

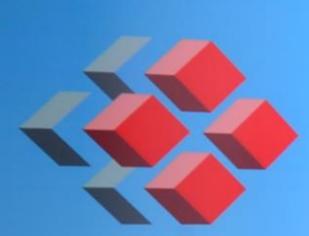


Real-time Mission Critical Supercomputing with Cray Systems

Napa Valley, May 19 2013 Jason Temple and Luc Corbeil, CSCS

ntroduction H

© CSCS 2013



MeteoSwiss' Context

- HPC Services
- Client definition
- Client needs

Design Considerations

- Partitions
- Scheduler
- Filesystems
- Network



CSCS

- Sonexion
- Lustre issues
 - Lustre Support
 - Silent Data
 Corruption
- Implications





HPC Services for MeteoSwiss

• Maintenance of a 24/7 mission critical infrastructure within a research environmnent

- Leverage existing infrastructure where it makes sense
 - Mid-term storage and archiving
 - User environment (homes, etc.)
- Put in place the required safeguards/failover mechanisms
 - Infrastructure, power and cooling
 - UPS
 - Hardware configuration
 - System configuration
 - Global systems monitoring (Nagios, Ganglia)
 - 24/7 on-call support (Pichetto and external inf. company)
- Close collaboration between both organizations at all levels
 - Management
 - Operations





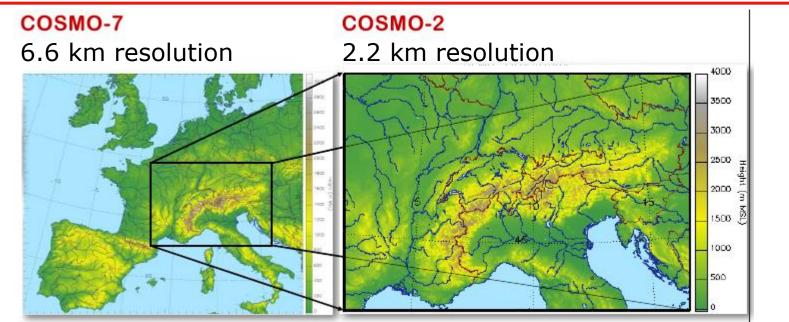


- "MeteoSwiss is the national weather and climate service for the Swiss public, for government, industry and science. With our public service we ensure the basic supply of weather and climate information in Switzerland."
- In addition, must provide on-demand monitoring for the Nuclear Regulatory Agency in the event of a nuclear incident somewhere in the world.





MeteoSwiss – some details



COSMO 2km: 8 times/day, within 25 minutes COSMO 7km: 3 times/day, within 25 minutes

Correct results required to issue weather warnings in a timely manner

Must be right the first time (unlike running linpack until success)



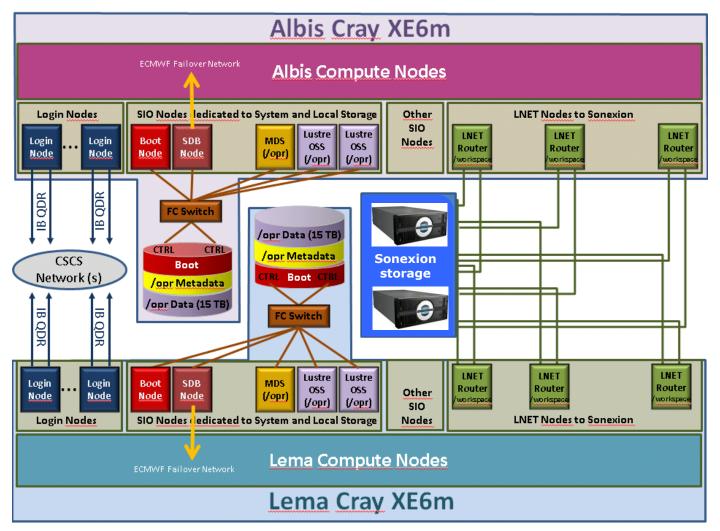


Two Distinct Partitions/One System ECOphies ECC. Albis Lema Cray XE6 11111 Cray XE6 Cray XE6 ///// 11111 5 Login Nodes 5 Login Nodes 18 compute blades 42 compute blades -24 cores each -24 cores each 72 nodes 168 nodes 144 sockets 336 sockets 1728 cores 4032 cores 16.5 TFLOPS 33.5 TFLOPS Two distinct systems (Albis/Lema)





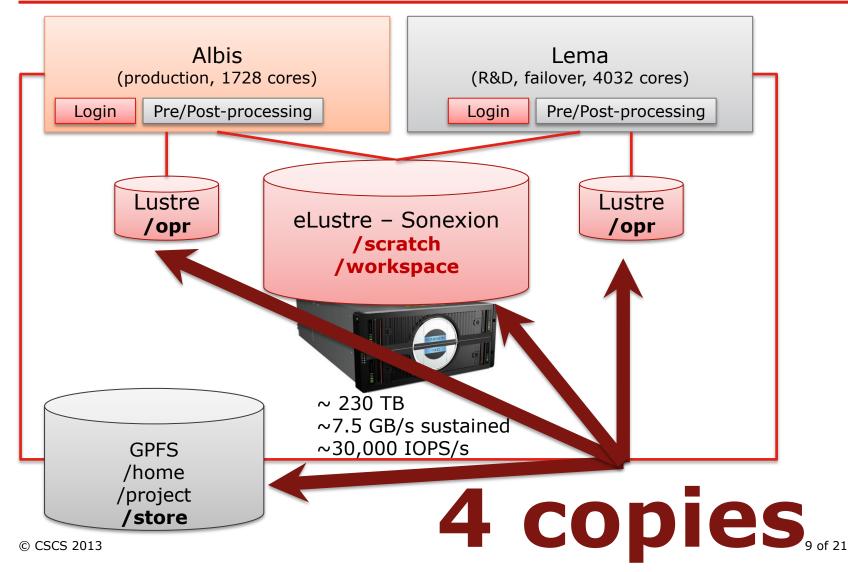
Albis and Lema Complex Configuration







Albis and Lema Filesystem Configuration





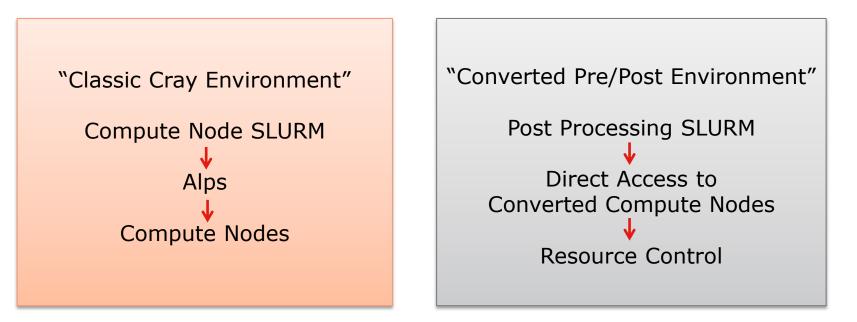


SLURM for Compute and Pre/Post Proc Scheduling

SLURM

- «Simple Linux Utility for Resource Management»
 - Open Source from LLNL
 - Free
 - Very configurable, extensible









Problems for Mission-Critical Supercomputing







Problems with the Sonexion 1300

- Management GUI is not very useful in our version
- Incorrect installation
 - Failover patch not on all servers
- Substandard switch hardware installed (unmanaged switches)
 - Occasionally froze, needed rebooting
- Difficult to administer
 - No external ports
 - Puppet/certificate setup non-trivial
 - No «reliable» performance metrics
 - Basically a black box
- Apparent communication problems between Cray and Xyratex
- No «smooth» upgrade path between 1.0 -> 1.2.1
- Silent Data Corruption (not isolated to Sonexion, Lustre in general) © CSCS 2013





Problems with the Sonexion 1300

Hostname 📤	Node Type	Power State	Mounted (1)	Targets (17)	HA Partner
r sonex00	MGS	💧 On	0	0	sonex01
L sonex01	MDS	Unknown	1	1	sonex00
r sonex02	OSS	🍐 On, Offline	0	4	sonex03
L sonex03	OSS	🍈 On, Offline	0	4	sonex02
r sonex04	OSS	🍈 On, Offline	0	4	sonex05
L sonex05	OSS	🍐 On, Offline	0	4	sonex04





Problems with the Sonexion 1300

Node Control Performance Log Browser Support Terminal Dashboard Health Configure	re
--	----

File Configure

fsl

	fs1-MDT0	000 201	13-04-22	2 15: 20:1	5.0	1		OST 20	13-04-22 15	:27:50.	0			OSS	2013-04-22	15:27:50	0
%CPU	c.	%KB		%Inod	es	0st Name	Read Rate	Write Rate	%CPU	% KB	%Inodes	Oss Name	Read Rate	Write Rate	%CPU	%Space Used	%Inodes Used
****		*o*o*o*		*ototek		fs1-0ST0000	жоюнок	жнонок	(WORKORK)	63.67	1.30	sonex02	жиснок	жононок	1.56	63.69	1.30
Operation	Samples	Sample	Avg	Std Dev	Units	fs1-0ST0001	жоюж	жнонок	WORKORK	жононок	жононок	sonex03	жиник	жононок	жононок	NOROROW	*CHORCE:
		/Sec	Value			fs1-0ST0002	жоюж	жнонок	1.56	63.67	1.30	sonex04	жиник	жононок	жононок	NOROROW	*CHORCE:
						fs1-0ST0003	жоюж	жнонок	1.56	63.69	1.30	sonex05	жнонок	жононок	жононок	NOROROW	*CHORCE:
						fs1-0ST0004	жоюж	жнонок	WORKING CONTRACTOR	жононок	жононок	AGGREGATE	0.00	0.00	жонононск	NOROROW	жонононк
						fs1-0ST0005	жжение	жнонок	300000K	жононок	жиник	MAXIMUM	жононок	жоюж	1.56	63.69	1.30
						fs1-0ST0006	жженик	жнонок	жиновок	жононок	жононок	MINIMUM	жыснок	жоюж	1.56	63.69	1.30
						fs1-0ST0007	жженик	жнонок	жиновок	жононок	жононок	AVERAGE	0.00	0.00	0.39	15.92	0.33
						fs1-0ST0008	жжение	жнонок	300000K	жононок	жиник						
						fs1-0ST0009	жжение	жнонок	300000K	жононок	жиние						
						fs1-OST000a	жжение	жнонок	300000K	жононок	жиник						
						fs1-OSTOOOb	жоюж	жысыск	WORKORK	жононок	жононок						
						fs1-OST000c	жононок	жнонок	WORKORK	жононок	жононок						
						fs1-OST000d	жононок	жнонок	WORKORK	жононок	жононок						
						fs1-0ST000e	жононок	жысыск	WORKORK	жононок	жононок						
						fs1-OST000f	жнонок	жнонок	XOROROK	жононок	жононок						
						AGGREGATE	0.00	0.00	жисновок	жонононск	жиновок						
						MAXIMUM	жжжж	жнонок	1.56	63.69	1.30						
						MINIMUM	жжжж	жнонок	1.56	63.67	1.30						
						AVERAGE	0.00	0.00	0.20	11.94	0.24						



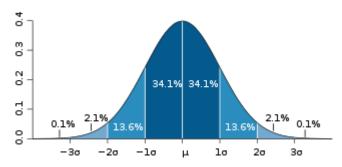


Reliability of Tools and Scientific Computing



How can you trust your scientific results if the tools you use are not 100% reliable?

Do you make many runs, then choose a 95% confidence level from the normal distribution?





Silent Data Corruption

- After going into production, MeteoSwiss started to experience data corruption
- Absolutely silent in the Lustre logs, at any level
- 3 different types of corruption
 - Zero size files resulting from a simple untar of text files
 - Corrupted data in the middle of files, either zeroes or random
 - Truncated files.
- Random occurrences in random types of files
- Caused MeteoSwiss to send corrupted product output files to their clients
- Problem lasted more than 10 months
 - First reported in June '12
 - Cray got involved in August
 - Still corruption in February '13



0xbc045c5 0xbc045f2 0xbc0461f 0xbc0461f



Examples of Data Corruption

A # M N 9 . 8 . 2 . +

н. 5

>

xbc04598	_SINP_lp	sinputlm	<u>_</u> C					_SINP_	lpsinput	:lm_c.bad					



ETTH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich

Examples of Data Corruption

		SINP lpsinputlm c	SINP lpsinputlm c.bad
	0xbc04598	000000000000000000000000000000000000000	000000000000000000000000000000000000000
	0xbc045c5	00000000000000000000000000000000000000	000000000000000000000000000000000000000
	0xbc046a6		
	0xbc046d3		
	0xbc04700	000000000000000000000000000000000000000	000000000000000000000000000000000000
	0xbc0472d	0000000000000001f00ff01ff0207027a02ff0267025a027e0364045305ff072207ff08220808081b086508	000000000000000000000000000000000000000
Carcelary i Carcelary i		ff097f0aff0bff0d430cff0bff09ff072403ff00ff000000000000000000000000000000	
000:003:001f05f111[29:31136f145[27f128:247] 11f119f143:23916c1691[27128] 000:00:000:000:000:000:000:000:000:000	0xbc04895		
0xbc1d12 5xbc131 0xbc1412 0xbc131 0xbc1412 0xbc131 0xbc331	0xbc048c2	00000000000000000000000000000000000000	000000000000000000000000000000000000000
0xbc1d429 0e 0e <	0xbc048ef	000000000401ff05ff11ff29083aff3eff4a5f27ff286c24ff1fff19ff435039ff6cff65ff623f613219ff1d	000000000000000000000000000000000000000
0xbc1d429 0e 0e <			
Chi Coldario On			
0x1c1433 0x10			
0xbcdadd			
0xbcdddz 0xbcddz			
00:00:00:00:00:00:00:00:00:00:00:00:00:			
0xbccda37 0:00:00:700 f 011:01:01:01:01:01:01:01:01:01:01:01:00:00			
Oble CdaB4 11d=:11i11111:11:01:00:00:00:00:00:00:00:00:00:0			
Oxbccdabi f11c=0:18 f134.dcff08 f104 f104=114:13 f15=c14 f131 ede=3095702 f100:100:00:00:00:000:000:000:000:000:00			
cdbcddadb cdbccdbdb cdbcdbdb cdbcdbdb cdbcdbdb cdbcdbdb cdbcdbdb cdbcdbdb cdbcdbdb <	0xbc04a84	ff1d5c1bff19ff164e12150cff066801ff000000000004002d007b00130019003301ff05490620064109ff15	
Outcadabb C00 C	0xbc04ab1	ff1c691e5d18ff134a0cff08ff0aff0e43114913ff155614ff131e0e64093f02ff0012000000000000000000000000000000	
cdbcdb3g cdbcdb3g cdbcdb4g cdbcdb4g <td< td=""><td>0xbc04ade</td><td><pre>000000000000000000000000000000000000</pre></td><td>000000000000000000000000000000000000000</td></td<>	0xbc04ade	<pre>000000000000000000000000000000000000</pre>	000000000000000000000000000000000000000
0xbc2db55 0x00:00:00:00:00:00:00:00:00:00:00:00:00	0xbc04b0b	00000000000000000000000000000000000000	000000000000000000000000000000000000000
0xbc2db55 0x00:00:00:00:00:00:00:00:00:00:00:00:00	0xbc04b38	00000000000000000000000000000000000000	000000000000000000000000000000000000000
0xbc04b2 0xbc04b2 <td< td=""><td></td><td></td><td></td></td<>			
0xbc0bbf 0xbc0bf			
0xbc0ds_00 0xbc0ds_00 <td></td> <td></td> <td></td>			
bxbcdrd1 ffd4ffd1l81a4c287:34ff4d0375tff1a519f17ff3c47ab724ff7d6ff6ffbff294728439ff274z26ff27 c000c000:000:000:000:000:000:000:000:00			
0xbc04c46 5323f253d224lb45ldf125f35f3b54542ff31bd3bd2e5e00c000000000000000000000000000000000			
0xbcd4z3 0x00x00x00x00x00x00x00x00x00x00x00x00x0			
xbc04ca x000000000000000000000000000000000000			
0xbc04cd 0x00000000000000000000000000000000000			
0xbc04cfa 0x00000000000000000000000000000000000			
0xbc04d27 0x00000000000000000000000000000000000	0xbc04ccd		
0xbc04d4d 0xbc00c00c00c00c00c00c00c00c00c00c00c00c00	0xbc04cfa	000000000000000000000000000000000000000	
0xbc04d4d 0xbc00c00c00c00c00c00c00c00c00c00c00c00c00	0xbc04d27	000000000000000000000000000000000000000	00000000000000000000000000000000000000
0xbc04d81 33005900ff00ff011201ff03ff064f0a5101c16141d0723ff2a6e3012352a37ff38ff376433ff2ff2b 000000000000000000000000000000000000	0xbc04d54	○○00○○00○○00○○00○○00○○00○○00○○00○○00○○	00000000000000000000000000000000000000
0xbc04dae ff26ff21ff1aff11ff095e022c0026005e00ff014300ff00ff01430ff08ff08ff08ff08ff08ff08ff08ff08ff08ff			
0xbc04db 7214170dff0b170dff11361474170618ff18ff17ff136a0e6b06ff01ff000000000000000000000000000000			
0xbc04e8 0x00000000000000000000000000000000000			
0xbc04e35 0x0000000c00c00c00c00c00c0000000000000			
0xbc04e62 0xbc04e62 <t< td=""><td></td><td></td><td></td></t<>			
0xbc04esf 000000000000000000000000000000000000			
0xbc04ebc 0xbc00000000000000000000000000000000000			
0xbc04ee9 0xbc04ee9 0xbc00000000000000000000000000000000000			
0xbc04f16 000000000000000000000000000000000000	0xbc04ebc		
0xbc04f43 ff1f042bff3856430550ff51ff51ff51ff51ff51ff51ff61482b7d497c447b093e2f40583a2c38ff3906414b3bff3cff2c 0000000000000000000000000000000000	0xbc04ee9		
	0xbc04f16	00000000000000000000000000000000000000	
	0xbc04f43	ff 1f 042bff3856430550ff51ff51ff51ff51ff51ff482b7d497c447b093e2f40583a2c38ff3906414b3bff3cff2c	000000000000000000000000000000000000000
	0xbc04f70	ff2cff26622bff3460406040ff3e363fff52ffff413fff44ff39ff ⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰⁰	000000000000000000000000000000000000000





Examples of Data Corruption

	backup.lfff0006000p	lfff00060000p.corrupt
0x2fffc9d	600d64ff68ff6c0a6e656fff6ff6c6965ff5c3b531c4cff49ff496a4aff4c374dff4f4d50ff51ff53ff55ff	600d64ff68ff6c0a6e656fff6fff6c6965ff5c3b531c4cff49ff496a4aff4c374dff4f4d50ff51ff53ff55ff
0x2fffcca	59105cff612165ff6aff707475ff7b037fffffffffffffffffffffffffffffffff	591 0 5cff612165ff6aff70 74 75ff7b 03 7ffffffffffffffffffffffffffffffffffff
0x2fffcf7	[[ff[[ff][25][ff]]1a][ff][34][ff][ff][0e][43][54][1b][ff][62][7b][6c][ff][ff][ff][ff][ff]	ffffffff[25ffffff1affffff34ffffffff0eff43f54ff1bfffff62ff7bf6cffffffffffffffffffffffff
0x2fffd24	f 08 04 ff 50 19 ff 2f ff 6d ff ff ff ff 7f 71 71	[08] 04] [f] 50] [9] [f] 2[] [f] 6d] [f] [f] [f] [f] 7] [7] [7] [1] [5b] [f] 44] [f] 43
0x2fffd51	666 [51] 62 [51] 62 [51] 64 [51] 65 [6] 65 [66 51 62 ff 1d 0e 27 ff
0x2fffd7e		
		151 ff ff ff 4b ff ff 22 16 6b 10 43 77 77 ff ff ff ff ff ff ff ff ff
0x2fffdab	f 7b f ff f 2e f 4f ff ff 12f 57 ff ff ff ff ff ff ff 4d ff ff ff ff ff ff 68 ff ff 26	17b1ff12e14f1ff112257fff11ff1ff1ff14ff14d1ff14ff14ff14ff14f14681ff126
0x2fffdd8	ffleffffffffffffffffffffffffffffffffff	ffleffffffff79ff28ffffff31ff3afffffffff5eff00fflcff74ffldffffffffffffffffffffffffffffffff
0x2fffe05	ffffff21ffffff41ff5cff55fffffffffffffffffffffff	ffffff21fffff41ff5cff55ffffffffffffffffff1dff51ff7bff08ff5aff70ffffff7cffffff13ffffff26ff33
0x2fffe32	ffffffffffff15ff15ff39fffffi35ff2bffffffffffffff4fffffff27ff54ff51ff29ff07ff2bffffffff	ffffffffff15ff15ff139ffffff35ff2bffffffffffffffffffffffffffff
0x2fffe5f	ff6bffffff24ff32ff59ffffff7bff71fffffffffffffffffffffffffff	ff6bfiffff24ff32ff59ffffff7bff71fffffffffffffffffffffffffff
0x2fffe8c	ffffff61ff29ffffffffffffffffffffffffffff	ffffff61ff29fffffffffffff0eff71ffffff70ff3eff19ffffff1cff47ff60ff62ff49ff3bff66fffffffffffff
0x2fffeb9	fffffff4cff0fffffffffffffffffffffffffff	fffffff4cff0fffffffffffffffffffffffffff
0x2fffee6	fffffffffffffffffaf7b0d76ff73ff7264726572ff72ff70ff6dff6a1466ff645562ff61ff5f6f5aff53 37 4a44	ffffffffffffffffafb0d76ff73ff7264726572ff72ff70ff6dff6a1466ff645562ff61ff5f6f5aff53 37 4a44
0x2ffff13	42ff3eff3e1e3eff3eff3dff3b 43 38ff36ff35ff361 5 37 0b 37ff38 78 38ff38 0a 36ff33ff30ff2e 51 2dff2f 59	42ff3eff3e1e3eff3eff3dff3b4338ff36ff35ff3615370b37ff387838ff380a36ff33ff30ff2e512dff2f59
0x2ffff40	31ff34ff37 7c 3aff3e 12 41ff451a48ff4dff531a59 7b 60ff67ff6e 5a 73 78 76 6a 77 46 76ff75ff74ff73ff	31ff34ff37 7c 3aff3e1241ff451a48ff4dff531a597b60ff67ff6e5a73 78 766a774676ff75ff74ff73ff73ff
0x2ffff6d	742b753c764f770d77ff77ff78ff79ff7a4d7aff7a14790f77ff75ff73ff71746f5c6d706bff6aff69ff69ff	742b753c764f770d77ff77ff78ff79ff7a4d7aff7a14790f77ff75ff73ff71746f5c6d706bff6aff69ff69ff
0x2ffff9a	694268ff68ff67ff66ff65ff64ff63ff62ff61ff60ff5fff5f035e715dff5dff5dff5dff5e085e655eff5eff	694268ff68ff67ff66ff65ff64ff63ff62ff61ff60ff5fff5f035e715dff5dff5dff5dff5e085e655eff5eff
0x2ffffc7	5e455dff5cff5bff5b0f5a0358ff57ff56665555545b5360524b511a4fff4eff4dff4dfc4cff4c544bff4aff	5e455dff5cff5bff5b0f5a0358ff57ff5666555545b5360524b511a4fff4eff4dff4d4c4cff4c544bff4aff
0x2fffff4		492f47ff45ff43ff41ff3fff3dff3bff394036ff32ff32ff316430992eff2d000000000000000000000000000000
	492147114511451145114111511150115910501154105011541152115104500920112077205020112a5029112952 28ff2814267123ff21021e651c561aff19ff19ff19f2195619365f2b63ff68686c5e6f5e714d724e72ff7251	
0x3000021		0.0000
0x300004e	71ff70ff6fff6eff6e026d0c6bff6a58691b681f672b65ff63ff61715eff5b4857ff54014f7849ff41ff39ff	
0x300007b	33302fff2f4430ff332035ff38ff3b6e3dff3f344077416341ff41ff41ff41ff41ff425b431e43ff445f442542ff	······································
0x30000a8	40ff3eff3d333cff3dff3fff42ff460e49434bff4eff51ff550e59115d6461ff66526a0a6cff6eff6fff6e14	000000000000000000000000000000000000000
0x30000d5	68ff605156ff4eff4a6b497f4aff4c734e765054520953ff561459095cff60ff65ff6a6f6f71744b78ff7d04	000000000000000000000000000000000000000
0x3000102	ff1aff3afffffffff58ffffff58ffffffffffffffff	000000000000000000000000000000000000000
0x300012†	ff3effffffffffffffffffffffffffffffffff	
0x300015c	ff4dffffffffffff6bff71ffffff72ffffffff58ff57ff37ff32ffffff3ffffffff6bff07ff16ffffff53	<pre>000000000000000000000000000000000000</pre>
0x3000189	ff2affffffffff5eff16ffffffffffffffffffffffff	000000000000000000000000000000000000000
0x30001b6	ffffff7cffffffffffffffffffffffffffffff	000000000000000000000000000000000000000
0x30001e3	ff08ffffff09ffffff6ffffffff49fffffffffffffff	000000000000000000000000000000000000000
0x3000210	ff7affffff11ffffff37ff0cff09ff69ffffff64ff1dff76ff1ffff1efffffffffffffffffffffffffff	000000000000000000000000000000000000000
0x300023d	fffffffffff64ffffff3efffffffffff53ff79ff5eff38ffffff08ff02ff6aff65ffffff6fffffff65ff29ff25	000000000000000000000000000000000000000
0x300026a	fffffff48fffffff4fffffffffffffffffffff	000000000000000000000000000000000000000
0x3000297	ffffffffffffffffffffffffffffffffffffff	000000000000000000000000000000000000000
0x30002c4	ff7aff4fff38ffffff4bff6ffffffffffffffffffffffffffff	000000000000000000000000000000000000000
0x30002f1	ff78ff3affffffffffffffffff62ffffff27ff0aff12ff51ffffffffffffffffffffff0a7f717bff784c751d72ff	000000000000000000000000000000000000000
0x300031e	71ff71 72 70ff6eff6c 00 68ff65ff64 3 a63ff63ff61ff5c 5a 54 4 e4bff44ff40ff3fff3fff3fff3e6c3c28391e	000000000000000000000000000000000000000
0x300034b	3628343a33ff34ff35ff36173602354b33ff31ff2fff2e082dff2f03314233ff36ff39ff3cff405043ff477e	000000000000000000000000000000000000000
0x3000378	4bff5off56ff5dff6dff6bff71ff757b771f7713761e74ff73ff733973ff747e75ff767f771777ff787a796f	000000000000000000000000000000000000000
0x30003a5	722d725479ff78ff76ff750f72ff70ff6eff6cff6b5e6a3c697668ff68ff684467ff66ff65ff64ff63ff62ff	
0x30003d2	/a2//a3//stores/arts/f5/df5/df5/df5/df5/df5/df5/df5/df5/df5/	
0x30003ff	5ff54ff53ff52545125450014f014e3f4dff4dfe4cff4c214b1049ff47ff460a440741ff3fff3e1b3c393a2d	
0x300042c	37ff35ff33 78 31ff30 69 2f 57 2e 47 2d 07 2bff2a 35 28ff27ff27 00 266 a 26 1 425ff2456224c1fff1dff1bff1aff	
0x3000420	la 75 1a 4 c1a 37 1a 0e 19ff5e 70 63ff68ff6cff6fff71ff72ff73ff73ff736c72ff71ff70ff6fff6eff6d 76 6c 6 1	
0x3000486	6b796aff695d6763647760ff5d415a43580c5617536f4f0c4879406a38ff331a30ff315033ff37323a5e3cff	
0x30004b3	3e6b3f70403240ff416341ff41ff4256430643ff444044d43ff41ff3f693cff3b6a3b233c043dff40ff4424 Azff4x50xxffxff56552556b5x655465ff62ff62ff62ff62ff62ff62ff62ff62ff62ff	
0x30004e0	47ff4a 59 4cff4f 56 52 62 56 0b 5a 46 5eff63ff67ff6b 01 6d 4d 6eff6eff6bff647 c 5aff51ff4bff49ff4aff4cff	
0x300050d	4f29517e53ff561759065cff60ff65666a696fff74ff78ff78ff7c3e7ffff5cff46ffffffff74ffffff0effff	
0x300053a	30 ff ff ff 42 ff 63 ff 35 ff ff ff ff ff 23 ff 15 23 23 5a ff ff ff ff	······································
0x3000567	ffffffffffffffffffffffffffffffffffffff	······································
0x3000594	ff6dff33ffffff23ff7bffffffffffffffffffffffff	000000000000000000000000000000000000000
0x30005cl	75 ff 19 ff 58 ff ff 1f ff ff ff ff	©©000000000000000000000000000000000000
0x30005ee	ff50ff0bff4eff41fffffffffffffffffffffffffffffff	000000000000000000000000000000000000000
0x300061b	ff13fffffffffffffffff07ff19fffffffffffffffff	000000000000000000000000000000000000000
0x3000648	ffffffffffffffffffffffffffffffffffffff	000000000000000000000000000000000000000
0x3000675	ff2dff6aff2fffffffffffff44ff3cfffffffffff77ffffffffff	000000000000000000000000000000000000000





Difficulties Capturing the Problem

- The fact that it was silent and random made it almost impossible to troubleshoot.
- Was not easily reproduceable, therefore, not easy to capture.
- CSCS managed to reproduce the zero-size file by untar issue one time after over 50,000 attempts, but nothing was seen in the logs
- MeteoSwiss was forced to fsync() almost every write operation in an attempt to flush the cache

– No discernable effect, other than slowing down I/O

- Most Vexing: Happened on internal Lustre, as well as on the Sonexion!!!
 Lustre versions 1.8.x and 2.0 (Sonexion 1300)
- Despite CSCS' & Cray's efforts, no serious progress on the case
- What next?





Using the Lustre Mailing Lists

- At our wits' end, a question describing out data corruption issue was sent to the mailing lists.
 - Almost immediately, we received a response from another Cray user that had experienced almost the exact same problems, with links to lustre bug reports
- This email coincides with sudden renewed interest on the part of Cray
- Weekly con-calls were implemented in order to corner the problem



Finally a Solution

- After more than 10 months of silent data corruption, Cray fast-tracked some more-than-year-old Lustre patches:
 - (from the patch readme files)
 - Handle network errors during bulk I/O.
 - Lookup returns wrong inode following rename by another client
 - Modify LND message send/recv rx timeout policy
- As of today, more than 2 months later, there have been no further incidents of corruption



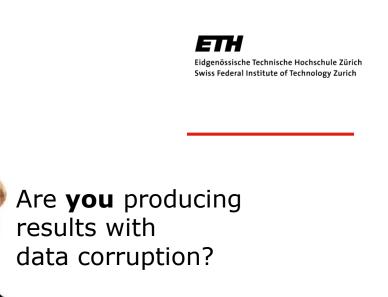


Implications

- When a company freezes or forks Lustre, it **freezes** it
 - «Slow» access to recent bug fixes
 - «Even slower» access to recent developements
 - E.g. HA failover
 - Our bugs
- Public mailing lists: to post or not to post?
 - CSCS primary duty is to protect MeteoSwiss operations
- Other centers around the world can be impacted by these problems
 - How many systems are sold to this day without these patches?
 - How are customers supposed to know?
- Lustre is always advertised as **scratch space**
 - Implying «don't trust it, it can be lost at any time», but it must still provide data integrity – «fast vs reliable»







© CSCS 2013





And Real-Time Mission Critical Supercomputing?

- Significant and respected scientific results are produced using Lustre
- For Real-Time operations, it must work the first time. All the time.
- No parallel filesystem is 100% reliable
 - But supportability is key, so issues are quickly addressed
 - The breach of trust occurs once the first byte of data is lost
- Sites must be made aware of major filesystem issues and be given the opportunity to mitigate
 - And reformatting the filesystem is not a viable upgrade path





How can we make this better?

- For CSCS:
 - Acceptance
 - Run the entire suite (not IOR)
 - Work with Cray to standardize bug reporting
 - Consider lobbying within OpenSFS to prioritize supportability

- For Cray:
 - Field Notices for critical issues
 - Admitting knowledge of a bug to clients is not a weakness
 - Consider lobbying within OpenSFS to prioritize supportability
 - Back-porting essential
 - Closer collaboration with Lustre entities (i.e. Xyratex)





Thank you for your attention.