Data Reliability (and ideas on how to improve it)

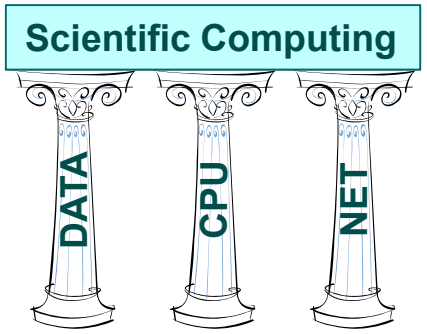
Alberto Pace

alberto.pace@cern.ch

CERN Data and Storage Services Group

Why data reliability ?

- **Scientific research in recent years has exploded the computing requirements**
  - Computing has been the strategy to reduce the cost of traditional research
  - Computing has opened new horizons of research not only in High Energy Physics
- **Data management is one of the three pillars of scientific computing**

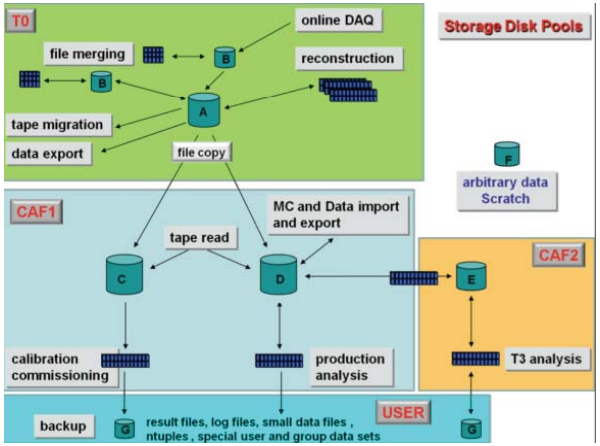


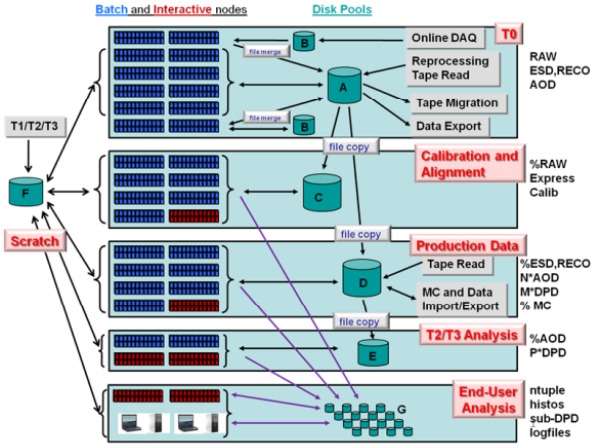
Data management

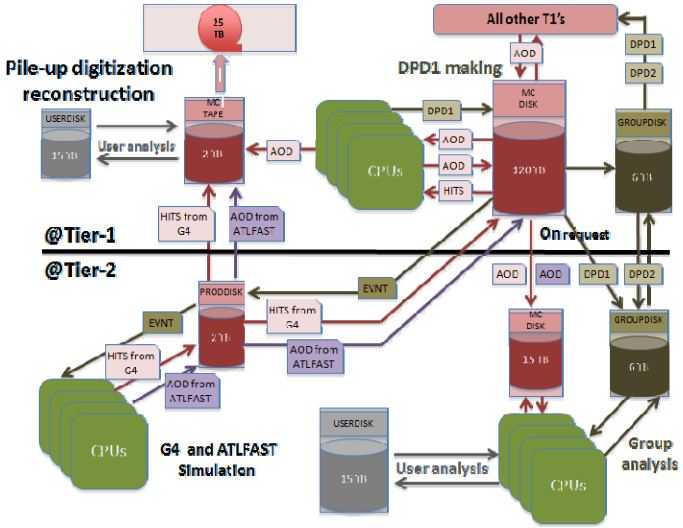
- **Data Management solves the following problems**
  - Data reliability
  - Data archives, history, long term preservation
  - Data distribution
  - Access control
  - In general:
    - Empower the implementation of a workflow for data processing

What is data management ?

- **Examples from LHC experiment data models**



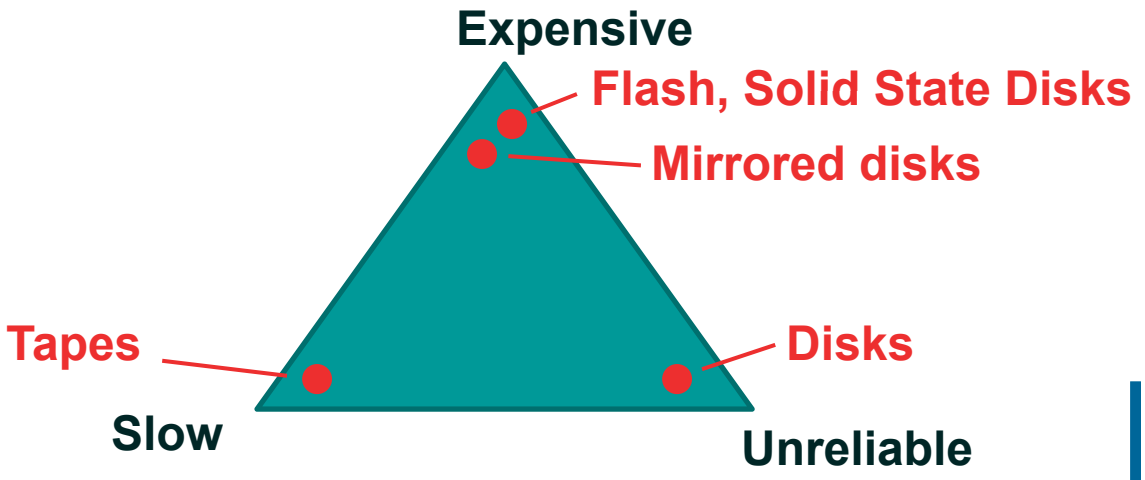




- **Two building blocks to empower data processing**
  - Data pools with different quality of services
  - Tools for data transfer between pools

Data pools

- **Different quality of services**
  - Three parameters: (Performance, Reliability, Cost)
  - You can have two but not three



Data Reliability

- **Reliability is related to the probability to lose data**
  - Def: “the probability that a storage device will perform an arbitrarily large number of I/O operations without data loss during a specified period of time”
- **Reliability of the “service” depends on the environment (energy, cooling, people, ...)**
  - Will not discuss this further
- **Reliability of the “service” starts from the reliability of the underlying hardware**
  - Example of disk servers with unmirrored disks: reliability of service = reliability of disks
- **But data management solutions can increase the reliability of the hardware at the expenses of performance and/or additional hardware / software**
  - Disk Mirroring
  - Redundant Array of Inexpensive Disks (RAID)

Reminder: types of RAID

- **RAID0**
  - Disk striping
- **RAID1**
  - Disk mirroring
- **RAID5**
  - Parity information is distributed across all disks
- **RAID6**
  - Uses Reed–Solomon error correction, allowing the loss of 2 disks in the array without data loss

http://en.wikipedia.org/wiki/RAID

Reed–Solomon error correction ?

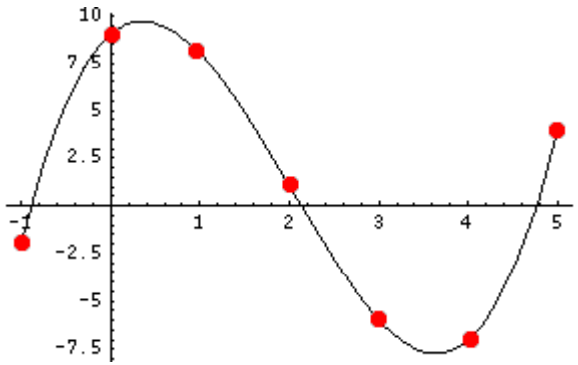
- .. is an error-correcting code that works by oversampling a polynomial constructed from the data
- Any k distinct points uniquely determine a polynomial of degree, at most, k − 1
- The sender determines the polynomial (of degree k − 1), that represents the k data points. The polynomial is "encoded" by its evaluation at n (≥ k) points. If during transmission, the number of corrupted values is < n-k the receiver can recover the original polynomial.
- Note: only when n-k ≤ 3 , we have efficient implementations
  - n-k = 0 no redundancy
  - n-k = 1 is Raid 5 (parity)
  - n-k = 2 is Raid 6 (Reed Solomon or double parity)
  - n-k = 3 is … (Triple parity)

http://en.wikipedia.org/wiki/Reed-Solomon

Reed–Solomon (simplified) Example

- 4 Numbers to encode: { 1, -6, 4, 9 } (k=4)
- polynomial of degree 3 (k − 1):

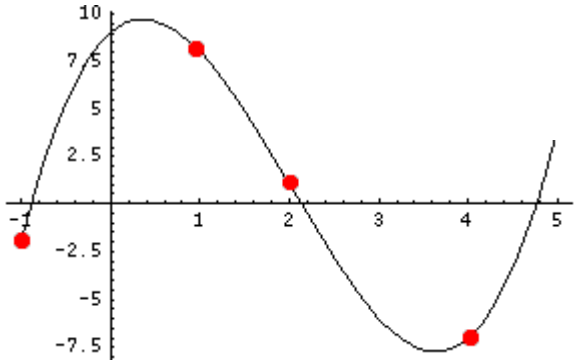
$$y = x^3 - 6x^2 + 4x + 9$$



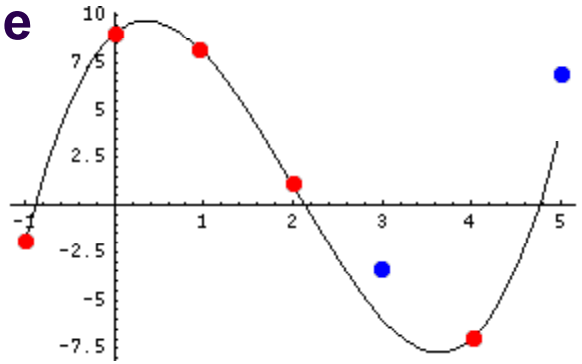
- We encode the polynomial with n=7 points
  { -2, 9, 8, 1, -6, -7, 4 }

Reed–Solomon (simplified) Example

- To reconstruct the polynomial, any 4 points are enough: we can lose any 3 points.



- We can have an error on any 2 points that can be corrected: We need to identify the 5 points "aligned" on the only one polynomial of degree 3 possible

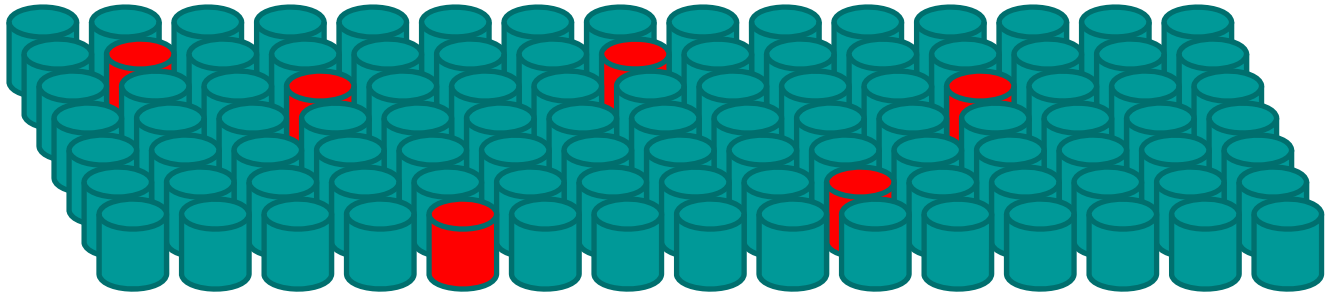


http://kernel.org/pub/linux/kernel/people/hpa/raid6.pdf

Reliability calculations

- **With RAID, the final reliability depends on several parameters**
  - The reliability of the hardware
  - The type of RAID
  - The number of disks in the set
- **Already this gives lot of flexibility in implementing arbitrary reliability**

Raid 5 reliability



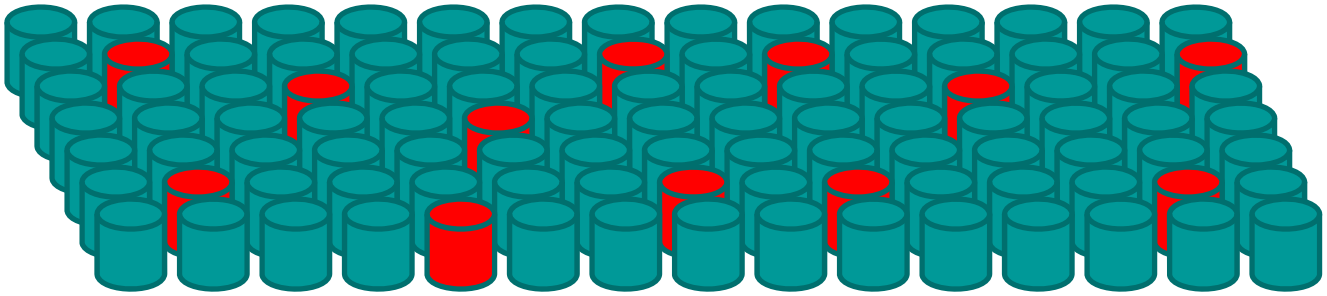
- Disk are regrouped in sets of equal size. If c is the capacity of the disk and n is the number of disks, the sets will have a capacity of

$$c\ (n-1)$$

example: 6 disks of 1TB can be aggregated to a “reliable” set of 5TB

- The set is immune to the loss of 1 disk in the set. The loss of 2 disks implies the loss of the entire set content.

Raid 6 reliability



- Disk are regrouped in sets of arbitrary size. If c is the capacity of the disk and n is the number of disks, the sets will have a capacity of

$$c\ (n-2)$$

example: 12 disks of 1TB can be aggregated to a “reliable” set of 10TB

- The set is immune to the loss of 2 disks in the set. The loss of 3 disks implies the loss of the entire set content.

Some calculations for Raid 5

- Disks MTBF is between 3 x 10^5 and 1.2 x 10^6 hours
- Replacement time of a failed disk is < 4 hours
- Probability of 1 disk to fail within the next 4 hours

$$P_f = \frac{Hours}{MTBF} = \frac{4}{3\times10^5} = 1.3\times10^{-5}$$

Some calculations for Raid 5

- Disks MTBF is between 3 x 10^5 and 1.2 x 10^6 hours
- Replacement time of a failed disk is < 4 hours
- Probability of 1 disk to fail within the next 4 hours

$$P_f = \frac{Hours}{MTBF} = \frac{4}{3\times10^5} = 1.3\times10^{-5}$$

- Probability to have a failing disk in the next 4 hours in a 15 PB computer centre (15'000 disks)

$$P_{f15000} = 1-(1-P_f)^{15000} = 0.18$$

Some calculations for Raid 5

- Disks MTBF is between 3 x 10^5 and 1.2 x 10^6 hours
- Replacement time of a failed disk is < 4 hours
- Probability of 1 disk to fail within the next 4 hours

$$P_f = \frac{Hours}{MTBF} = \frac{4}{3\times10^5} = 1.3\times10^{-5}$$

p(A and B) = p(A) * p(B/A)

if A,B independent : p(A) * p(B)

- Probability to have a failing disk in the next 4 hours in a 15 PB computer centre (15'000 disks)

$$P_{f15000} = 1-(1-P_f)^{15000} = 0.18$$

- Imagine a Raid set of 10 disks. Probability to have one of the remaining disk failing within 4 hours

$$P_{f9} = 1-(1-P_f)^{9} = 1.2\times10^{-4}$$

Some calculations for Raid 5

- Disks MTBF is between 3 x 10^5 and 1.2 x 10^6 hours
- Replacement time of a failed disk is < 4 hours
- Probability of 1 disk to fail within the next 4 hours

$$P_f = \frac{Hours}{MTBF} = \frac{4}{3\times10^5} = 1.3\times10^{-5}$$

p(A and B) = p(A) * p(B/A)

if A,B independent : p(A) * p(B)

- Probability to have a failing disk in the next 4 hours in a 15 PB computer centre (15'000 disks)

$$P_{f15000} = 1-(1-P_f)^{15000} = 0.18$$

- Imagine a Raid set of 10 disks. Probability to have one of the remaining disk failing within 4 hours

$$P_{f9} = 1-(1-P_f)^{9} = 1.2\times10^{-4}$$

- However the second failure may not be independent from the first one. Let's increase its probability by two orders of magnitude as the failure could be due to common factors (over temperature, high noise, EMP, high voltage, faulty common controller, ...)

$$P_{f9corrected} = 1-(1-P_f)^{900} = 0.0119$$

Some calculations for Raid 5

- Disks MTBF is between 3 x 10^5 and 1.2 x 10^6 hours
- Replacement time of a failed disk is < 4 hours
- Probability of 1 disk to fail within the next 4 hours

$$P_f = \frac{Hours}{MTBF} = \frac{4}{3\times10^5} = 1.3\times10^{-5}$$

p(A and B) = p(A) * p(B/A)

if A,B independent : p(A) * p(B)

- Probability to have a failing disk in the next 4 hours in a 15 PB computer centre (15'000 disks)

$$P_{f15000} = 1-(1-P_f)^{15000} = 0.18$$

- Imagine a Raid set of 10 disks. Probability to have one of the remaining disk failing within 4 hours

$$P_{f9} = 1-(1-P_f)^{9} = 1.2\times10^{-4}$$

- However the second failure may not be independent from the first one. Let's increase its probability by two orders of magnitude as the failure could be due to common factors (over temperature, high noise, EMP, high voltage, faulty common controller, ...)

$$P_{f9corrected} = 1-(1-P_f)^{900} = 0.0119$$

- Probability to lose computer centre data in the next 4 hours

$$P_{loss} = P_{f15000} \times P_{f9corrected} = 6.16\times10^{-4}$$

- Probability to lose data in the next 10 years

$$P_{loss10\,yrs} = 1-(1-P_{loss})^{10\times365\times6} = 1\text{-}10^{-21} \cong 1$$

Same calculations for Raid 6

- **Probability of 1 disk to fail within the next 4 hours**

$$P_f = \frac{Hours}{MTBF} = \frac{4}{3\times10^5} = 1.3\times10^{-5}$$

- **Imagine a raid set of 10 disks. Probability to have one of the remaining 9 disks failing within 4 hours (increased by two orders of magnitudes)**

$$P_{f9} = 1-(1-P_f)^{900} = 1.19\times10^{-2}$$

- **Probability to have another of the remaining 8 disks failing within 4 hours (also increased by two orders of magnitudes)**

$$P_{f8} = 1-(1-P_f)^{800} = 1.06\times10^{-2}$$

- **Probability to lose data in the next 4 hours**

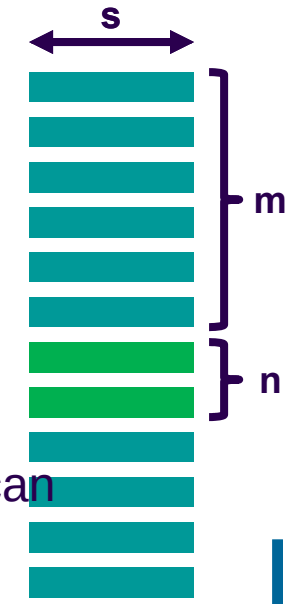
$$P_{loss} = P_{f15000}\times P_{f99}\times P_{f98} = 2.29\times10^{-5}$$

- **Probability to lose data in the next 10 years**

$$P_{loss10\,yrs} = 1-(1-P_{loss})^{10\times365\times6} = 0.394$$

Arbitrary reliability

- **RAID is “disks” based. This lacks of granularity**
- **For increased flexibility, an alternative would be to use files ... but files do not have constant size**
- **File “chunks” is the solution**
  - Split files in chunks of size “s”
  - Group them in sets of “m” chunks
  - For each group of “m” chunks, generate “n” additional chunks so that
    - For any set of “m” chunks chosen among the “m+n” you can reconstruct the missing “n” chunks
  - Scatter the “m+n” chunks on independent storage



Arbitrary reliability with the "chunk" based solution

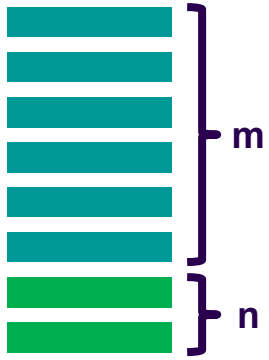
- **The reliability is independent form the size "s" which is arbitrary.**
  - Note: both large and small "s" impact performance
- **Whatever the reliability of the hardware is, the system is immune to the loss of "n" simultaneous failures from pools of "m+n" storage chunks**
  - Both "m" and "n" are arbitrary. Therefore arbitrary reliability can be achieved
- **The fraction of raw storage space loss is n / (n + m)**
- **Note that space loss can also be reduced arbitrarily by increasing m**
  - At the cost of increasing the amount of data loss if this would ever happen

Analogy with the gambling world

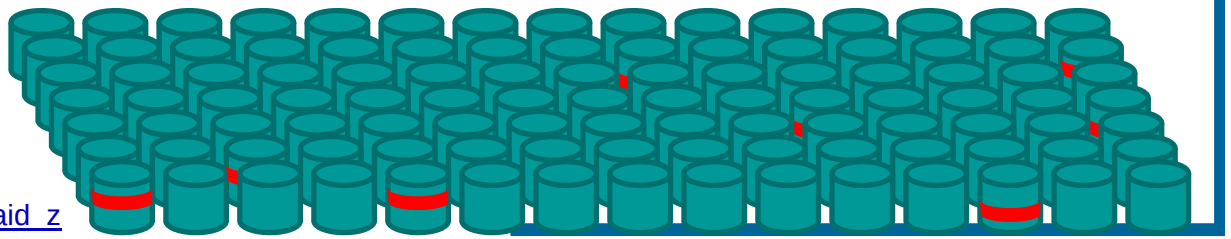
- We just demonstrated that you can achieve “arbitrary reliability” at the cost of an “arbitrary low” amount of disk space. By just increasing the amount of data you accept loosing when this happens.
- In the gambling world there are several playing schemes that allows you to win an arbitrary amount of money with an arbitrary probability.
- Example: you can easily win 100 Euros at > 99 % probability ...
  - By playing up to 7 times on the “Red” of a French Roulette and doubling the bet until you win.
  - The probability of not having a “Red” for 7 times is $(19/37)^7 = 0.0094$)
  - You just need to take the risk of loosing 12’700 euros with a 0.94 % probability

Amount		Win		Lost	
Bet	Cumulated	Probability	Amount	Probability	Amount
100	100	48.65%	100	51.35%	100
200	300	73.63%	100	26.37%	300
400	700	86.46%	100	13.54%	700
800	1500	93.05%	100	6.95%	1500
1600	3100	96.43%	100	3.57%	3100
3200	6300	98.17%	100	1.83%	6300
6400	12700	99.06%	100	0.94%	12700

Practical comments



- **n can be 1 or 2**
  - 1 = Parity
  - 2 = Parity + Reed-Solomon, double parity
  - 3 = Reed Solomon, ZFS triple parity
- **Although possible, n > 2 has a computational impact (for Reed Solomon) that affects performances**
- **m chunks of any (m + n) sets are enough to obtain the information. Must be saved on independent media**
  - Performance can depend on m (and thus on s, the size of the chunks): The larger m is, the more the reading can be parallelized
  - Until the client bandwidth is reached



http://blogs.sun.com/ahl/entry/triple_parity_raid_z

How to tune m and n

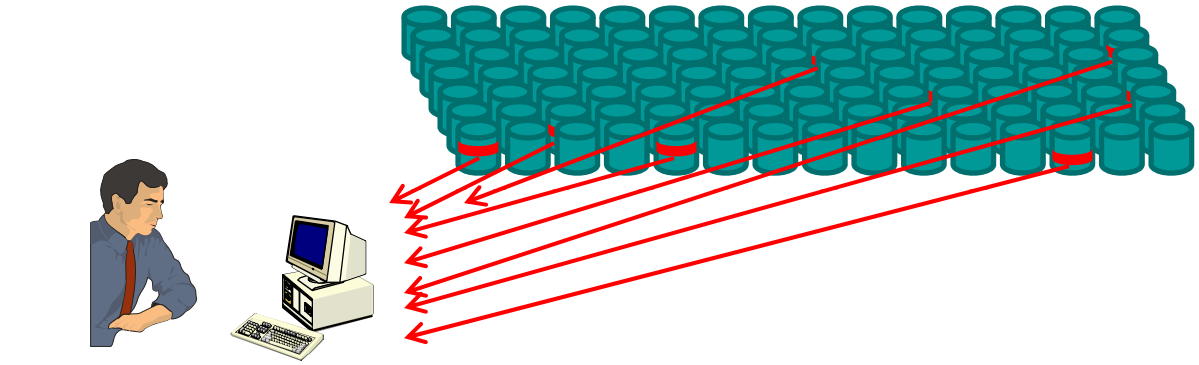
1. **Start from the published MTBF of your hardware vendor**
2. **Decide what is the "arbitrary" replacement time of a faulty hardware**
3. **Chose m and n in order to expect the reliability that you want**
4. **While running the system you MUST constantly monitor and record**
   - Hardware failures (so that you constantly re-estimate if the true MTBF)
   - Replacement time of faulty hardware
- **With these new parameters you recalculate new m and n**

Chunk transfers

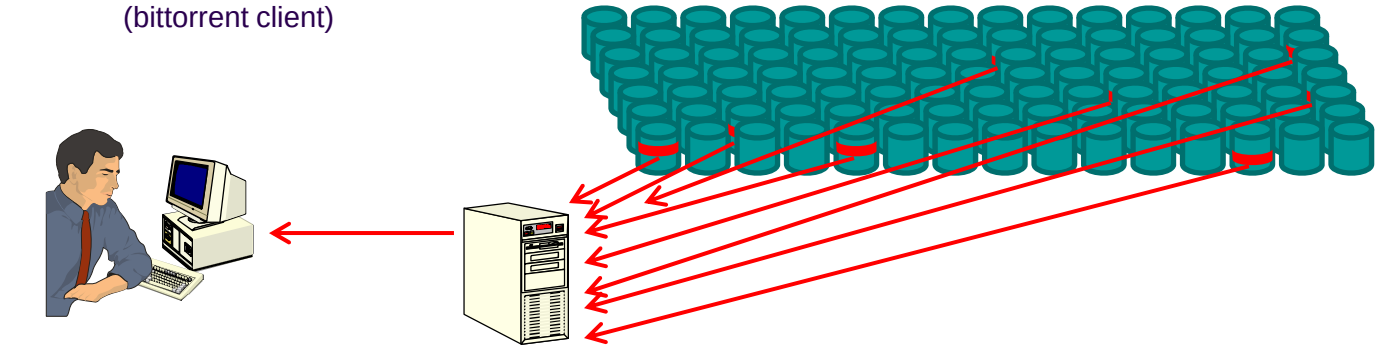
- **Among many protocols, Bittorrent is the most popular**
- **An SHA1 hash (160 bit digest) is created for each chunk**
- **All digests are assembled in a “torrent file” with all relevant metadata information**
- **Torrent files are published and registered with a tracker which maintains lists of the clients currently sharing the torrent’s chunks**
- **In particular, torrent files have:**
  - an "announce" section, which specifies the URL of the tracker
  - an "info" section, containing (suggested) names for the files, their lengths, the list of SHA-1 digests
- **Reminder: it is the client’s duty to reassemble the initial file and therefore it is the client that always verifies the integrity of the data received**

http://en.wikipedia.org/wiki/BitTorrent_(protocol)

Reassembling the chunks



Data reassembled directly on the client (bittorrent client)

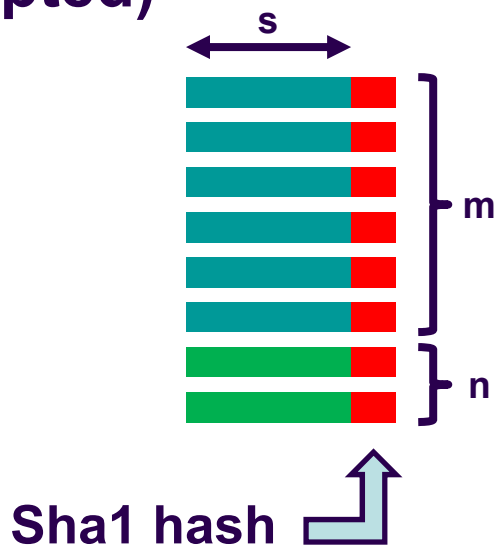


Reassembly done by the data management infrastructure

Middleware

Identify broken chunks

- The SHA1 hash (160 bit digest) guarantees chunks integrity.
- It tells you the corrupted chunks and allows you to correct n errors (instead of n-1 if you would not know which chunks are corrupted)



Summary

- **Several components, many of them independent**
  - Name Servers and databases
  - (I/O Scheduling)
  - (Data Access protocols and SRM)
  - Reliability
  - Data Replication
  - Data Caching
  - Access Control and Security
    - Authentication, Authorization, Accounting
  - Monitoring, Alarms
  - Quota
- **Allow to build an architecture to transform "data technologies" into "data management services"**