandra



Metadata Servers

- **Required by traditional file systems (POSIX) to translate names to sectors**
  - Hard to scale, heavy HA requirements, expensive
- **Can we use a name as a key?**
  - Place the data into a scaled key value store?
  - Eliminate costly metadata servers?

Object Storage

- **A storage system where objects (files) are read, written, replaced, but never changed.**
  - e.g. Amazon S3
- **Allows log structure with a minimum of garbage collection**
- **New tier of storage**
  - Lowest cost for online storage (not tape)
  - Huge aggregate performance (High throughput, OK latency)

Open Questions

- **Should tiering decisions be automated or *better* left as an economic decision?**
  - Is the complexity worth it
- **Can Hadoop clusters be general purpose?**
  - Amdahl's Law
- **Is there a general paradigm for turning drives off?**
  - given complexity and access time

Predictions

- **Tape ➜ Log Structured File System on Tape**
  - Can allow multiple people to be streaming to the same tape
- **Distributed File System ➜ Scaled Object Store**
  - Lower cost, higher performance
- **Fast disk ➜ 100% Flash**
  - Database, POSIX
- **RAM ➜ Remotely accessed as Key/Value Store**
  - Implemented in hardware

Conclusion

- **Storage devices are continuing to get denser**
  - At a constant cost per device
- **Flash Devices are taking over fast disk**
  - Hybrids valuable?
- **Object Stores are replacing Distributed File Systems**
  - Success of S3
- **A Key/Value API for Storage**
  - Reduces or eliminates the metadata server
  - More natural for the log structure of storage devices

谢谢您

www.huawei.com