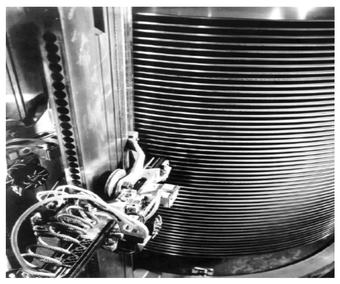
IBM Almaden Research Center - Storage Systems

Almaden Research - History of Innovations



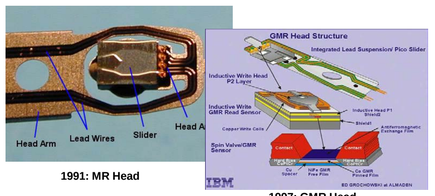
1955: Disk Drive

PubID	Publisher	PubAddress
03-4472822	Random House	123 4th Street, New York
04-7733903	Wiley and Sons	45 Lincoln Blvd, Chicago
03-4859223	O'Reilly Press	77 Boston Ave, Cambridge
03-3920886	City Lights Books	99 Market, San Francisco

AuthorID	AuthorName	AuthorBDay
345-28-2938	Haile Selassie	14-Aug-92
392-48-9965	Joe Blow	14-Mar-15
454-22-4012	Sally Hemmings	12-Sept-70
663-59-1254	Hannah Arendt	12-Mar-06

ISBN	AuthorID	PubID	Date	Title
1-34532-482-1	345-28-2938	03-4472822	1990	Cold Fusion for Dummies
1-38482-995-1	392-48-9965	04-7733903	1985	Macrame and Straw Tying
2-35921-499-4	454-22-4012	03-4859223	1952	Fluid Dynamics of Aquaducts
1-38278-293-4	663-59-1254	03-3920886	1967	Beads, Baskets & Revolution

1970: Relational Database
1974: SQL Query Language



1991: MR Head

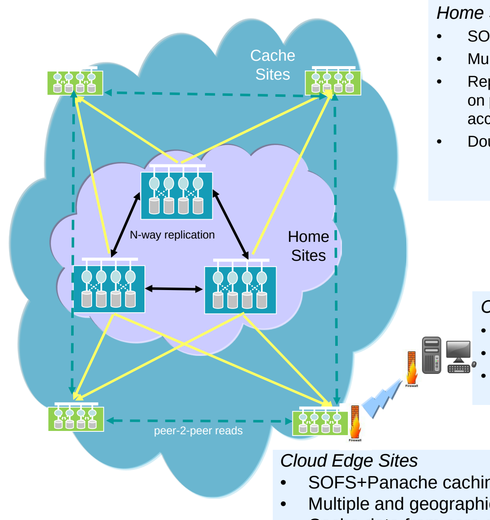
1997: GMR Head

1998: Microdrive

2000: National Medal of Technology
Leadership in Data storage technologies

Panache – A Parallel File System Cache for Global File Access

- Persistent cache in client-side GPFS file system
  - Global wide-area read-write cache designed for scalability and performance spanning multiple sites
  - Integrated with GPFS for consistent access from all nodes of the cluster
  - Over the WAN consistency is configurable
- pNFS for parallel data transfer over WAN
- Disconnected operations
  - Application updates to cache are written back asynchronously
  - Writebacks are deferred if disconnection occurs
    - Updates are journalled for later writeback
  - Supports whole-file or partial-file caching
- Storage Cloud
  - Backup
    - Use cache for data replication
  - Disaster Recovery
    - Recover from site failures
  - Peer-to-Peer
    - Seamless data movement among sites
  - Consolidation
    - Provide single file system view of numerous legacy filers
  - Migration
    - Online cross-vendor data migration



Home Sites
- SOFS+Panache-enabled clusters
- Multiple and geographically distributed
- Replicate/Migrate data between sites based on policy. (during planned upgrades or access based)
- Double as Edge site

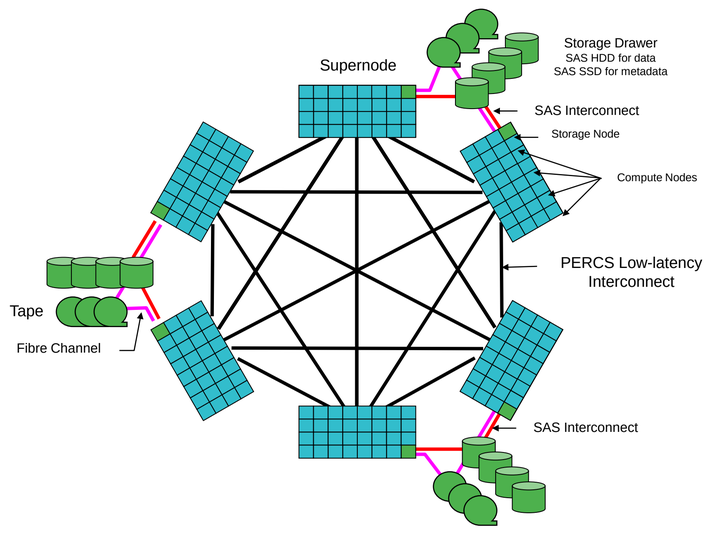
Customer Premise
- Small write-thru cache
- Connect through edge cache
- NFS, CIFS, etc

Cloud Edge Sites
- SOFS+Panache caching clusters
- Multiple and geographically distributed
- Cache data from core sites
- Cache both reads and writes
- Data async written to core
- Per fileset configurable consistency

File Systems for Petascale Supercomputers

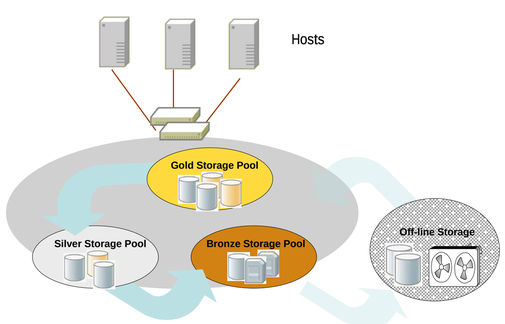
Blue Waters System at NCSA

- PERCS: "Productive Easy-to-use Reliable Computer System"
  - Balance between hardware, software, storage, networking, scaling, and productivity
  - Sustained Petaflop performance
  - 8x more memory per core than other HPCS systems
  - GPFS Perseus for storage controller
- NCSA Blue Waters PERCS
  - Collaboration between IBM, NCSA, State of Illinois, and partners
  - IBM Power7 processor
  - Shared memory and storage
  - 200k processor cores
  - 10 Petabyte GPFS storage system
  - Operational in 2011

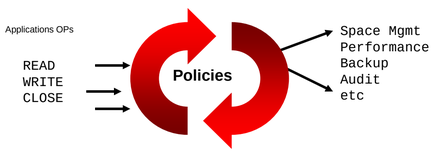


Petascale Data Management

- Petabytes on-line, Exabytes off-line
- Billions to Trillions of files
- Hundreds of Thousands of nodes



Policy Controlled Event Triggers



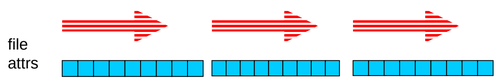
Policy controlled event trigger notification

open, close, create, destroy, read, write, chattr,…

Policy filters events based on file attributes

Events delivered to multiple data managers

Scale-out Policy Scan for Candidate Selection

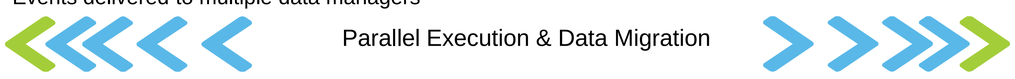


Sequential access to attributes & extended attributes

Parallel "MapReduce" on file metadata

Statistical sampling -- find "good" candidates quickly

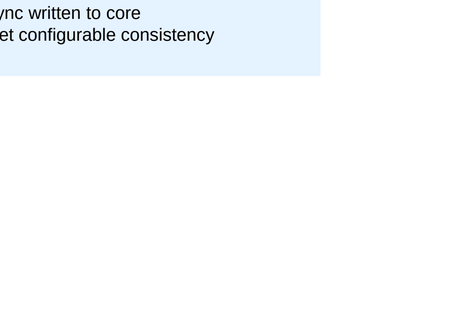
Low-priority -- consume idle cycles / bandwidth



Perseus: Advanced software RAID for GPFS

50-disk arrays to 100,000-disk supercomputers

- Software RAID for scalable GPFS (NSD)
- Declustered RAID implementation
  - Spread data strips randomly across all array disks
  - Performance will be minimally affected by rebuilding array
- 2/3-fault tolerant erasure codes
  - "RAID-D2" or "RAID-D3"
  - Software Reed-Solomon
  - Optional 3/4-way mirroring
- End-to-end checksum
- Runs on generic servers with direct-attach disks
- Supersedes traditional external RAID controller
  - Reduces storage subsystem costs by 30 - 60 %



- Improved file system performance
  - With 100k disks, a storage array is always rebuilding
    - 100k disks * 24 / 400 khrs => 6 rebuilds per day
- Improved data integrity
  - RAID-5 is non-starter with 100k disks: MTTDL ~ 9 days!
    - Hard error rate of 1-in-10^{15} bits implies data loss every ~26th rebuild, or once every 26 / 6 = 4 days
  - RAID-D2 (8+2P stripes): MTTDL ~ 100 years
  - RAID-D3 (8+3P stripes): MTTDL ~ 130 million years!