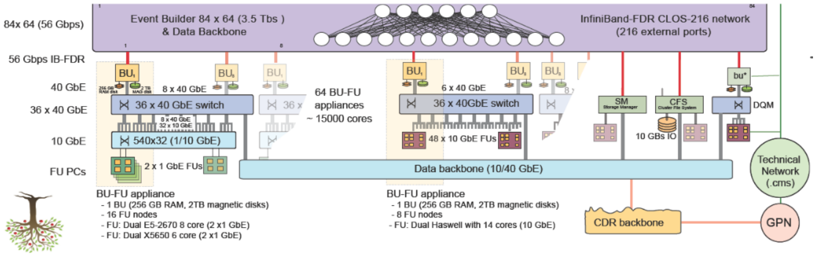
Online data handling and storage at the CMS experiment

Lavinia Darlea, on behalf of CMS DAQ Group

MIT/DAQ CMS

April 10, 2015

Introduction



Storage and Transfer System in the DAQ chain

- input: end of the DAQ chain, as described by Emilio* in the previous talk
- last part of the data flow: ensure safe storage and transfer to Tier0

**E. Meschi, File-based data flow in the CMS Filter Farm, CHEP 2015*

Storage and Transfer System Role

3 Implementation Stages

- merge the filter units output as to obtain 1 data and 1 metadata file/LS/stream
- buffer the data should the connection with Tier0 at CERN be lost
- copy the final files according to their intended destination:
  - Tier0 for the main data streams
  - various sub–detectors for online consumption: DQM, EventDisplay
  - store locally for local calibration of various sub–detectors
- ensure hand–shake with Tier0 for proper accountability

Storage and Transfer System Requirements

Merger System

- "merge" data at the BU level such as to obtain 1 file/BU/LS/Stream (mini–merger)
- centralize and merge all the BU outputs such as to obtain 1 file/LS/Stream (macro–merger)
- latency: a maximum of 2LS (1LS = 23s) delay in the macro–merger is considered acceptable
- provide input for the online monitoring system – 1 additional metadata file per data file (see next talk by Srecko*)
- not only "concatenate", but deal with special files, such as histograms and jsn files

**S. Morovic, A scalable monitoring for the CMS Filter Farm based on elasticsearch, CHEP 2015*

Storage and Transfer System Requirements

Storage and Transfer

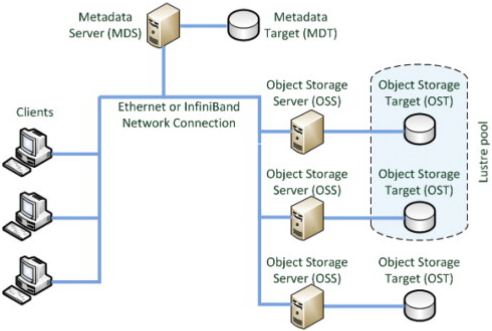
- buffer a minimum of 3 days of continuous running (estimated 250TB)
- aggregated SM input from the 62 BUs is expected to reach a maximum of 2GB/s – mini merger write to LFS (Lustre FS)
- the macro–merger needs to consume this data online (2GB/s read the fragments, 2GB/s write the final merged file): 4GB/s(*)
- the transfer system is expected to transfer most of the data to Tier0 at 1GB/s
- overall: LFS needs to guarantee a total of sustained 7GB/s parallel r/w

Mergers

2 available options

- “A”dditive
  - mini–mergers write a file/BU/LS/Stream, macro–merger merges them and makes them available for the TS
  - easy debugging, reliable, “standard” logic
- “C”opyless
  - mini–mergers write in parallel in the final file, macro–merger checks for completion and makes it available for the transfer system
  - reduce the required bandwidth with 4GB/s, fast due to parallel writing in the same file, more sensitive to corruption

Storage


Metadata Server (MDS)
Metadata Target (MDT)
Clients
Ethernet or InfiniBand Network Connection
Object Storage Server (OSS)
Object Storage Target (OST)
Object Storage Server (OSS)
Object Storage Target (OST)
Lustre pool
Object Storage Server (OSS)
Object Storage Target (OST)

Lustre FS architecture

- MDT: E2724, 16 drives of 1TB in one volume group plus 8 hot spares
- 2 OST controllers E5560 + expansion shelves DE6600, each 60 disks of 2TB
- servers: 6 DELL R720

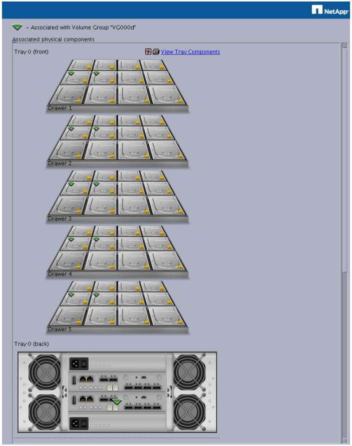
Front OST

Disk shelves

Storage

High Availability

- all devices are dual powered (normal and UPS)
- all servers configured in active/passive failover mode
- volumes repartition to provide full shelf failure redundancy
- LFS availability: 40GE and InfiniBand (56Gb) data networks



– Associated with Volume Group "VG000d"

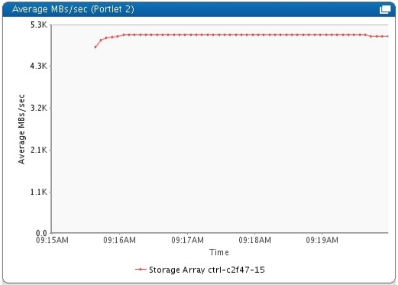
Associated physical components

Tray 0 (front)

Tray 0 (back)

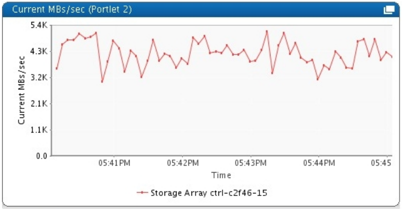
Volumes configuration

Storage – Bandwidth Validation



Commissioning Acceptance

Proven steady 10GB/s rate in r/w mode



Merger emulation

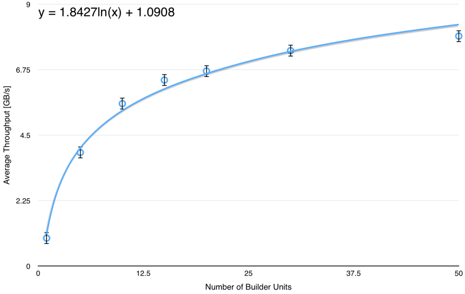
Proven steady 7.5GB/s rate

Transfer System

Inheritance and Progress

- use old transfer system as a base
- new features have been added
  - identify and set the final destination of each stream per run
  - new logic in the bookkeeping and hand–shake protocol between the CMS site and Tier0

Conclusion

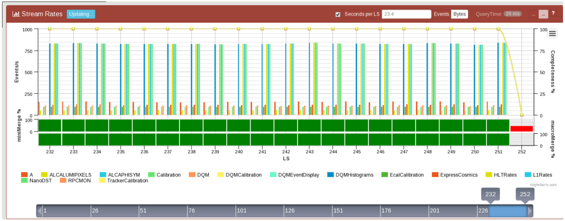


LFS bandwidth benchmarking

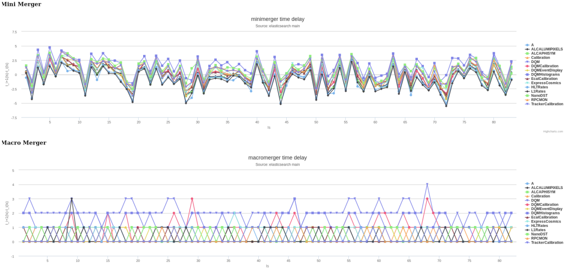
LFS Validation

- usable space of 350TB
- tests done with different number of BU units
- obvious non–linear behaviour with the number of BUs
- saturation is expected around 8.5GB/s

Conclusion



Mergers monitoring sample



Mini Merger

Macro Merger

Mergers delays sample

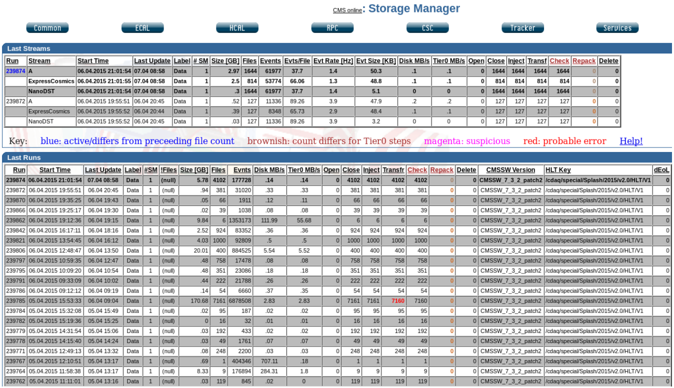
Mergers Validation

- stable behaviour in 3 months of cosmic runs
- general latencies within the requirements
- proven reliability and availability

Conclusion

Transfer System

- successfully upgraded to transfer DAQ2 merged files
- work in progress: benchmarking and central management



CMS online: Storage Manager

Last Streams

Run	Stream	Start Time	Last Update	Label	# SM	Size [GB]	Files	Events	Evts/File	Evt Rate [Hz]	Evt Size [KB]	Disk MB/s	Tier0 MB/s	Open	Close	Inject	Transf	Check	Repack	Delete
239874	A	06.04.2015 21:01:54	07.04 08:58	Data	1	2.97	1644	61977	37.7	1.4	50.3	.1	.1	0	1644	1644	1644	1644	0	0
	ExpressCosmics	06.04.2015 21:01:55	07.04 08:58	Data	1	2.5	814	53774	66.06	1.3	48.8	.1	.1	0	814	814	814	814	0	0
	NanoDST	06.04.2015 21:01:54	07.04 08:58	Data	1	.3	1644	61977	37.7	1.4	5.1	0	0	0	1644	1644	1644	1644	0	0
239872	A	06.04.2015 19:55:51	06.04 20:45	Data	1	.52	127	11336	89.26	3.9	47.9	.2	.2	0	127	127	127	127	0	0
	ExpressCosmics	06.04.2015 19:55:52	06.04 20:44	Data	1	.39	127	8348	65.73	2.9	48.4	.1	.1	0	127	127	127	127	0	0
	NanoDST	06.04.2015 19:55:52	06.04 20:45	Data	1	.03	127	11336	89.26	3.9	3.2	0	0	0	127	127	127	127	0	0

Key: blue: active/differs from preceeding file count brownish: count differs for Tier0 steps magenta: suspicious red: probable error Help!

Last Runs

Run	Start Time	Last Update	Label	#SM	!Files	Size [GB]	Files	Evnts	Disk MB/s	Tier0 MB/s	Open	Close	Inject	Transfr	Check	Repack	Delete	CMSSW Version	HLT Key	dEoL
239874	06.04.2015 21:01:54	07.04 08:58	Data	1	(null)	5.78	4102	177728	.14	.14	0	4102	4102	4102	4102	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0
239872	06.04.2015 19:55:51	06.04 20:45	Data	1	(null)	.94	381	31020	.33	.33	0	381	381	381	381	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0
239870	06.04.2015 19:35:25	06.04 19:43	Data	1	(null)	.05	66	1911	.12	.11	0	66	66	66	66	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0
239866	06.04.2015 19:25:17	06.04 19:30	Data	1	(null)	.02	39	1038	.08	.08	0	39	39	39	39	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0
239862	06.04.2015 19:12:36	06.04 19:15	Data	1	(null)	9.84	6	1353173	111.99	55.68	0	6	6	6	6	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0
239842	06.04.2015 16:17:11	06.04 18:16	Data	1	(null)	2.52	924	83352	.36	.36	0	924	924	924	924	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0
239821	06.04.2015 13:54:45	06.04 16:12	Data	1	(null)	4.03	1000	92809	.5	.5	0	1000	1000	1000	1000	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0
239806	06.04.2015 12:48:47	06.04 13:50	Data	1	(null)	20.01	400	884525	5.54	5.52	0	400	400	400	400	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0
239797	06.04.2015 10:59:35	06.04 12:47	Data	1	(null)	.48	758	17478	.08	.08	0	758	758	758	758	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0
239795	06.04.2015 10:09:20	06.04 10:54	Data	1	(null)	.48	351	23086	.18	.18	0	351	351	351	351	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0
239791	06.04.2015 09:33:09	06.04 10:02	Data	1	(null)	.44	222	21788	.26	.26	0	222	222	222	222	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0
239786	06.04.2015 09:12:12	06.04 09:19	Data	1	(null)	.14	54	6660	.37	.35	0	54	54	54	54	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0
239785	05.04.2015 15:53:33	06.04 09:04	Data	1	(null)	170.68	7161	6878508	2.83	2.83	0	7161	7161	7160	7160	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0
239784	05.04.2015 15:32:08	05.04 15:49	Data	1	(null)	.02	95	187	.02	.02	0	95	95	95	95	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0
239782	05.04.2015 15:19:36	05.04 15:25	Data	1	(null)	0	16	32	.01	.01	0	16	16	16	16	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0
239779	05.04.2015 14:31:54	05.04 15:06	Data	1	(null)	.03	192	433	.02	.02	0	192	192	192	192	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0
239778	05.04.2015 14:15:40	05.04 14:24	Data	1	(null)	.03	49	1761	.07	.07	0	49	49	49	49	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0
239771	05.04.2015 12:49:13	05.04 13:32	Data	1	(null)	.08	248	2200	.03	.03	0	248	248	248	248	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0
239767	05.04.2015 12:10:51	05.04 13:17	Data	1	(null)	.69	1	404346	707.11	.18	0	1	1	1	1	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0
239764	05.04.2015 11:58:38	05.04 13:17	Data	1	(null)	8.33	9	176894	284.31	1.8	0	9	9	9	9	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0
239762	05.04.2015 11:11:01	05.04 13:16	Data	1	(null)	.03	119	845	.02	0	0	119	119	119	119	0	0	CMSSW_7_3_2_patch2	/cdaq/special/Splash/2015/v2.0/HLT/V1	0

Questions?