Summary of the Storage Technology Session

(Convener Andrei Maslennikov)

HEPIX Spring Workshop 2008

After C5-Meeting 23.5.2008

Andreas-Joachim Peters
CERN-IT DM-SMD

A normal life...

A physicists life...

Storage Session Summary – After C5 Meeting - 23.5.2008

Meeting Agenda

Wednesday, 07 May 2008

[28] **Final Report from File Systems Working Group**
by Andrei MASLENNIKOV (CASPUR)
(Council Chamber: 09:45 - 10:15)

[29] **Options for medium-/long-term improvements to LHC mass storage and data management**
by Dirk DUELLMANN (CERN-IT)
(Council Chamber: 10:15 - 10:45)

Coffee break
(10:45 - 11:15)

[30] **Storage elements at BNL**
by Robert PETKUS (Brookhaven National Laboratory)
(Council Chamber: 11:15 - 11:45)

[34] **CASTOR Status and Plans**
by Sebastien PONCE (CERN)
(Council Chamber: 11:45 - 12:15)

[32] **Towards the new data management solution at CNAF**
by Vladimir SAPUNENKO (CNAF)
(Council Chamber: 12:15 - 12:45)

[43] **Handling large datasets at Google: Current systems and future directions**
by Sascha BRAWER (Google)
(Council Chamber: 14:00 - 14:30)

[42] **FZK storage news**
by Silke HALSTENBERG (FZK)
(Council Chamber: 14:30 - 15:00)

[8] **The unbearable slowness of tapes**
by Charles CURRAN (CERN)
(Council Chamber: 15:00 - 15:30)

Coffee break
(15:30 - 16:00)

[17] **Setting up a simple Lustre Filesystem**
by Stephan WIESAND (DESY)
(Council Chamber: 16:00 - 16:30)

[19] **Experience and Lessons learnt from running high availability databases on Network Attached Storage**
by Nilo SEGURA CHINCHILLA (CERN)
(Council Chamber: 16:30 - 17:00)

[23] **Lustre cluster in production at GSI**
by Walter SCHöN (GSI)
(Council Chamber: 17:00 - 17:30)

HEPiX

CERN
4 talks

BNL

INFN CNAF

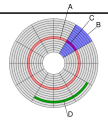
Google Schweiz

KIT Karlsruhe Institute of Technology

DESY

GSI

Topic Distribution

- Filesystems  6 talks
- Tape  3 talks
- HA  3 talks
- Hw R&D  3 talks
- Sw R&D 1 talk
- DB  1 talk

*many talks had overlapping topics

Storage Solutions in CC presentations

- **CASTOR**
  - CERN (CNAF)
- **dCache**
  - FZK
- **xrootd**
  - BNL

- **BlueArc**
  - BNL
- **LUSTRE**
  - GSI/DESY
- **GPFS**
  - FZK/CNAF

Trends observed ...

- Filesystems become more important in computing center storage systems
  - **LUSTRE** was the 'big winner' (atleast for 3 speakers)
    - 1st rank in HEPIX FSWG tests
    - HEP installation at GSI (0.3 PB – 60 server - 6 GB/s)
    - DESY presented simple setup recipe
  - **GPFS**
    - **GPFS + TSM** backend as new storage solution at CNAF
    - 2nd rank in HEPIX FSWG tests

Trends observed ...

- Storage Hardware
  - **low-end**
    - GSI (no SAN-disks)
  - **mid-range**
    - FZK: NEC D3-10 (attached arrays: FC)
  - **high-end**
    - BNL
      - SUN X4500/ZFS
    - CERN
      - SUN (NAS/DB)
  - **new storage medias/network**
    - BNL tested SSD disks /FZK tested 10GE for disk servers

HEPIX FSWG Final Report

Approach to evaluate **existing** storage solutions

- The goal was to review the available file system solutions and storage access methods, and to divulge the know-how and practical recommendations among HEP organizations and beyond

CASPUR	A.Maslennikov (Chair), M.Calori (Web Master)
CEA	J-C.Lafoucriere
CERN	B.Panzer-Steindel
DESY	M.Gasthuber, Y.Kemp, P.van der Reest,
FZK	J.van Wezel, C.Jung
IN2P3	L.Tortay
INFN	G.Donvito, V. Sapunenko
LAL	M.Jouvin
NERSC/LBL	C.Whitney
RAL	N.White
RZG	H.Reuter
SLAC	A.Hanushevsky, A.May, R.Melen
U.Edinburgh	G.A.Cowan

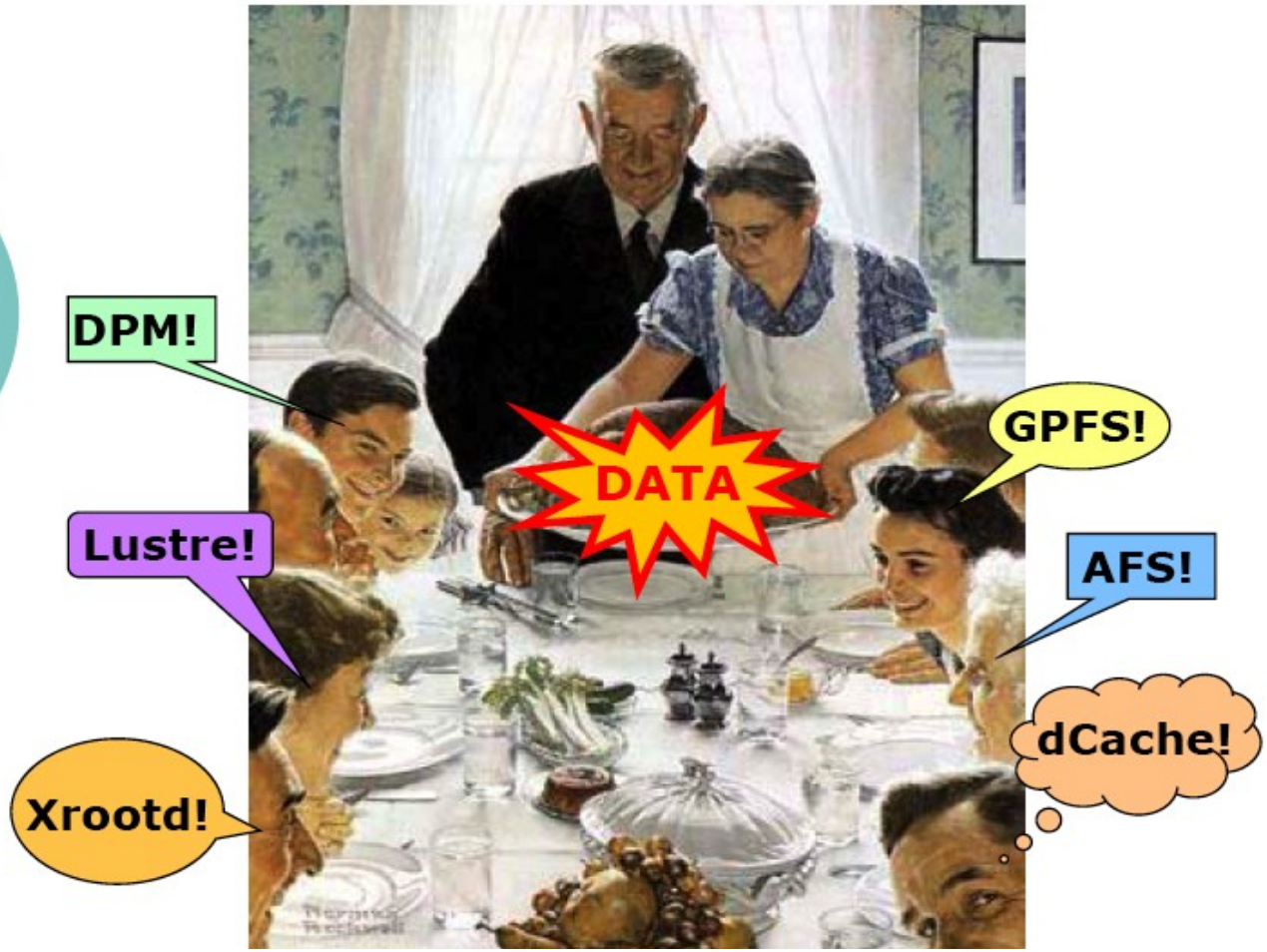
Comparison of:

AFS,GPFS, Lustre,dCache, DPM,xrootd

- Selected a reduced set of architectures to look at:
  - File Systems with Posix Transparent File Access (AFS, GPS, Lustre);
  - Special Solutions (dCache, DPM and Xrootd)

CASTOR2? Not included!

HEPIX FSWG
The test ... or who get's most of the cake



- **Same hardware** used for all
  - 10 standard CERN disk server
  - 60x8 core CPU server
- **Same tests** used for all
  - *Assume:* results are correct for the performed tests

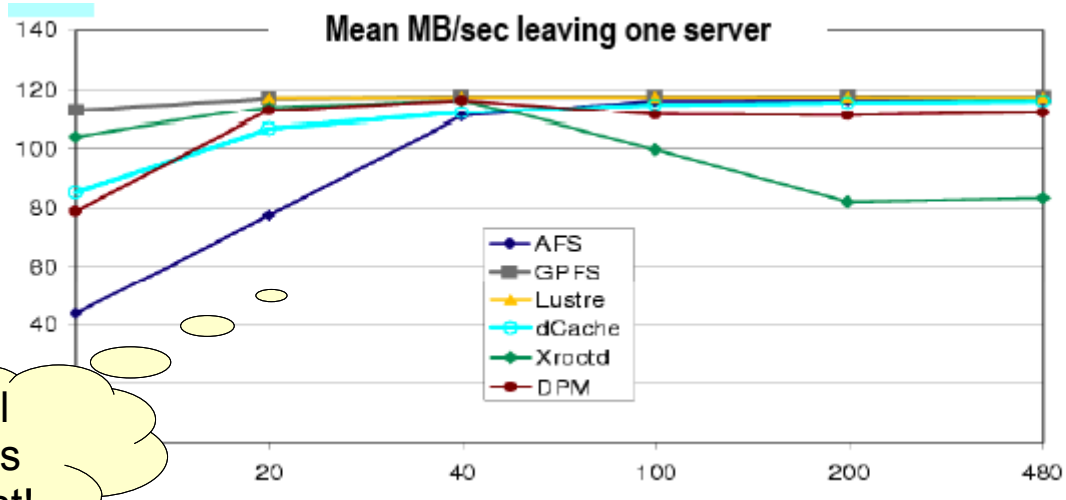
HEPIX FSWG

1. **"Acceptance Test":** 50 thousand files of 300 MB each were written on 10 servers

	Lustre	dCache	DPM	Xrootd	AFS	GPFS
Average MB/sec entering a disk sever	117	117	117	114	109	96

'write' works well for most!

2. **"Sequential Read Test":** 10,20,40,100,200,480 simultaneous tasks were reading a series of 300-MB files sequentially, with a block size of 1 MB.



	Number of jobs					
	10	20	40	100	200	480
AFS	3812	6751	9622	10069	10008	9894
GPFS	9794	10102	10144	10130	10073	9921
Lustre	9774	10138	10151	10117	10089	9935
dCache	5254	7959	9323	9744	9770	9531
Xrootd	8955	9801	10009	8545	7028	6953
DPM	4644	7872	9693	9390	9652	9866

Sequential 'read' works well for most!

HEPIX FSWG

3. "Pseudo-Random Read Test":

It is quite complicated to describe

100,200,480 simultaneous tasks were reading a series of 300-MB files. Each of the tasks was programmed to read randomly selected small data chunks from within the file; the size of a chunk to read was set to be 10,25,50 or 100 KB and remained constant while 300 megabytes were read. Then the next file was read out, with a different chunk size. Each of the files was read only once.

The chunk sizes were selected in a pseudo-random way:
10 KB (10%), 25 KB (20%), 50 KB (50%), 100 KB (20%).

	Number of jobs		
	100	200	480
AFS	6766	3802	1815
GPFS	13728	9575	6502
Lustre	12109	12062	11908
dCache	3185	4356	5530
Xrootd	3036	4194	5223
DPM	3216	4513	5988

Numbers of 300-MB files processed

	Number of jobs		
	100	200	480
AFS	79	112	87
GPFS	114	75	69
Lustre	117	117	117
dCache	35	49	65
Xrootd	34	47	60
DPM	35	48	64

Average MB leaving a server per second

Test favours caching systems?

1st Conclusion of the WG: We should run real experimental analysis code using real data, but WG lacked time/resources.

HEPIX FSWG Conclusions

Investigation of

6 storage solutions

POSIX solutions **easily compete** special **HEP sol.** (in some use cases)

- The HEPiX File Systems Working Group was set up to investigate the storage ss solutions and to provide practical recommendations to HEP
- p made an assessment of existing storage architectu cted information on them, and performed a simple com sis for 6 of the most diffused solutions. It leaves behind a start-up web e dedicated to the storage technologies.
- The s ies done by the group confirm that shared, scalable file system with Posix file access semantics may easily compete in performance wi the special storage access solutions currently in use at HEP sites, at lea in some of the use cases.
- r short list of recommended TFA file systems contains GPFS and Lustr latter appears to be more flexible, may be slightly more performi is free. The group hence recommends to consider deployment of the re file system in venue of a shared data store for large compute clusters.
- Initial compa tive studies performed on a common hardware base had revealed the need to further investigate the role of storage architecture as a part of a complex compute cluster, against the real LHC analysis codes.

We rank and recommend:
1. LUSTRE
2. GPFS

We understood: **we need to run real life applications not tests!**

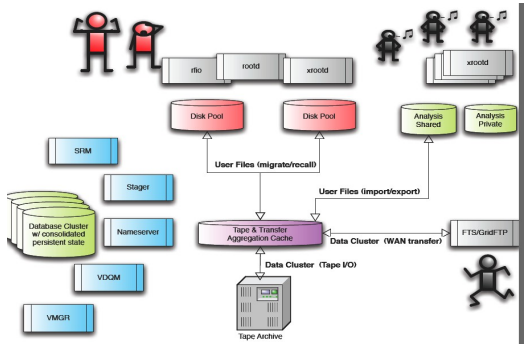
Improvement Options for LHC Mass Storage and Data Management

Dirk Düllmann

HEPIX spring meeting @ CERN,

7 May 2008

- Report on role, mandate and status of IT-DM R&D project
  - mid-/longterm developement of CERN datamanagement
  - 1st roadmap in summer
  - *Currently only tests & discussions*



Storage Elements at BNL

- > **2 PB** added in **2008**, >**4 PB** in **2009**
  - Storage demand growths faster than CPU for ATLAS
- BNL favours **SUN Solaris** + SunFire 4500
- Interesting R&D/testing with **SSD disks** and HEP application
  - Performance Comparison between SSD & Disk with PROOF/XROOTD analysis

Purchased (10) Mtron 3.5" SATAII SSDs, (1) per testbed system
- 64GB, 120MB/sec read, 90MB/sec write sustained performance
- Random access time = 0.1 ms (SATA HDD = ~10ms)
- Write endurance >140 years @ 50GB/day
- MTBF = 1 million hours
- 7-bit error correction code

  - **Gain Factor 6** in processing time to draw a single histogram
  - Currently: random write performance of devices very poor, but coming:
    - Fusion-IO: ioDrive promises 600 MB/s random write + 700 MB/s random reads *(PCI-X card – up to 640 GB .. < 30 $/GB .. expensive!)*

Castor status and plans

Sebastien Ponce, Hepix, May 7th 2008

- ... I imagine people at CERN know well ... but
  - at present mostly consolidation
  - somehow 'waiting' for first decisions of R&D DM project for future directions

Toward new HSM solution using GPFS/TSM/StoRM integration

Vladimir Sapunenko (INFN, CNAF)
Luca dell'Agnello (INFN, CNAF)
Daniele Gregori (INFN, CNAF)
Riccardo Zappi (INFN, CNAF)
Lunca Magnoni (INFN, CNAF)
Elisabetta Ronchieri (INFN, CNAF)
Vincenzo Vagnoni (INFN, Bologna)

GPFS/HSM@CNAF

- **D1T1 prototype** tested for 2 month
  - 'Positive' results but needs more/larger testing
  - 1st Production usage by LHCb in CCRC 08
- **D0T1 prototype**
  - 'Encouraging' results
- **D1T0**
  - Since Febrary in production for Atlas
- Tape integration via **ILM* policies** (GPFS)

*ILM = Information Lifecycle Management

The Unbearable Slowness of Tape

C. Curran, CERN

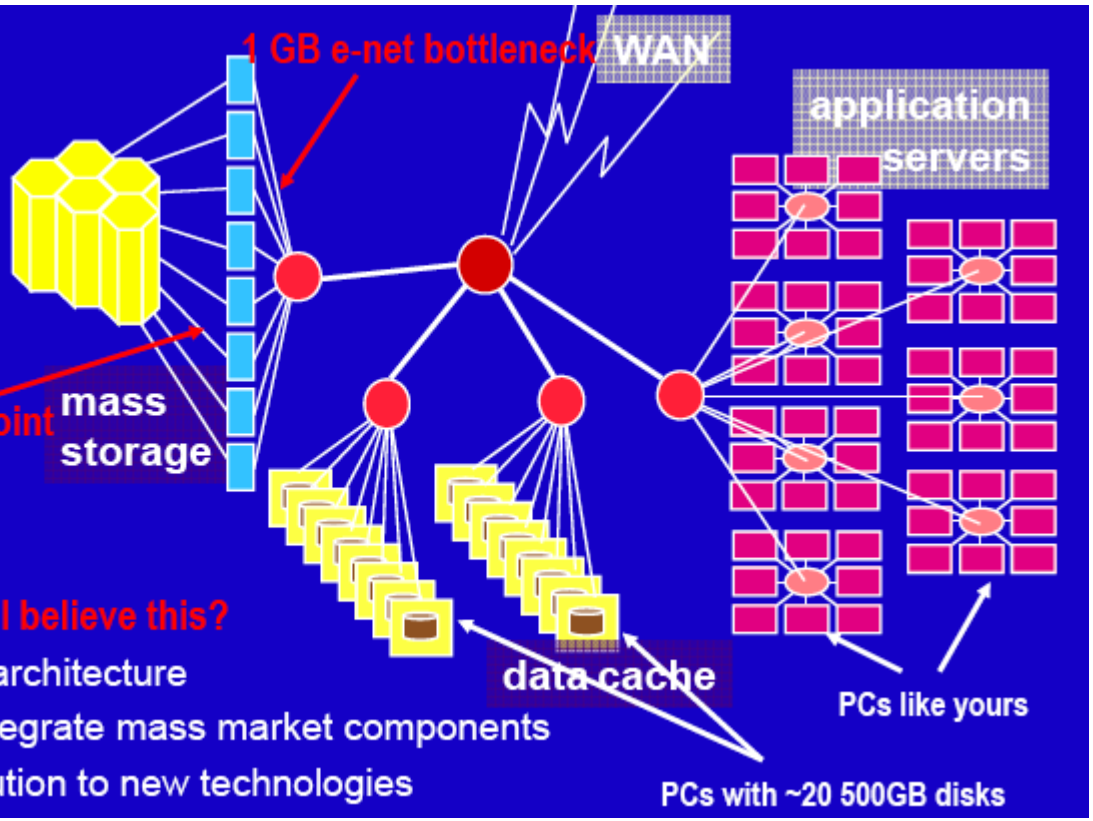
HEPIX

CERN, May 2008 (version 6.5.2008 10h00)

„We have 120 HQ tapes, should see 10-12 Gb/s .... but we don't .....“

It's (y)our fault

Repack = 'dd'



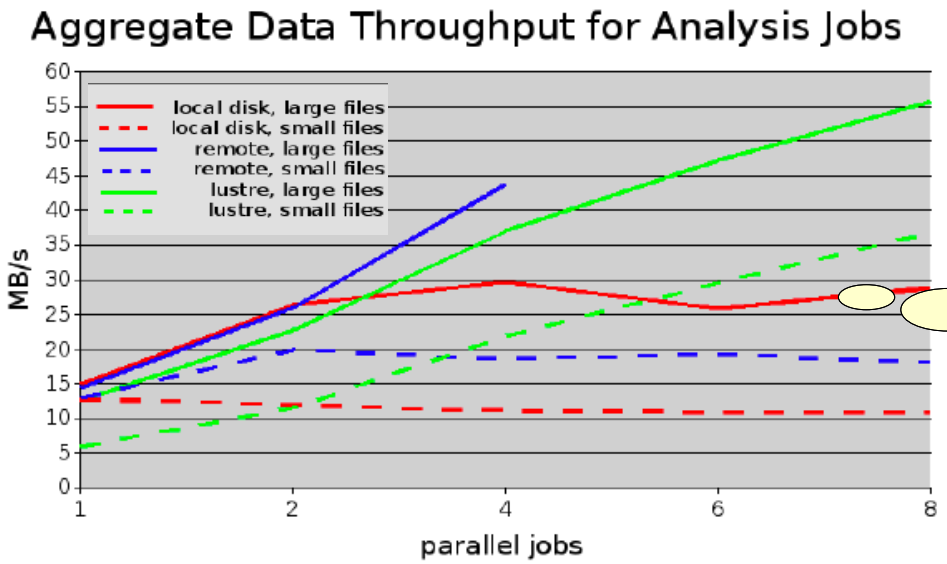
But we can do better!

- can go from 58% to 88% tape writing efficiency!

- can go from <20% to 39% tape read efficiency!

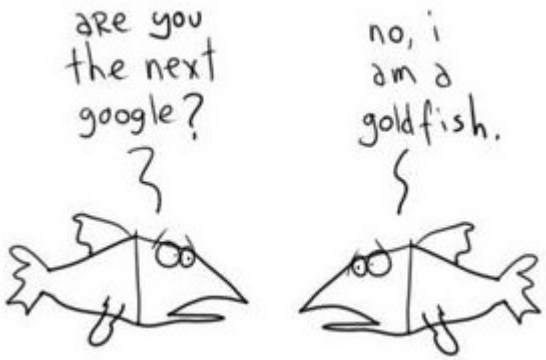
Lustre Cluster at GSI

Walter Schön, GSI



- **GSI LUSTRE Cluster**
  - 60 disk server (120 SATA arrays)
  - 0.3 PB Raid5
  - 6 Gb/s aggregate i/o
  - Current system 660 Euro/TB
    - future 400 Euro/TB
  - No disk server redundancy/replicas
  - HA for head node in production
  - Judgement
    - not everything is yet paradise !

Tests with ALICE analysis small & big files:
LUSTRE scales well with number of cores



Invited Talk

The biggest & proven successful storage system presented in the workshop!

Handling Large Datasets at Google: Current Systems and Future Directions

Sascha Brawer
sascha@google.com

(Original talk by Jeff Dean)

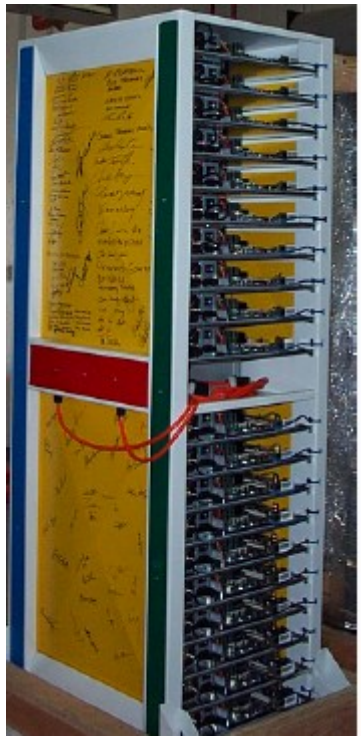
Google

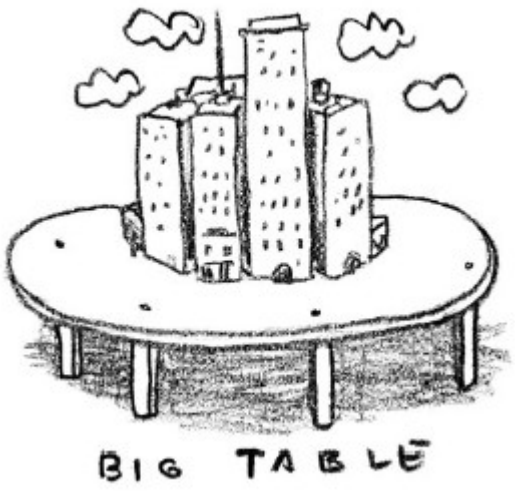
Google Schweiz

- ***Distributed System***
  - PB datasets (offline processing)
  - TB datasets (online applications)
- ***Key Components***
  - Scheduling System (batch queue)
  - GFS – Google Filesystem (200+ GFS cluster)
    - replicated file chunks
    - biggest cluster 5000+ machines, 5+ PB, 10.000+ clients
  - Big Table (DB)
  - Map Reduce (job framework)

Google Schweiz

- Hardware Philosophy
  - Low-cost machines
  - Everything uses trivial parallelism
  - Performance/$ matters – not Performance/machine
  - Many centers around the globe
  - Very frequent failures managed by software
  - Inhouse rack design





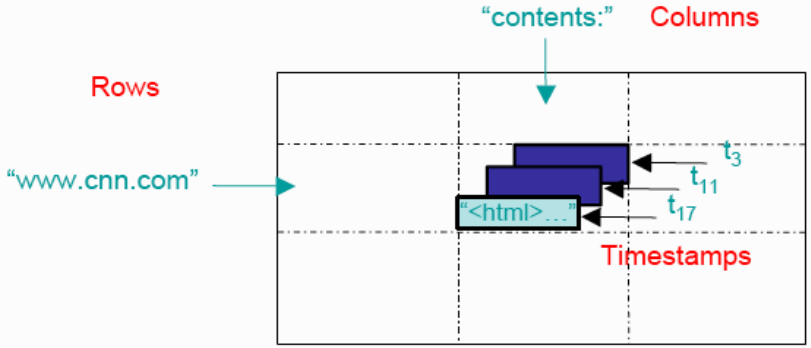
Google Schweiz

- 500 Big Table cells
- Largest 6+Pb of data
  - 3000+ machines

- Busiest cell 500000+ ops/s sustained 24/7

BigTable Data Model

- Multi-dimensional sparse sorted map
  (row, column, timestamp) => value



Google

Google Schweiz

Data Processing: MapReduce

- Google's batch processing tool of choice
- Users write two functions:
  - Map: Produces (key, value) pairs from input
  - Reduce: Merges (key, value) pairs from Map

	Mar 05	Mar 06	Sep 07	Apr 08
Number of jobs	72K	171K	2,217K	2,993K
Average time (seconds)	934	874	395	453
Machine years used	981	2,002	11,081	15,815
Input data read (TB)	12,571	52,254	403,152	634,920
Intermediate data (TB)	2,756	6,743	34,774	56,960
Output data written (TB)	941	2,970	14,018	25,260
Average worker machines	232	268	394	124

1 month = 1 year EGEE (2007)

1 month = 1 year EGEE (2007)

Google Schweiz

Next Generation Infrastructure

Truly global systems to span all our datacenters

- Global namespace with many replicas of data worldwide
- Support both consistent and inconsistent operations
- Users specify high-level desires:
  - "99%ile latency for accessing this data should be <50ms"
  - "Store this data on at least 2 disks in EU, 2 in U.S. & 1 in Asia"

- Increased utilization through automation
- Automatic migration, growing and shrinking of services
- Lower end-user latency
- Provide high-level programming model for data-intensive interactive services

Google

Summary

- FS gain ground in general at CCs
  - **HEPIX FSWG** recommendation: 1.**Lustre** 2.**GPFS**
    - no special HEP solution
    - tests should use experimental frameworks in the future – we have to test what people do
- Tape System@CERN
  - Tape inefficiency is homemade, but can be improved
- CNAF follows **GPFS/TSM** road
- No change in storage medias in HEP now (change in ratios)
  - Tape is not yet dead – Flash is currently too expensive

There are other big storage challenges than LHC out there:
Google ... a story of success!