

IT-ST: The Storage group

Alberto Pace on behalf of the IT-ST group



The Storage group



Mandate

- Ensure a coherent development and operation of storage services at CERN for all aspects of (physics) data
- Three goals
 - Storage
 - Distribution
 - Preservation



Three sections

- Analytics & Developments
 - Design and develop central storage services and their evolution. Operate and support the Analytics and FTS services
 - Leader: Dirk Duellmann
- File & Disk Operations
 - Operate and support the storage and file system services for physics (AFS, CEPH, CERNBOX, CASTOR and EOS)
 - Leader Massimo Lamanna
- Tape, Archives & Backups
 - Design, operate and support the archive and backup services (includes robotics, drive and media, infrastructure for backup).
 - Leader: German Cancio Melia





(IT-ST-FDO)





(IT-ST-AD)



IT-ST-FDO

Presented by Xavier Espinal

Storage Services for Physics and more

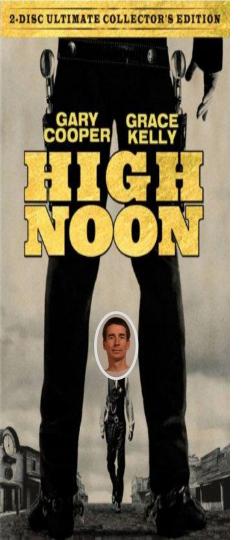




IT-ST-FDO

Storage Services for Physics and more IT-ST-FDO





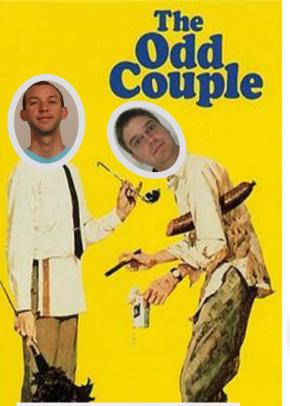
AFS (phase-out)

- Difficult task, but achievable
 - 520 TB (+16% this year vs 34% last year), 3.5 billions files (+16% vs +20% last year)
- We have solutions for nearly all current AFS usages
 - Successfully migrating web/projects/personal files to EOS and software to /cvmfs
- No hard line to "unplug" (Run3?)
- Starting a discussion with all relevant service managers concerning **home directories**:
 - Ixplus, Ixbatch, linux/win/Mac support, IT-CDA, IT-CM,...



Filer service

- The current Filer service solves the native NFS (posix) needs
- 80TB allocated and +30 projects:
 - Twiki, Puppet, MIC,LSG, gridCE, myproxy,gitlab, boinc, openshift, Indico, Inspire, etc.
- Currently running on a single NFS server over CEPH-RBD does not provide High Availablilty. We are evaluating possibilities to provide HA:
 - pNFS cluster
 - CEPH-FS



Wiki » Ceph Advisory Board »

CAB 2016-01-12 ¶

Present:

- Patrick McGarry, Red Hat (chair)
- Sage Weil, Red Hat (tech committee chair)
- Dan van der Ster, CERN (Academic Liason)

ceph @ cern

Beesly (5 PB + 433 TB, v0.94.9): Cinder (block storage volumes) Glance (images repo) Rados GW (object storage interface: S3, Swift)

-(\$)-

openstack[~]



Dwight (0.5 PB, v10.2.3): Preprod cluster for development (client side) Testing, upgrades and crazy ideas

Erin (4.2 PB, v10.2.3): New cluster for CASTOR: disk buffer/cache in front of tape drives

Flax (0.4PB, v.10.2.3 - early stage): Ceph-FS HPC cluster for QCD studies

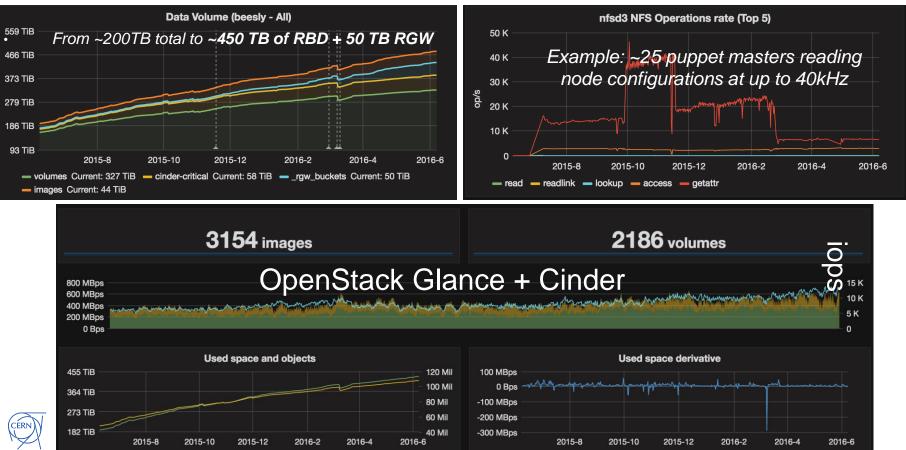
Gabe (1PB, v.10.2.3): New S3 Object Store IPV6 only

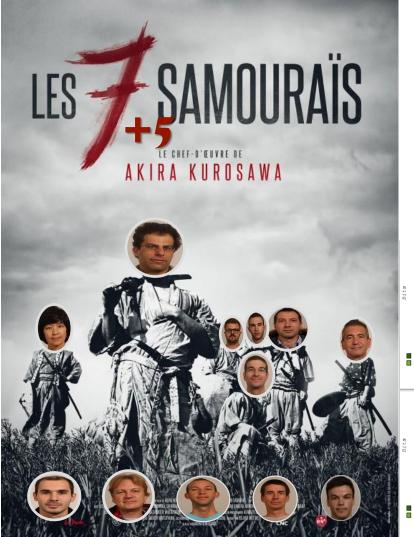


Bigbang (~30 PB, master): Playground for short term scale tests Usually when we receives new hardware











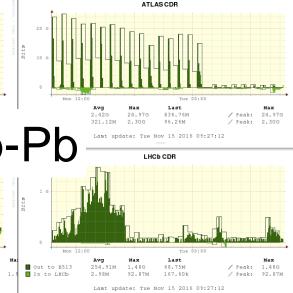
CERN Advanced STORage manager

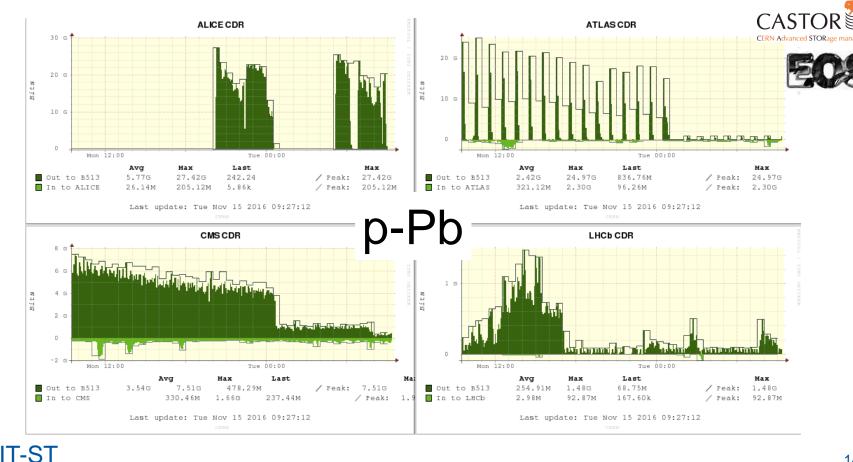


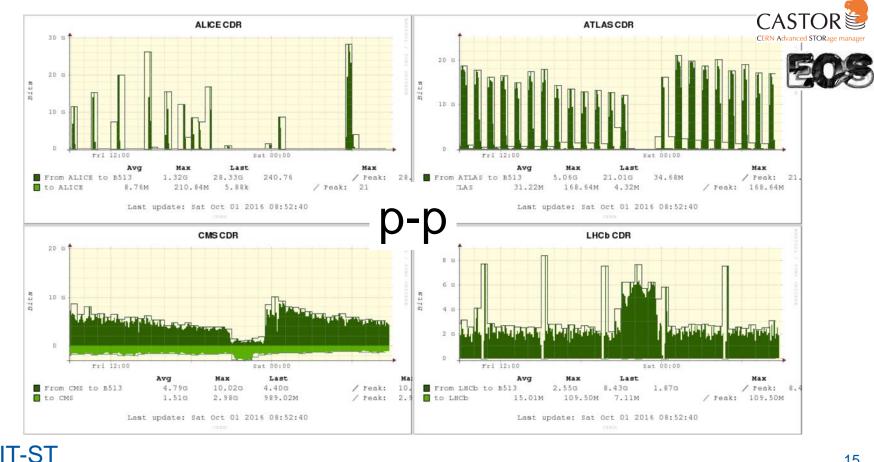


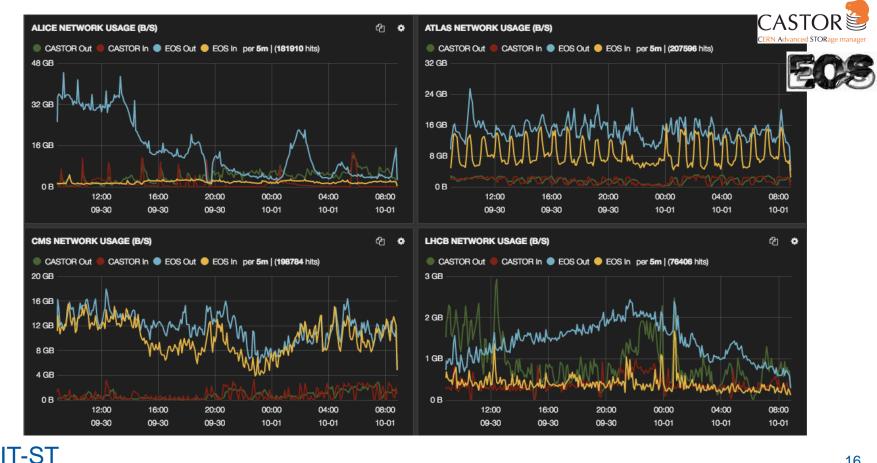
Last update: Tue Nov 15 2016 09:27:12







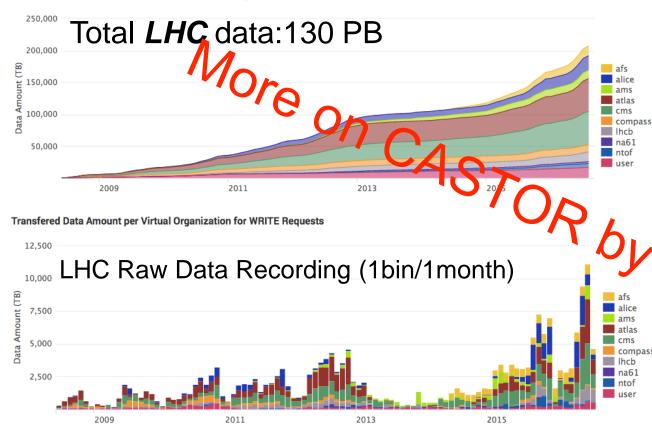




16



Accumulative Transfered Data Amount per Virtual Organization for WRITE Requests



Biggest physics-repo worldwide : 180PB and 500M files

Evolved towards a **Tape oriented** system during LS1

CEPH backend in production (Alice, Repack)

Cold by definition: high throug by thigh latency

Future evolution for a pluggable tape backend, see CTA (Julien) ¹⁷





Easily scalable (#disk #servers) Performant and manageable LHC main storage platform

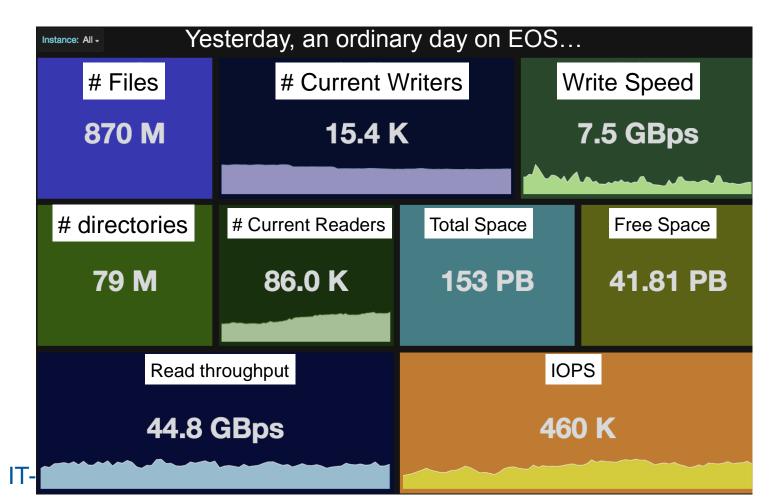


Concurrency peaks observed at +25000 Writes +100000 Reads

Throughput peaks at +80GB/S Reads







FROM THE CREATORS OF FINDING NEMO

Same PIXAR one

Large Scale Storage for (not only) LHC data



Users $\frac{6200}{(\Delta^{7d}+60)}$

#files 125M

#dirs 15M

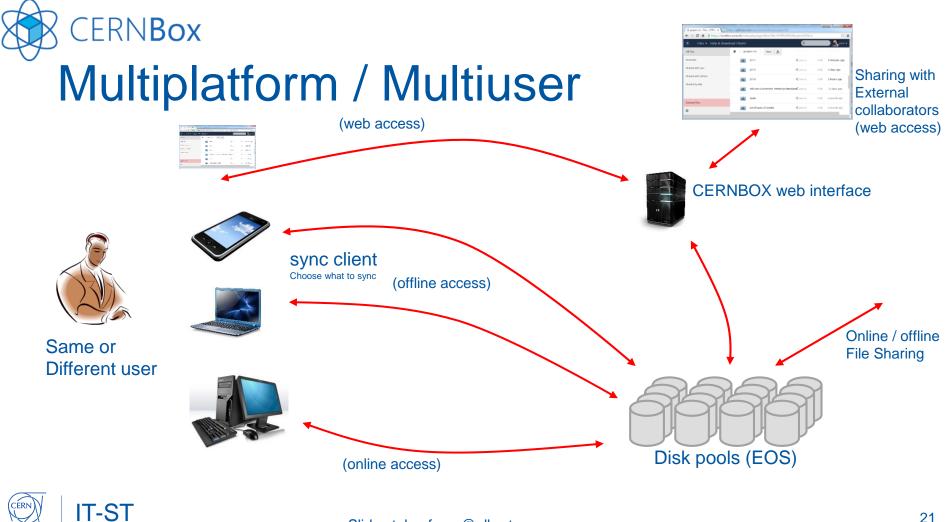
Quota 1TB/user

Used Space 310TB

Deployed Space 1.5PB Share Offline work
Community
storage
Sync SWAN

Collaborate

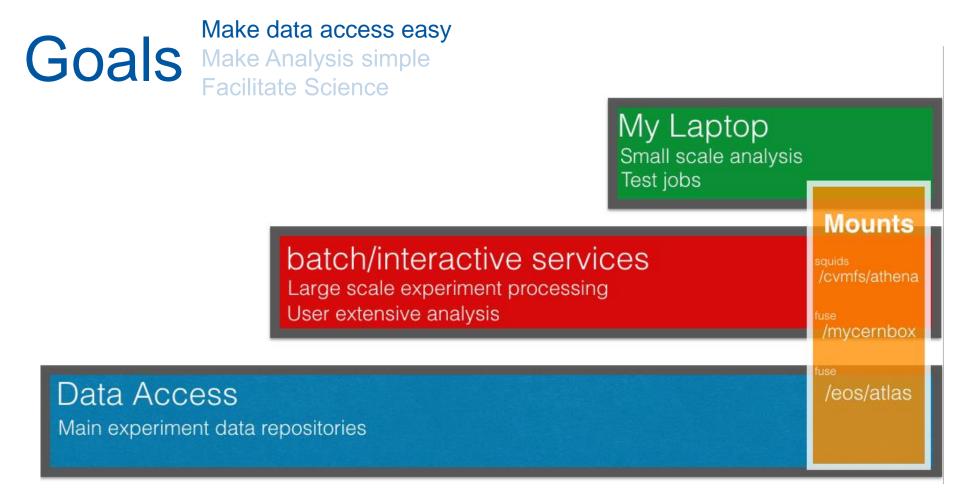














Goals Make data access easy Make Analysis simple Facilitate Science

Physicist code: **topmass.**kumac on his laptop on **/mycernbox** and sync'd via **cernbox** client Physicist identify an interesting dataset /eos/atlas/phys-top

Submit jobs to Ixbatch/wlcg to process the data EOS Fuse: /eos/atlas/phys-top EOS Fuse: /mycernbox/topmass.kumac Experiment SW: /cvmfs/athena

Results (ntuples) aggregated on /mycernbox/topmass are synced on his laptop as the ↓ if desired jobs are being completed Working on final plots on his laptop and Latex-ing the paper directly on /mycernnbox/topmass/paper

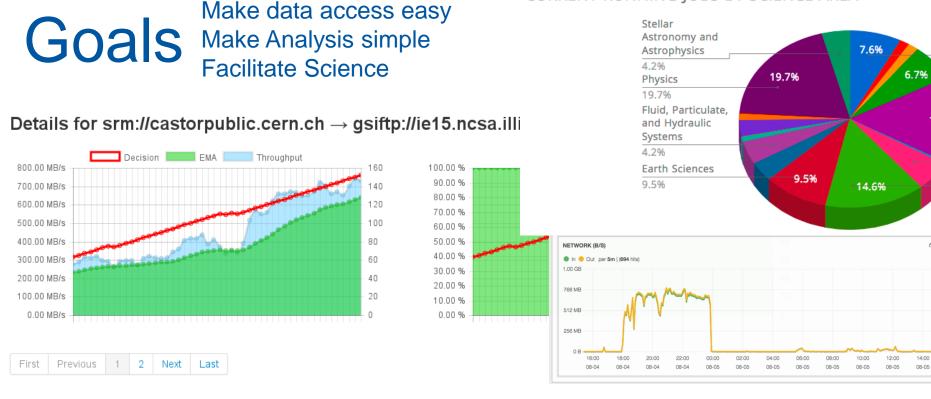




Mapping Proton Quark Structure in Momentum and Coordinate Space using PetaByte Data-Sets from the COMPASS Experiment at CERN.



CURRENT RUNNING JOBS BY SCIENCE AREA



		Success rate (last				
Timestamp	Decision	Running	Queue	1min)	Throughput	EMA
2016-08- 05T13:57:24	154	152	1898	100.00%	735.688 MB/s	648.032 MB/s

CERN

IT-ST

16%

2 0



Goals

Make data access easy Make Analysis simple Facilitate Science

- Scale-out filesystem underneath the ownCloud app, using the eosd fuse interface for file IO

aspera **RSYNC**

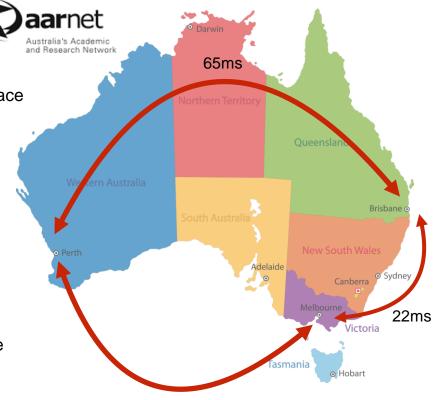
Geo-distributed setup: Brisbane, Melbourne, Perth

- ~1PB (scale to ~20PB next year)



Australian National University

- Australian National University, in Acton Canberra: mirror archives of both genome sequences and open or freely available software distributed among three sites



28



Goals

Make data access easy Make Analysis simple Facilitate Science

Components of the JRC Earth Observation Data Processing Platform (JEODPP)



rocessing <u>iterface</u> rect access interface rough HTCondor heduler for periment/projects

Patch

Joint Research Centre @V.Vasilev,F.Eyraud (JRC)

> Interactive Processing Interface web access interface for end users based on Jupyter, Leaflet and custom built image processing libraries



Distributed file system



Computing Clients

FILESYSTEM EOS 2 MGM servers 10 FST servers 240 6 TB disks

TOTAL SPACE: 1.44 PB

Current usage: 24.5M files 336TB

PROCESSING 9 2U processing servers 16 1U processing Servers

TOTAL: 600 cores 7+TB memory

IT-ST-TAB

Presented by Julien Leduc





AGE



Storage Services for Physics and more





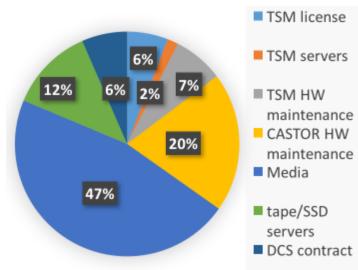
IT-ST-TAB

Tape, Archives and Backups (TAB)

- Mandate
 - Design, operate and support the archive and backup services.
 - This includes the tape-based software backend for CASTOR, tape robotics, drive and media for physics, infrastructure for backup and restore of file servers and databases.
 - "Tape from A to Z"
- People
 - 7 staff + visitors + students
 - 1 external contract
- Budget

IT-ST

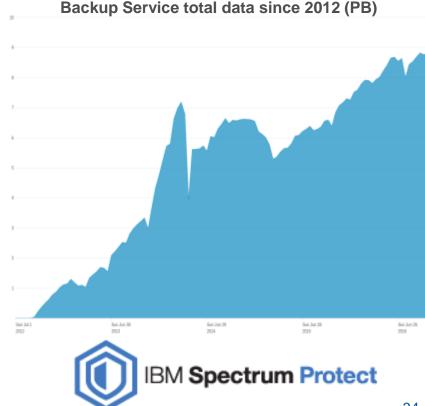
• 1.7MCHF in 2016



Backup Service

- Operate, maintain and monitor **CERN** Backup infrastructure:
 - Critical service backups
 - High reliability, sustained daily traffic: 70TB/day
- Infrastructure:
 - 2 tape libraries (9k tapes)
 - 55 tape drives
 - 20 TSM servers

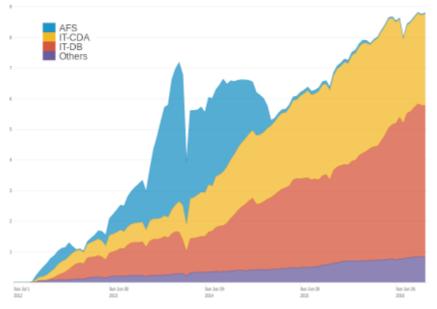




Backup Service

- TSM cost is proportional to backed up volume
 - Minimize backup volume in view of upcoming license retendering
- Ongoing efforts to limit annual growth rate: 24% (50% in the past):
 - Moved AFS backups to CASTOR
 - Worked with IT-CDA to remove redundant copies
 - Working with IT-DB to move Oracle backups out of TSM



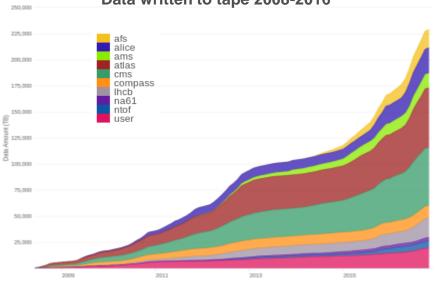


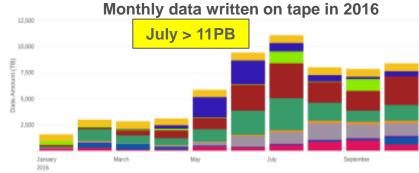


Data Archiving

- Development, operation and monitoring of CERN's tape data archive:
 - Storage and long term (ad aeternum) archiving of CERN physics data
 - High throughput, high reliability, cost efficient
- Infrastructure:
 - 7 tape libraries, 83 tape drives / tape servers, 23k tapes
 - Current use: 180PB
 - Current capacity: 0.6 EB
 - Max throughput: 25GB/s

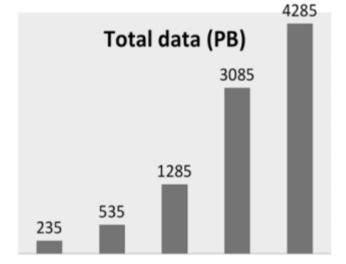






Data written to tape 2008-2016

- Anticipating the future ...
 - 2017-2018: 100PB/year
 - 2021++ : 150PB/year
- ... and preserving the past

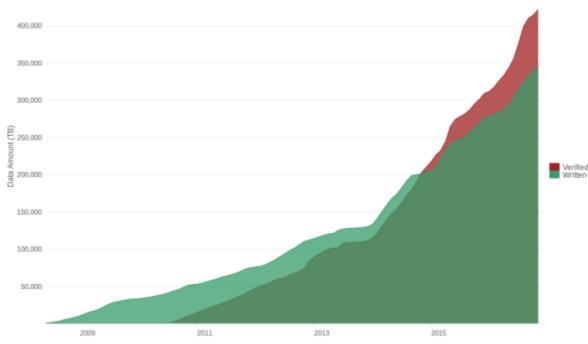


2015-2017 2018-2020 2021-2023 2024-2026 2027-2029

- Protect against bit rot (data corruption, bit flips, environmental elements, media wear out and breakage ...)
- Migrate data across technology generations, avoiding obsolescence
 - Some of our data is 40 years old

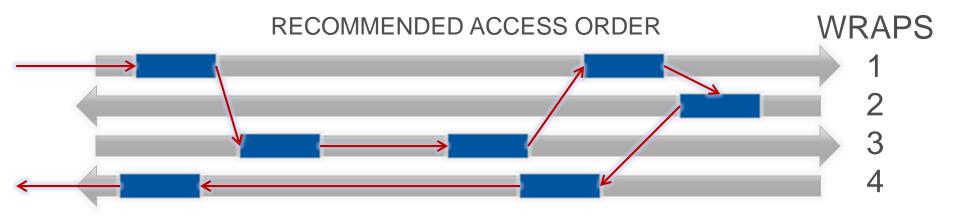


Systematic media verification



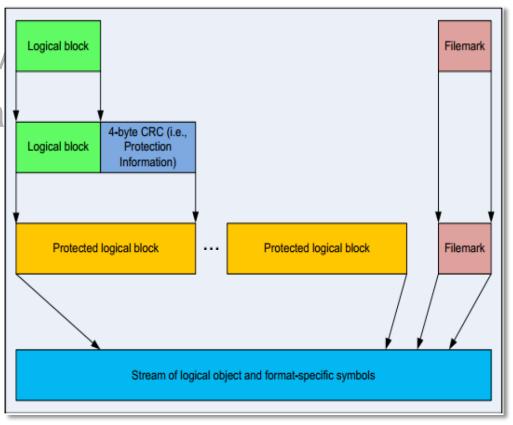


- Systematic media verification
- Optimising media access





- Systematic media v
- Optimizing media a
- Improve reliability





- Systematic media verification
- Optimizing media access 2016-01-26T05:46:03.594249+01:00 tpsrv220 tapeserverd[3335]: LVL=Info TID=3350 MSG="Logging volume statistics" firmwareVersion="460E" lifetimeBOTPasses="1486" • In bring lifetimeMOTPasses="1556" lifetimeVolumeMounts="202" lifetimeVolumeRecoveredReadErrors="167" Not good.. lifetimeVolumeRecoveredWriteErrors="30" lifetimeVolumeUnrecoveredReadErrors="4" Really bad! lifetimeVolumeUnrecoveredWriteErrors="2" validity="1" volumeManufacturingDate="20110603"

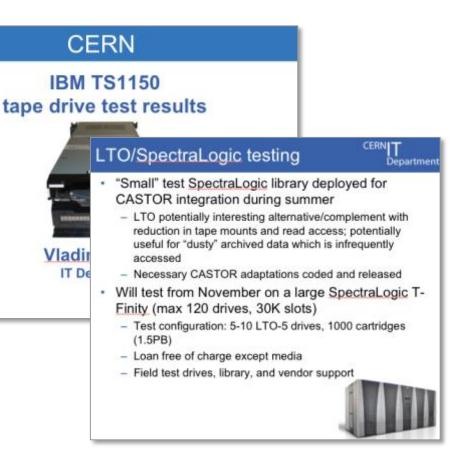


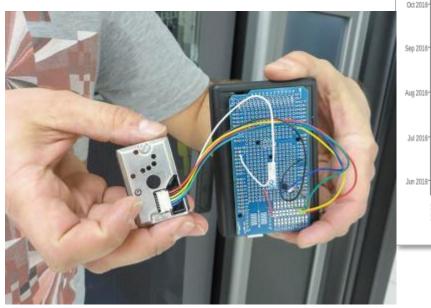
- Systematic media
- Optimizing media
- Improve reliability
- In depth performant
- Collaborations

IT-ST



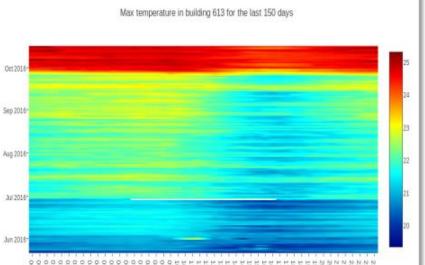
ORACLE



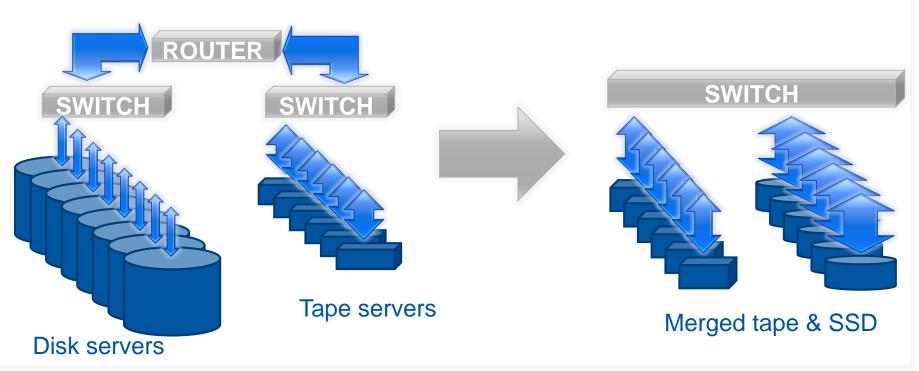


Environmental protection

Large dust particles in building 613 for the last 150 days







Large scale media migration



Data Archiving: Evolution

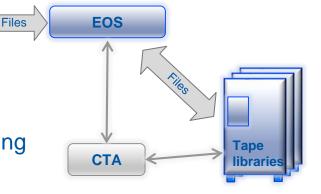
- EOS + tape ...
 - EOS is the strategic storage platform
 - Tape is the strategic long term archive medium
- EOS + tape =

T-ST

- Meet CTA : the CERN Tape Archive
- Streamline data paths, software and infrastructure



- CTA is glued to the back of EOS
- EOS manages CTA tape files as replicas
- CTA contains a catalogue of all tape files
- CTA provides optimised, preemptive scheduling



Data Archiving: Evolution

- CTA Timeline
 - End 2016: First functional prototype release
 - April 2017: First release for additional copy use cases
 - 2018: Production-ready version
 - Easy migration path from CASTOR to EOS+CTA
 - Only metadata needs to be migrated
 - CASTOR tape format will be reused



IT-ST-AD

Presented by Andreas Peters







Storage Services for Physics and more





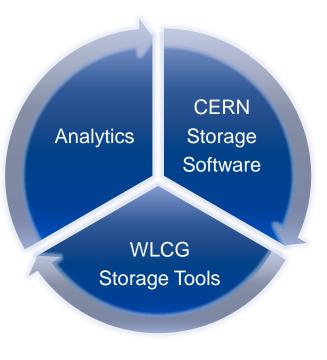
IT-ST-AD

Section Activities

- Analytics
 - Activities/Working Group CPU/IO log analysis
 - Hadoop Service
- Storage Software Development
 - CERN
 - CASTOR^{XRootD}
 - EOS / CernBOX
 - XRootD Client & Release Management
 - WLCG

IT-ST

- DPM
- Dynafed
- GFAL & DAVIX
- FTS3 Dev&Ops

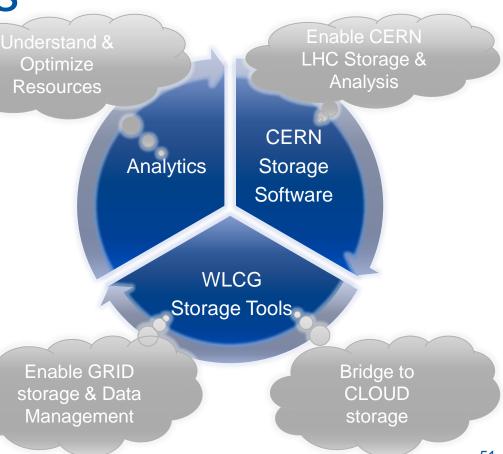


Section Activities

- Analytics
 - Activities/Working Group CPU/IO log
 - Hadoop Service
- Storage Software Development
 - CERN
 - CASTOR^{XRootD}
 - EOS / CernBOX
 - XRootD Client & Release Management
 - WLCG

IT-ST

- DPM
- Dynafed
- GFAL & DAVIX
- FTS3 Dev&Ops



Section Members

Andreas J. Peters

Dirk Düllmann Christian Nieke Rainer Többicke Luca Menicchetti



Elvin A. Sindrilaru

Analytics

CERN Storage Development



WLCG Storage Oliver Keeble Development

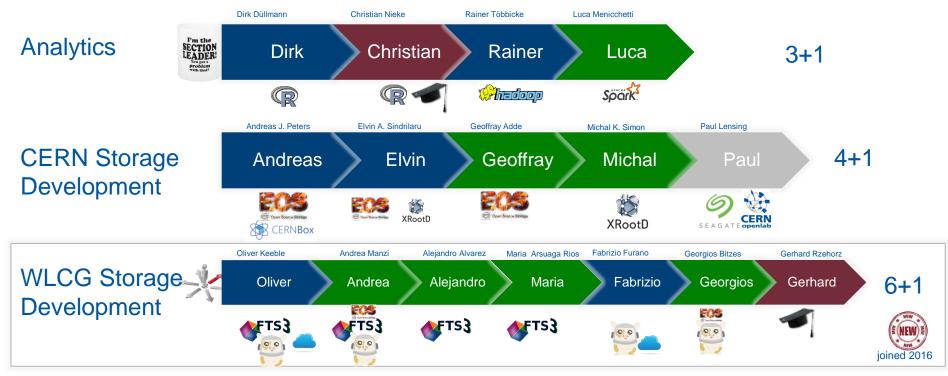
Andrea Manzi Alejandro Alvarez Fabrizio Furano Georgios Bitzes Gerhard Rzehorz Maria Arsuaga Rios

Geoffray Adde Michal K. Simon

Paul Lensing



Section Structure





Analytics – Hadoop Service



3 production cluster - joint activity of IT-ST & IT-DB

	Ixhadoop	analytix	hadalytic		
Purpose	general, de facto Atlas	monitoring, log analysis	beams dep, new development 14 2 (55) 6M 2500		
Nodes	20	38			
Size (used) PB (%)	1.5 (60)	2.5 (66)			
Files	35 M	8 M			
Jobs/week	2500	6600			

Highlights

- Analytix repository for EOS log files
- SWAN hadoop integration
- CERNbox/EOS access for map/reduce and **Spark**
- HDFS backup to Castor

Future Challenge

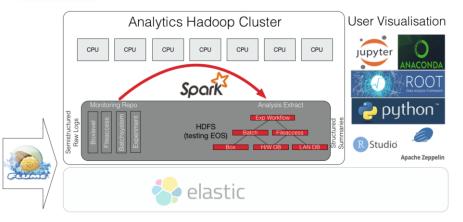
Hadoop on EOS disk server



Analytics – Log Storage & Analysis



... feed log records via IT monitoring chain



Hadoop & Friends - CHEP 2016

... fast interactive log data access

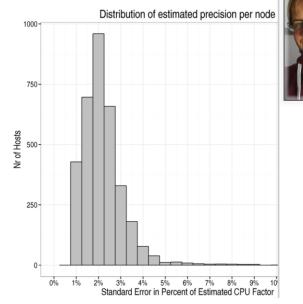
				RStudio					
	File Edit Code View	Plots Session Build Debug	Profile Tools Help					dd 🕞 🥘	
-	2.4.8818	- Go to Ne/function	Addins +					Project: (None)	
0 00	s-wb.Rmd × 9]tl.R × []uid ×				-0	Environment History Spark		-0	
	1 1 1 1 7 9 1 Prev	iew • @ •		10 Insert • 1 0 8 1 🗃	Run • 😪 • 🖄	da 🤗 SparkUl 📑 Log 🖲 🖓			
37	the second second second					local			
					2 * ×	0 eos			
38 -	· · · · (r)				0.2.1	path : chr "/eos/cms/store/	tØstr"		
	## count all eos accesses !					ruid : int 68813 24576 6881	3 6734		
40	eos % filter(ruid != 1) 5	6% count()				rgid : int 1399 1399 1399 1			
74					15 A A	td : chr "doemon.15761:4400			
	Source: query [?? x 1]					host : chr "lxfsre09a03.cer			
	Database: spark connection	naster-local[24] app-sparkly	yr local=TRUE			fid : int 422581508 4228064			
	10 C					fsid : int 10754 9482 15835 ots : num 1.47e+09 1.47e+09			
	n sdbl>				otms : num 304 876 257 680				
	1 89163713					cts : num 1.47e+89 1.47e+89			
42						ctms : num 128 363 30 807 4	82		
	· · · · (r)				0.2.1	rb ; num 0.00 0.00 0.00 Z.1	Ze+0.,		
		## top user query over the total period from spark					wb : num 7.38e+08 8.35e+05 7.04		
45	by_uid <- eos %>%					sfwdb : num 0.00 0.00 0.00			
40	group_by(secname) %>%	select(rb,wb,secname, ruid) %>%				sbwdb : num 0.00 0.00 0.00			
48									
), tot.rb-sum(rb), no.opens-ni	()) %+%			Elles Plots Parkages Hein Vie			
49	top_n(100, tot.rb)), tot.rb-sum(rb), no.opens-ni	O) %×%			Files Plots Packages Help Vie			
49 50	top_n(100, tot.rb)		()) %-%			2 New Folder 2 Upload 2 Delet		More -	
49			O) 8-8			New Folder Q Upload Q Delet	te 🛃 Rename	More -	
49 50 51 52 53	<pre>top_n(100, tot.rb) # get the data into a data df <- collect(by_uid) # print as a table</pre>		O) 16.8			2 New Folder 2 Upload 2 Delet		More -	
49 50 51 52 53 54	<pre>top_n(100, tot.rb) # get the data into a data df <- collect(by_uid)</pre>		O) 168			New Folder Upload Delet Home urc sparklyr-test Aname	te 🛃 Rename	More -	
49 50 51 52 53	<pre>top_n(100, tot.rb) # get the data into a data df <- collect(by_uid) # print as a table df</pre>		0.3 848			Vew Folder Q Upload Q Delet	te 凄 Rename	More - (Modified Oct 12, 2016, 8:48	
49 50 51 52 53 54	<pre>top_n(100, tot.rb) # get the data into a data df <- collect(by_uid) # print as a table df</pre>		toLrb	no.opens	.8.4.X	New Folder Upload Delet Mome Mare Sparklyr-test Name " glognore Image: Sparklyr test Image: Sparklyr test	Size 63.8 104.7 KB	Modified Oct 12, 2016, 8:48 Oct 13, 2016, 4:51	
49 50 51 52 53 54	top.n(100, tot.rb) # get the data into a data df <- collect(by.uid) # print as a table df secname 	from tot.wb	tot.rb edbl-	<dbb< td=""><td>.4.4.8</td><td>New Folder Upload Delet Nome are sparkly-test A Name </td><td>5ize 63 8 104.7 K8 17.7 K8</td><td>More - Modified Oct 12, 2016, 8:44 Oct 13, 2016, 4:51 Oct 13, 2016, 5:13</td></dbb<>	.4.4.8	New Folder Upload Delet Nome are sparkly-test A Name	5ize 63 8 104.7 K8 17.7 K8	More - Modified Oct 12, 2016, 8:44 Oct 13, 2016, 4:51 Oct 13, 2016, 5:13	
49 50 51 52 53 54	top_n(100, tot.rb) # get the data (nto a data df <- collect(by_uid) # print as a table df schw cmsprd	from totwb cotox 1.089070e+16	tot.rb edblo 1.039794e+16	33082382	.4.4.X	New Folder Upload Delet Amme Anne Anne	5ize 63 8 104.7 KB 17.7 KB 20.9 KB	More - Modified Oct 12, 2016, 8:44 Oct 13, 2016, 4:51 Oct 13, 2016, 5:13 Oct 18, 2016, 9:33	
49 50 51 52 53 54	top_n(100, tot.rb) # get the data into a data df <- collect(by_uid) # print as a table df secname athra cmsprd eos	Frame tot.wb (db) 1.059276+15	tot.rb etblə 1.039794e16 1.455796e15	odbb 33082382 12266595	.0.1.X	Yew Folder Upload Delet Nore Nore	63 8 104.7 KB 17.7 KB 20.9 KB 775.7 KB	More - Modified Oct 12, 2016, 8:44 Oct 13, 2016, 4:51 Oct 13, 2016, 5:13 Oct 18, 2016, 9:33 Oct 18, 2016, 10:0	
49 50 51 52 53 54	top_n(100, tot.rb) # get the data (nto a data off < collect(by_uid) # print as a table df example eas physick k	Frame sot.wb -cdto- 1.059070e+16 1.457026+15 4.636716e+15	toLrb edbb- 1.039794e+16 1.455796e+15 1.377929e+15	33082382 12266595 3748689		Openet Opene Opene Opene <td>5ize 63 8 104.7 KB 17.7 KB 20.9 KB</td> <td>More - Modified Oct 12, 2016, 8:44 Oct 13, 2016, 4:51 Oct 13, 2016, 5:13 Oct 18, 2016, 9:33 Oct 18, 2016, 10:0</td>	5ize 63 8 104.7 KB 17.7 KB 20.9 KB	More - Modified Oct 12, 2016, 8:44 Oct 13, 2016, 4:51 Oct 13, 2016, 5:13 Oct 18, 2016, 9:33 Oct 18, 2016, 10:0	
49 50 51 52 53 54	top_n(100, tot.rb) # get the data into a data df < collect(by_uid) # print as a table df 	Frome Color 1.0880704-16 1.457026+15 4.638718+15 0.000006+00	sot.rb - cpbb- 1.039794e-16 1.455796e-15 1.025526e-15	4dbb 33082382 12266595 3748689 348747	4.4.3	New Folder Optional Optional A Nome Lipskin/Fred A Nome Lipskin/Fred B Nome Lipskin/Fred B Notes Lipskin/Fred B Notes Lipskin/Fred B Notes Lipskin/Fred B Rotes Lipskin/Fred	e e Rename Size 63 8 104.7 KB 17.7 KB 20.9 KB 775.7 KB 1.9 KB	Modified Oct 12, 2016, 8:44 Oct 13, 2016, 4:51 Oct 18, 2016, 9:32 Oct 18, 2016, 10:0 Oct 18, 2016, 10:0	
49 50 51 52 53 54	top_n(100, tot.rb) # get the data into a data df < collect(by_uid) # print as a table df secame cmspd esi phedex huapch syddat	From: toLmb dBib 1.089070e16 1.457024e15 4.635714e15 0.00000e+00 0.00000e+00	toLrb adds 1.039794e-16 1.457796e-15 1.377929e-15 1.025526e-15 9.210677e-14	4005 33082382 12266595 3748689 348747 219722	8 + ×	the rade 0 byte 0 byte there is a subsystem there is a sub	e Rename Size 63 8 104.7 KB 10.7 KB 20.9 KB 775.7 KB 1.9 KB 205 8	Modified Oct 12, 2016, 8:44 Oct 13, 2016, 4:51 Oct 13, 2016, 5:12 Oct 18, 2016, 9:32 Oct 18, 2016, 10:0 Oct 18, 2016, 10:0 Oct 18, 2016, 10:0 Oct 12, 2016, 8:44	
49 50 51 52 53 54	top_r(100, tot.m) # get the data into a data dr < coltection_a data # print as a table # esta- esta- phetex huangch spadrat cmmod	from: color- 1.0867/024-15 4.0357784-15 0.0000004-00 0.000004-00 1.198904-10	tot.rb edbh 1.030794e-16 1.455796e-15 1.025556e-15 9.210677e-14 5.46222ae-14	edbb- 33082382 12266595 3748689 348747 219722 161611	a s x	Prender U polat O este Anne Anne	e e Rename Size 63 B 104.7 KB 17.7 KB 20.9 KB 775.7 KB 1.9 KB 205 B 1.4 KB	More - Modified Oct 12, 2016, 8:44 Oct 13, 2016, 45:1 Oct 13, 2016, 51:1 Oct 18, 2016, 10:1 Oct 18, 2016, 10:0 Oct 18, 2016, 10:0 Oct 12, 2016, 8:44 Oct 12, 2016, 8:49 Oct 12, 2016, 2:29	
49 50 51 52 53 54	top_r(100, tot.mb) d get the data into a data d get collect(by_r(1d)) d get collect(by_r(1d)) d get est sectame est base ba	frame to:wb 1.0897070e16 1.437020e15 4.335718e15 0.00000e400 0.00000e400 1.198904e10 0.00000e400	tot.rb +c000 1.039794e-16 1.457798e-15 1.025556e-15 9.210677e-14 5.862523e-14 5.862523e-14	4865 33082382 12266595 3748689 348747 219722 161611 136124	8 8 X	the rade 0 byte 0 byte there is a subsystem there is a sub	e Rename Size 63 8 104.7 KB 10.7 KB 20.9 KB 775.7 KB 1.9 KB 205 8	More - Modified Oct 12, 2016, 8:44 Oct 13, 2016, 45:1 Oct 13, 2016, 51:1 Oct 18, 2016, 10:1 Oct 18, 2016, 10:0 Oct 18, 2016, 10:0 Oct 12, 2016, 8:44 Oct 12, 2016, 8:49 Oct 12, 2016, 2:29	
49 50 51 52 53 54	top_r(100, tot.m) f of the data into a data df collect(by_uid) error compared error compared error photex humagch spatrat compared photex humagch spatrat compared photex photex humagch spatrat compared photex humagch spatrat compared photex humagch spatrat humagch spatrat humagch humag	from 1.055070e+16 1.455070e+16 1.455726e+15 4.05000e+00 0.00000e+00 1.19894e+10 0.00000e+00	tot.rb wibbs 1.030794e-16 1.475798e-15 1.075929e-15 9.210677e-14 5.862528e-14 5.863534e-14 4.575327e-14	-0007 33082362 12266595 3748689 348747 219722 161611 136124 685484	8-1-X	Anne Anne	e e Rename Size 63 B 104.7 KB 17.7 KB 20.9 KB 775.7 KB 1.9 KB 205 B 1.4 KB	More - Modified Oct 12, 2016, 8:44 Oct 13, 2016, 45:1 Oct 13, 2016, 51:1 Oct 18, 2016, 10:1 Oct 18, 2016, 10:0 Oct 18, 2016, 10:0 Oct 12, 2016, 8:44 Oct 12, 2016, 8:49 Oct 12, 2016, 2:29	
49 50 51 52 53 54	top_r(100, tot.mb) d get the data into a data d get collect(by_r(1d)) d get collect(by_r(1d)) d get est sectame est base ba	frame to:wb 1.0897070e16 1.437020e15 4.335718e15 0.00000e400 0.00000e400 1.198904e10 0.00000e400	tot.rb +c000 1.039794e-16 1.457798e-15 1.025556e-15 9.210677e-14 5.862223e-14 5.862223e-14	4865 33082382 12266595 3748689 348747 219722 161611 136124	d-4.8	Anne Anne	e e Rename Size 63 B 104.7 KB 17.7 KB 20.9 KB 775.7 KB 1.9 KB 205 B 1.4 KB		
49 50 51 52 53 54	top_r(100, tot.mb) df == collect(by_r(10) df == collect(by_r(10) df == collect(by_r(10) df == cmprd ess photo humg(ch symbol symbol symbol biologic photo biologic symbol sym	Frame	tot.rb +cbi 1.059794+16 1.057994+15 1.025594+15 9.210677e-14 5.882534+14 4.3582534+14 4.35727e-14 4.459310e-14	-dbb- 33082382 12266595 3748689 348747 219722 161611 136124 685484 119105		Anne Anne	e e Rename Size 63 B 104.7 KB 17.7 KB 20.9 KB 775.7 KB 1.9 KB 205 B 1.4 KB	More - Modified Oct 12, 2016, 8:44 Oct 13, 2016, 45:1 Oct 13, 2016, 51:1 Oct 18, 2016, 10:1 Oct 18, 2016, 10:0 Oct 18, 2016, 10:0 Oct 12, 2016, 8:44 Oct 12, 2016, 8:49 Oct 12, 2016, 2:29	
49 50 51 52 53 54 55	top_((10), tot.m) top_((10), tot.m) df collect(0), und) df collect(0), und) errors comport ers photox humgch spatrat compol ploast georgia alant? twood 1-10 of 100 rows	Frame	tot.rb +cbi 1.059794+16 1.057994+15 1.025594+15 9.210677e-14 5.882534+14 4.3582534+14 4.35727e-14 4.459310e-14	60b5 33062282 3748689 348747 219722 161611 136124 685484 119105 609334	6 _ 10 Next	Anne Anne	e e Rename Size 63 B 104.7 KB 17.7 KB 20.9 KB 775.7 KB 1.9 KB 205 B 1.4 KB	More - Modified Oct 12, 2016, 8:44 Oct 13, 2016, 45:1 Oct 13, 2016, 51:1 Oct 18, 2016, 10:1 Oct 18, 2016, 10:0 Oct 18, 2016, 10:0 Oct 12, 2016, 8:44 Oct 12, 2016, 8:49 Oct 12, 2016, 2:29	
49 50 51 52 53 54 55 55 55	top_r(100, tot.m) is part the data into a data df ~ collect(by_uid) prime as a table df cmprd ess phetex humgch sgadat sgadat sgadat shant picoll authint strong	Frame	tot.rb +cbi 1.059794+16 1.057994+15 1.025594+15 9.210677e-14 5.882534+14 4.3582534+14 4.35727e-14 4.459310e-14	60b5 33062282 3748689 348747 219722 161611 136124 685484 119105 609334		Anne Anne	e e Rename Size 63 B 104.7 KB 17.7 KB 20.9 KB 775.7 KB 1.9 KB 205 B 1.4 KB	Morilled Oct 12, 2016, 8-48 Oct 13, 2016, 513 Oct 13, 2016, 513 Oct 18, 2016, 513 Oct 18, 2016, 10:0 Oct 18, 2016, 10:0 Oct 12, 2016, 8-44 Oct 12, 2016, 2-25	
49 50 51 52 53 54 55 55 56 -	top_f(100, tot.m) top_f(100, tot.m) fr - collect(Dy_i(d) fr - collect(Dy_i(d) fr - co	Frame 1.08070416 1.437024415 4.43717415 0.000004400 1.19894410 0.000004400 0.00000400 0.00000400 0.00000400	tot.rb +cbi 1.059794+16 1.057994+15 1.025594+15 9.210677e-14 5.882534+14 4.3582534+14 4.35727e-14 4.459310e-14	60b5 33062282 3748689 348747 219722 161611 136124 685484 119105 609334	6 _ 10 Next	Anne Anne	e e Rename Size 63 B 104.7 KB 17.7 KB 20.9 KB 775.7 KB 1.9 KB 205 B 1.4 KB	More - Modified Oct 12, 2016, 8:4 Oct 13, 2016, 4:5 Oct 13, 2016, 9:1 Oct 18, 2016, 9:1 Oct 18, 2016, 10 Oct 18, 2016, 10 Oct 12, 2016, 8:4 Oct 12, 2016, 8:4 Oct 12, 2016, 2:2	



Analytics – Performance Projects

PhD Christian Nieke

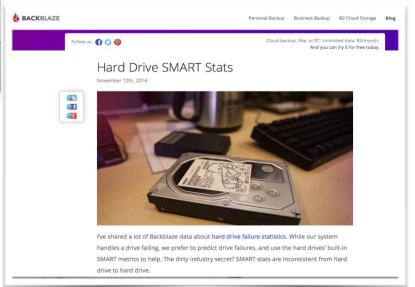
IT-ST



Passive CPU benchmark

performance prediction error better than 5%

Study proposed by Dirk Düllmann and Darrell Long (University of California)

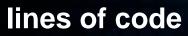


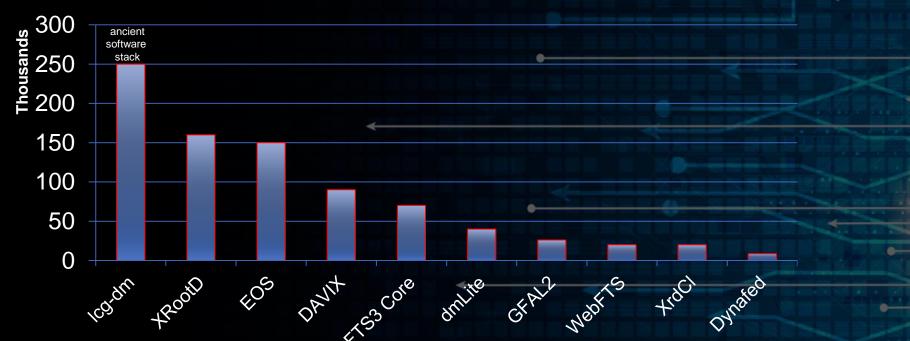
Harddrive failure analysis

Evaluate enterprise, consumer, network drives For service planning & predictive maintenance



Open Source Software Projects







60

XRootD

What is that? What is that for?

- provides high performance, scalable fault tolerant access to data repositories
- main protocol for WLCG WAN access latency optimized
 - >>99% of LAN data access at CERN via XRootD protocol
 - tight integration in ROOT framework
- provides a data oriented client-server C++ framework(plugin structure) used by
 - EOS
 - LSST
- Participation via collaboration core development SLAC

Our Responsibilities

- client development since 2010
- release management & distribution

Good to know ...

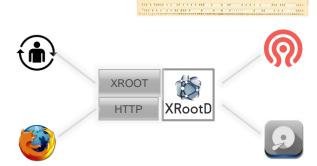
- since 2010 XRootD V3 production (e.g. EOS)
- since 2014 XRootD V4 IPV6, FileCache, HTTP(S) bridge, CEPH bridge production (e.g. CASTOR, Ixplus, LHC exp.) wire protocol security added V4.5
- **V5** expected in 2017

Future Challenges...

IT-ST

multi-source support, TPC²

GITHUB https://github.com/xrootd/xrootd.git 1000 source files – 160k lines of code





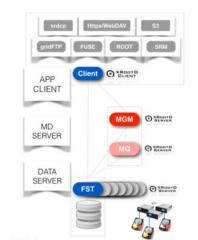
EOS



What is that?

IT-ST

- Large scale data storage system (150 PB@CERN)
 - Multi protocol XRootD HTTP(S) WebDAV, FUSE
 - Strong Authentication KRB5 X509
 - Flexible provisioning and life-cycle management
 - Large user community features (limits, quotas)
- since 2010 developed in-house in storage group
- since 2011 production service at CERN
 - +11 HEP EOS sites not at CERN LIBNL, ORNL, SUBATECH, IHEP, SASKE, FNAL, UNAM, MEPHI, SINICA, INFNTS, UNIVE
- simple scalable architecture
 - Meta data scale-up 0.5b files
 - Data scale-out 15k disks
 - Messaging 1.5Mhz pipelined





October 2016 150 PB disk space >110 PB used

• 1143 storage nodes

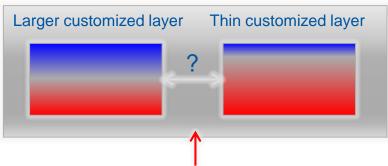
1y growth rates: +85% more files +50% more space used



EOS

What is the current development strategy?

- Evolve flexible storage infrastructure as *scale-out solution* towards exabyte scale
 - one storage many views on data analysis/analytics service integration
 - suitable for LAN & WAN deployments
 - enable cost optimized deployments
- Evolve file-system interface /eos
 - multi-platform access using standard protocol bridges CIFS, NFS, HTTP
- Evolve scale-out storage back-end support
 - object disks, object stores, public cloud, cold storage
- Evolve namespace scalability
 - scale-out implementation
- **Review** regularly break-point between in-house and community open-source layers







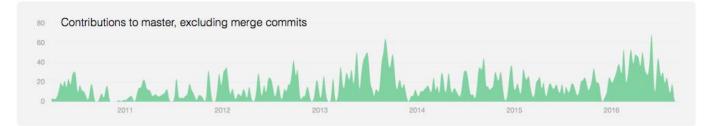


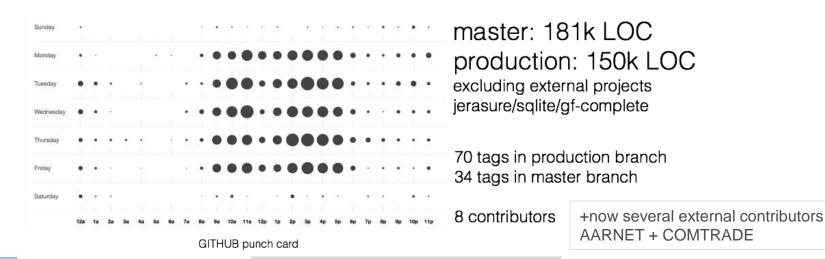


EOS Development

IT-ST







GITHUB https://github.com/cern-eos/eos

EOS Challenges



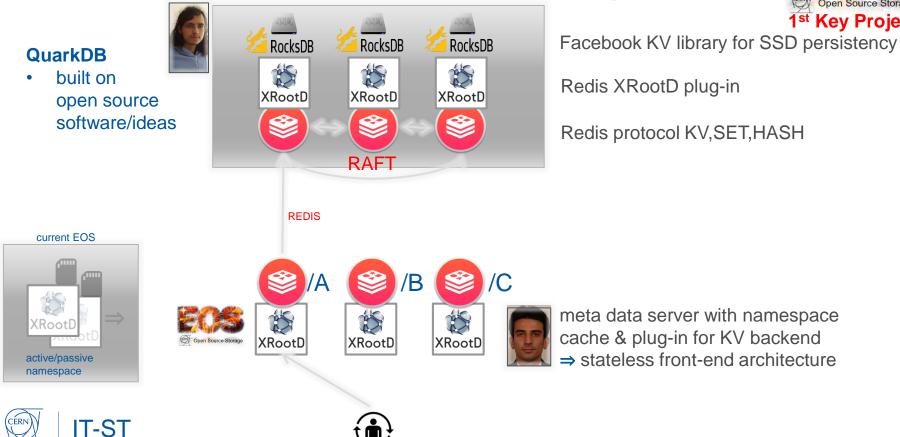
- 1 Meta data *scale-up* \Rightarrow *scale-out*
 - in-memory namespace ⇒ in-memory scale-up namespace cache + persistency in scale-out KV store
 1st key project: EOS Citrine & QuarkDB
- 2 Remote access APIs ⇒ filesystem API

2nd key project: EOS FUSE rewrite



EOS Scale-Out Namespace





65

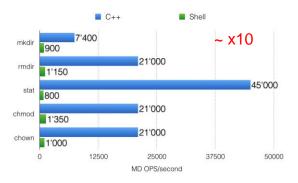
EOS FUSE improvements



- FUSE was added as an external layer to an existing server API [path-inode translations]
 - 1st implementation 2011 clone of XRootD FUSE xrootdfs C-interface
 - 2nd implementation 2016 introducing many latency optimizations and pure C++ gained 10x + more POSIX
 - meta-data consistency very difficult/race conditions without server side changes MD updates still synchronous
 - 3rd implementation on the way gain 10x + POSIX
 - client leases, async meta-data pipelining

T-ST

- optional persistent meta-data & data cache
- libfuse2 ⇒ libfuse3 21k seq IOPS ⇒ 500k seq IOPS [wb cache]







What is that?



- CERN **D**isk **P**ool **M**anager used in many GRID sites (outside CERN)
- originally storage software fork from CASTOR1 code developed since 16 years
- the Tier-2 GRID storage solution deployed at 160 sites world-wide providing 70 PB of storage space
- First refactoring included dmLite stack as plug-in architecture in 2011
- Latest release **1.9** includes now DOME component 2016
 - allows to to remove old legacy code & daemons ⇒ simplify and consolidate the architecture
 - better modularization
 - single internal communication protocol
 - SRM less operation
 - quota, space reporting, caching hooks
- Caching laboratory for 'unmanaged' Tier-2 storage, low-effort deployment, long-term consolidation
- DPM Workshop Paris https://indico.cern.ch/event/559673/



•

Key Components to Cloud Storage

- DAVIX
 - WebDAV/S3 C++ client integrated in ROOT
 - developed since 2012 focus on long term stability
 - latest features
 - compatible with Azure & AWS v4 authentication
 - performance optimizations by range coalescing and request parallelization
- DynaFED

IT-ST





- Grid file access library
 - provides protocol abstraction layer [⇒ FTS]

https://gitlab.cern.ch/dmc/gfal2

- Provides in-memory cached namespace federation of multiple HTTP enabled storage systems
 - Provides a Grid⇒Cloud bridging functionality translating X509 identities to signed URLs using DAVIX



FTS³

What is that?

 3rd generation of file transfer service building core platform for scheduled LHC data transfers

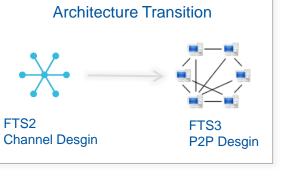
http://fts3-service.web.cern.ch

- developed since a 2001 at CERN
- 1st production release 2013
- key component in WLCG
- · transfers hundreds of peta bytes per year worldwide
- Latest release 3.5
 - Transfer Optimization based on network performance analysis
 - Decision visualization

Future Challenges

IT-ST

- Boost Scalability of database centric design
 - Initiated R&D project Flutter trying more radical redesign solutions
- HPC integration (dual credential support)
- Alarms & Containerization





FTS Optimizer Visualization

FTS 3

Details for srm://ccsrm.in2p3.fr \rightarrow srm://atlassrm-fzk.gridka.de







Optimizer Decision Visualization

First Previous 1 Next Last

Timestamp	Decision	Running	Succ	ess rate (last 1min)	Throughput	EMA	Diff	Explanation
2016-10-18T11:06:2	13	21	3940	100.00%	48.359 MB/s	48.426 MB/s	-1	Good link efficiency, throughput deterioration
2016-10-18T11:05:1	14	14	3948	100.00%	89.522 MB/s	48.434 MB/s	1	Good link efficiency, current average throughput is larger than the preceding average
2016-10-18T11:04:0	13	14	3968	100.00%	42.308 MB/s	43.869 MB/s	-1	Good link efficiency, throughput deterioration
2016-10-18T11:02:5	14	13	3986	100.00%	70.434 MB/s	44.042 MB/s	1	Good link efficiency, current average throughput is larger than the preceding average
2016-10-18T11:01:4	13	14	4006	99.00%	57.463 MB/s	41.109 MB/s	1	Good link efficiency, current average throughput is larger than the preceding average

Optimizer Status Explanation



FTS Operations



- AD responsible for operations since January '16
 - 2 cluster production & pilot
 - Moved to CC7
 - IPV6 enabled
 - Service Monitoring

https://monit.cern.ch/app/kibana#/dashboard/_project-FTS-Service-Level

• Transfer Log archival in EOS on the way





FTS @ CERN last week stats

O MONIT FTS Dashboard Navigation

1 ×

Home - Overview - Transfer Plots - Matrix View - Failures - Custom Views - Servers Configuration



WebFTS & LastMile

- WebFTS
 - end-user transfer portal
 - targets small VOs
- LastMile
 - end-point less FTS transfers
 - transfer data to/from workstations/laptops run a client – not a server



https://webfts.cern.ch

What is WebFTS? WebFTS is a file transfer and management solution which allows users to invoke reliable, managed data transfers on distributed infrastructures. Created following simplicity and efficiency oriteria, WebFTS allows the user to access and interact with multiple storage elements. Their content becomes browsable and different filters can be applied to get a set of files to accessing the detailed status of the different transfers and resubmitting any of them with only one click. The "transfer engine" used is FTS3, the service responsible for distributing the majority of LHC data across WLCG infrastructure. This Content webFTS with reliable, multi-protocol (gridftp, srm, http, xrootol, adaptive), optimised data transfers

IT-ST-AD Collaborations

EOS white-paper – now core software contributor

EOS OpenKinetic interface

EOS collaboration

EOS/CERNbox collaboration

DPM Grid storage in the cloud



File Serve

File Serve

Dynafed demonstrator

COMTRADE

European

Joint Research Centre

Open Telekom Cloud Commission

aarnet

Australia's Academic and Research Network

> University of Victoria (CA), INFN and Belle-II https://indico.cern.ch/event/394788

Data Intensive Cloud Research PhD focused on modeling ATLAS performance

IT-ST



www.cern.ch

IT-Storage Plans for 2017

- EOS as the strategic storage platform
 - · New, scalable name server (new REDIS-based implementation)
 - Improved FUSE to satisfy more requirements
- CERNBOX as the strategic service to access EOS storage
 - Offer migration paths out of AFS (target = EOS)
 - · Start a discussion with relevant service managers on the home directory service
 - · Define an architecture to host lxplus home directory
 - Improve Windows support, Samba service for Windows Clients
- CERN Tape Archive
 - First internal release interfacing EOS planned end 2016 using IBMLIB1 (test library)
 - Apr 2017: D1T1 and D0T1 workflows available in EOS
- TSM
 - New license to be negotiated, investigate alternative options for Oracle backups
- FTS: improvements on manageability and scalability
 - Files to/from laptops
- DPM: Consolidation

IT-ST

- Going SRM-less with DPM
- CEPH the backend storage solution
 - Understand if CEPHFS can replace the Filer service

Strategic Vision for IT-ST

- Build a flexible storage infrastructure
 - Unlimited storage for LHC experiments, at exabyte scale
 - Disk pools (EOS on-demand reliability, on-demand performance)
 - CERN Tape Archive (high reliability, low cost, data preservation)
 - Backbone of the CERN cloud (CEPH block storage for Openstack and Filers)
 - Cluster(s) for analytics (provided with Hadoop)
 - A generic home directory infrastructure

IT-ST

- Fuse mounts, NFS and SMB exports, (and HADOOP).
- CERNBOX (Sync client and Web/HTTP/DAV Access)
- Maintain and enhance (grid) data management tools
 - To empower global scientific data workflows
 - Data transfers (FTS), data caches (DPM, CVMFS), ...



www.cern.ch