Update on EOS productisation - Comtrade 360's results

Luca Mascetti
Elvin Sindrilaru
CERN IT Storage and Data Management

luca.mascetti@cern.ch
elvin.alin.sindrilaru@cern.ch

Gregor Molan
Comtrade 360's AI Lab

gregor.molan@comtrade.com

COMTRADE 360: CERN openlab Associate member

In 2015 Comtrade 360 joined CERN openlab as associate member

https://openlab.cern/project/eos-productisation

EOS

EOS Architecture

Open-Source Storage designed and developed in CERN IT

Elastic, Adaptable and Scalable for data recording, user analysis and data processing

High-available and low latency namespace
- namespace persisted on a distributed key-value store
- working entries cached in-memory

High available and reliable file storage, based on (cheap) JBODs:
- File replication across independent nodes and disks
- Erasure coding to optimize costs and data durability

MGM : meta data server
MQ : message queue
NS : persistent namespace
FST : file storage server

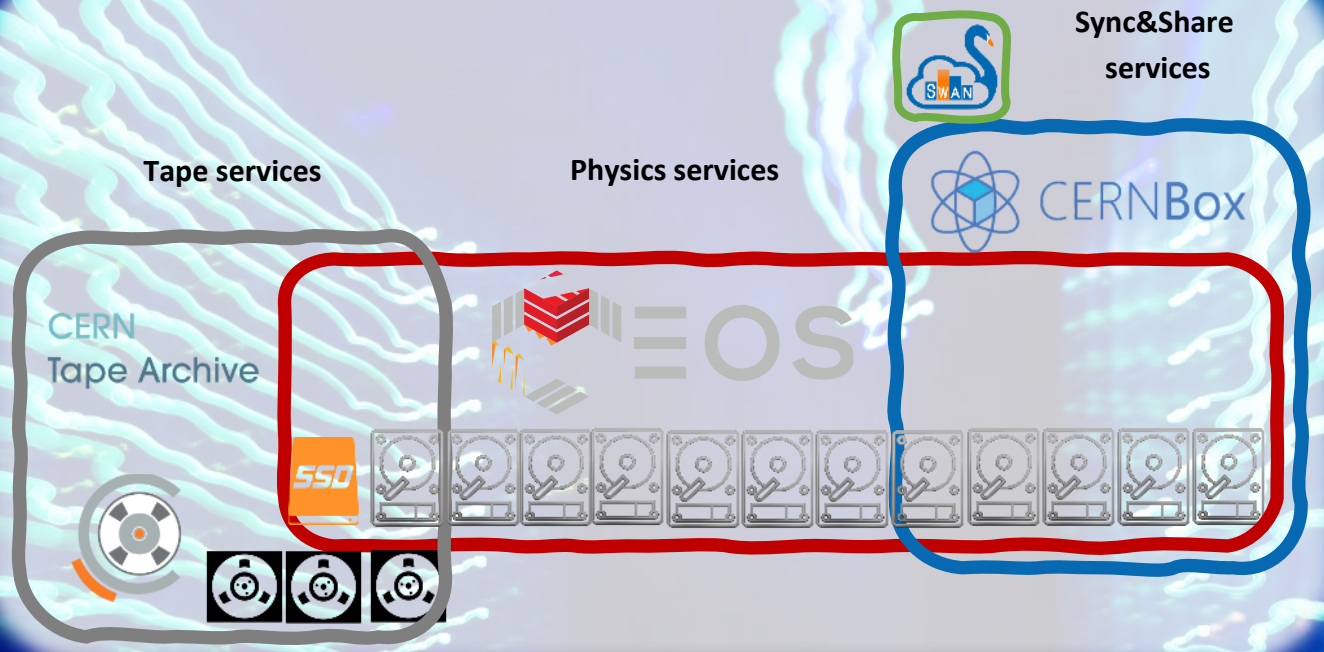
EOS @ CERN

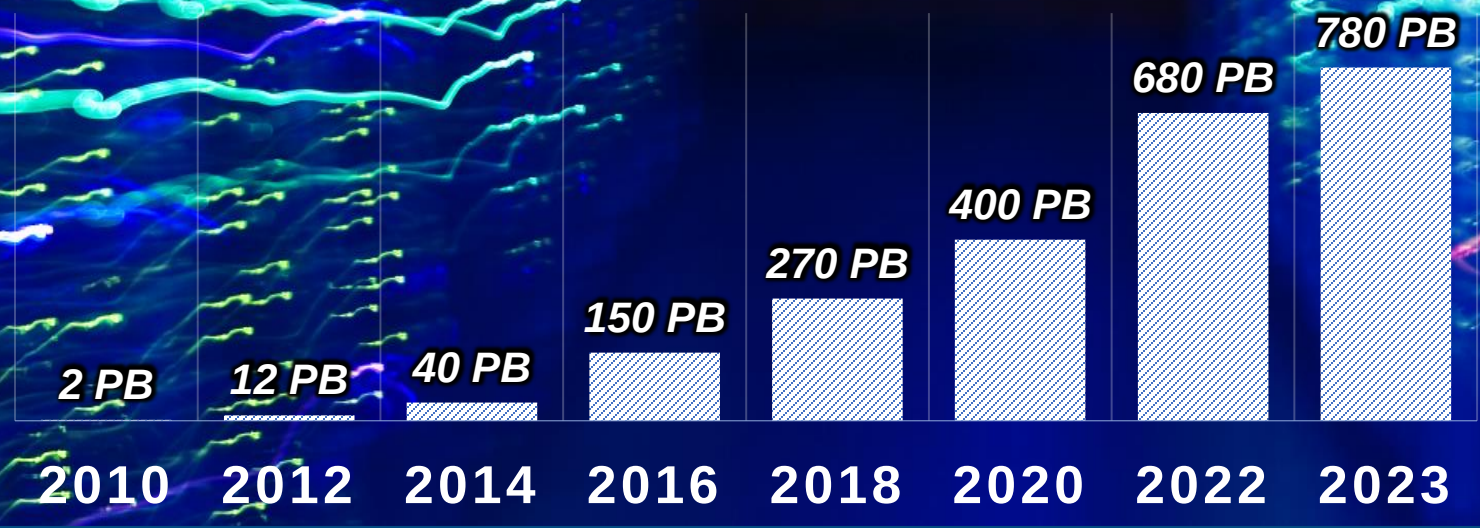
Total Space
780 PB

Files Stored
~8 Bil

 # Storage Nodes
~1300

 # Disks
~60000





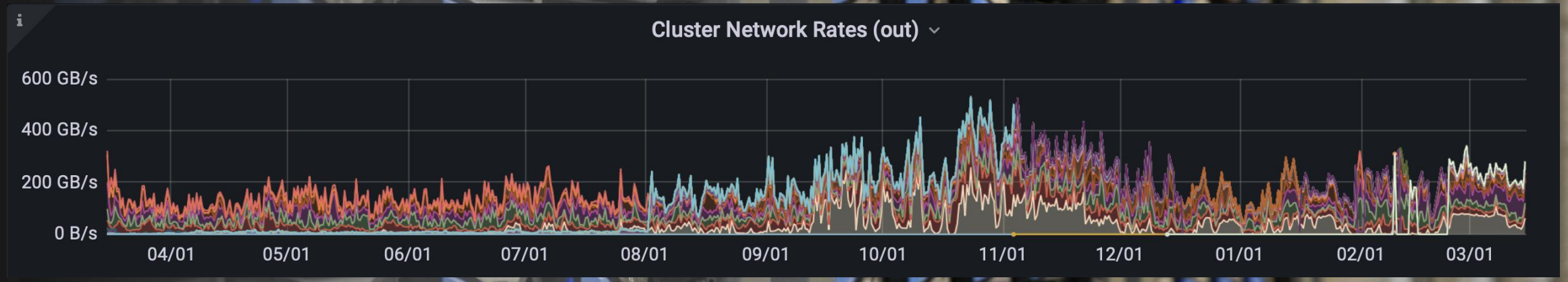
EOS CERN Services

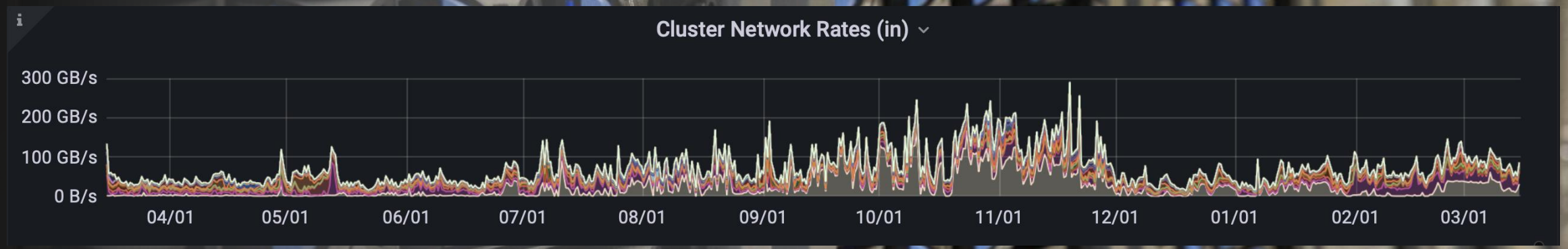
EOS Physics for experiment data

CERNBox for end-user data and sync&share

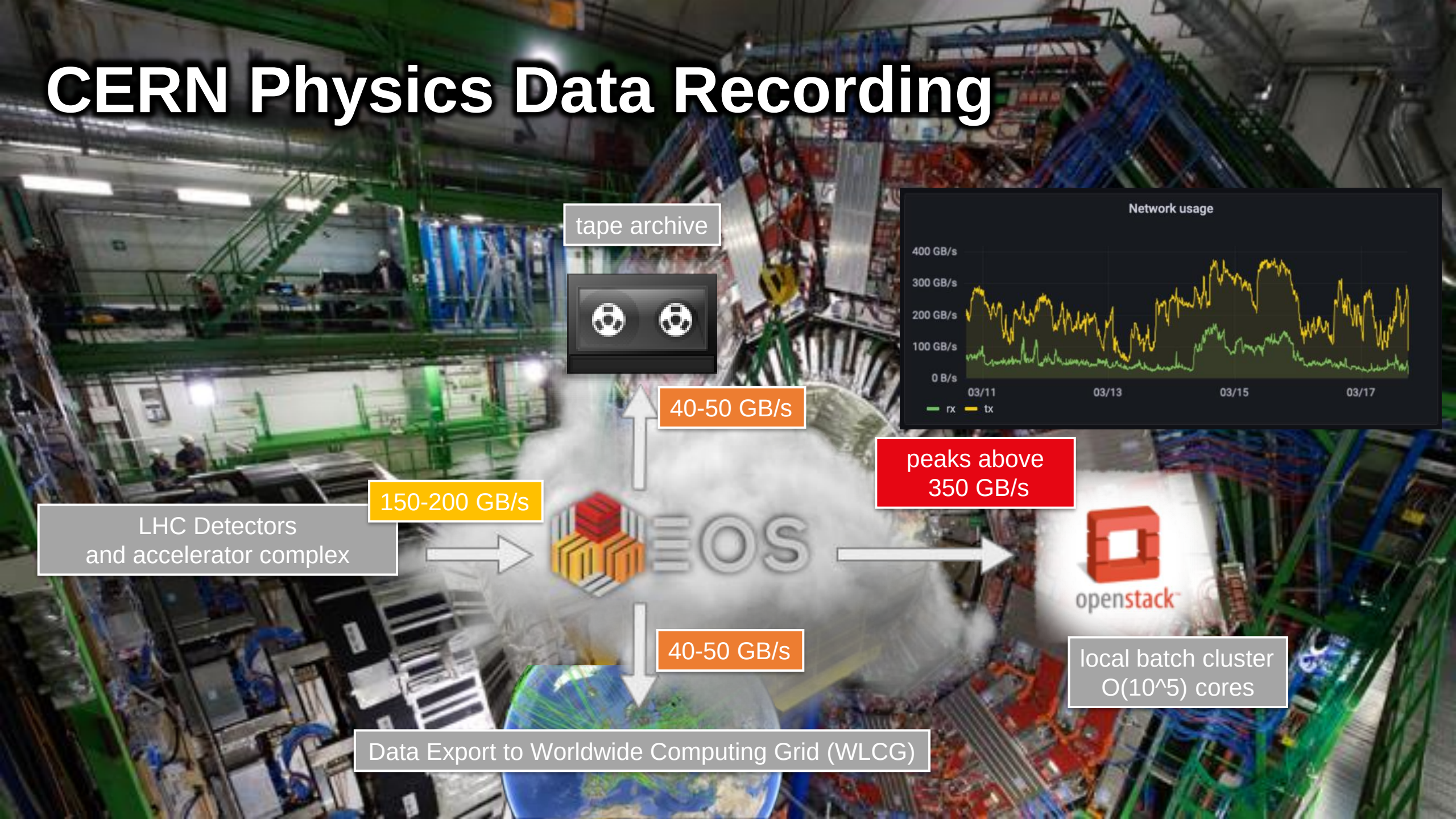
2022 stats

Total amount of files read	Total amount of bytes read	Total amount of files written	Total amount of bytes written
21.8 Bil	4.08 EB	5.73 Bil	631 PB





CERN Physics Data Recording



Comtrade at CERN

A. **History**
B. **Activities**
C. Recent R&D results
   i. EOS Windows Native Client
   ii. EOS-Drive for Windows
   iii. Distributed FS Comparison

Comtrade Group / Comtrade 360

A. Comtrade at CERN - History

2014-05-16

- The meeting with Slovenian scientists at Comtrade (Gregor Molan)

2014-10-01

- The first official meeting at CERN
  Alexis Lope-Bello, Viktor Kovačević, Gregor Molan

2014-12-16

- Started negotiations for the CERN EOS project

2015-10-20

- CERN openlab
  The signed framework agreement
  The signed project agreement

B. Comtrade at CERN - Activities

- The official industry partner for CERN EOS development.
- Physical participation in CERN software development (before Covid)
- Active presentations at CERN since 2019
- Joined presentation at EXPO 2020 in Dubai

- 2023
  - The transformation from “CERN openlab partner” to “CERN partner”.

Comtrade at CERN

A. History
B. Activities
C. Recent R&D results
   i. EOS Windows Native Client
   ii. EOS-Drive for Windows
   iii. Distributed FS Comparison

Comtrade Group / Comtrade 360

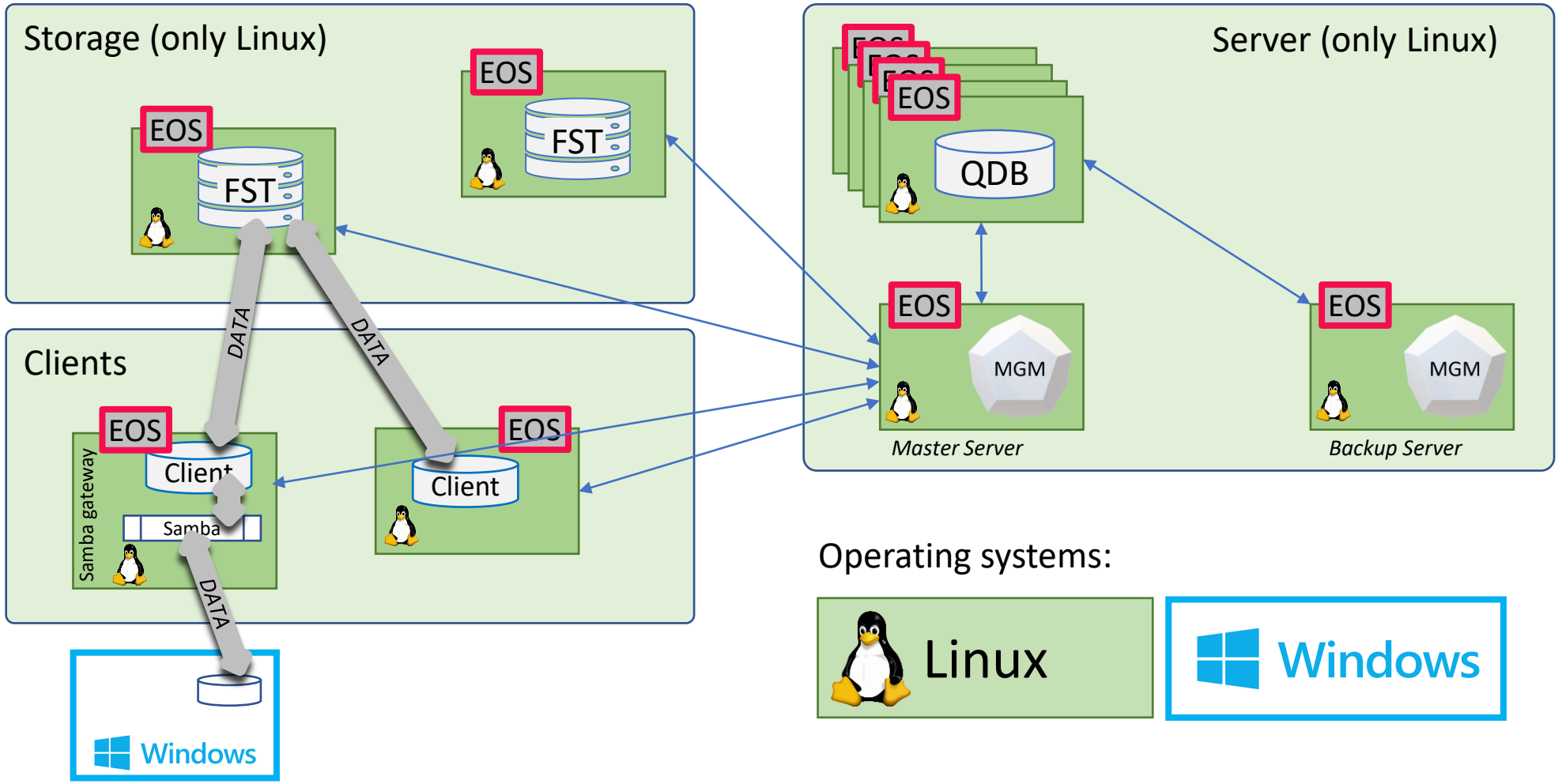
i. EOS CLI on Windows

- Not a port using Windows Subsystem for Linux (WSL)
- Completely new Windows EOS client
- New solution for networking issues on Windows
- New solution for security issues on Windows
- Technologies for EOS-wnc
  - Protocol Buffers
  - gRPC
  - cURL

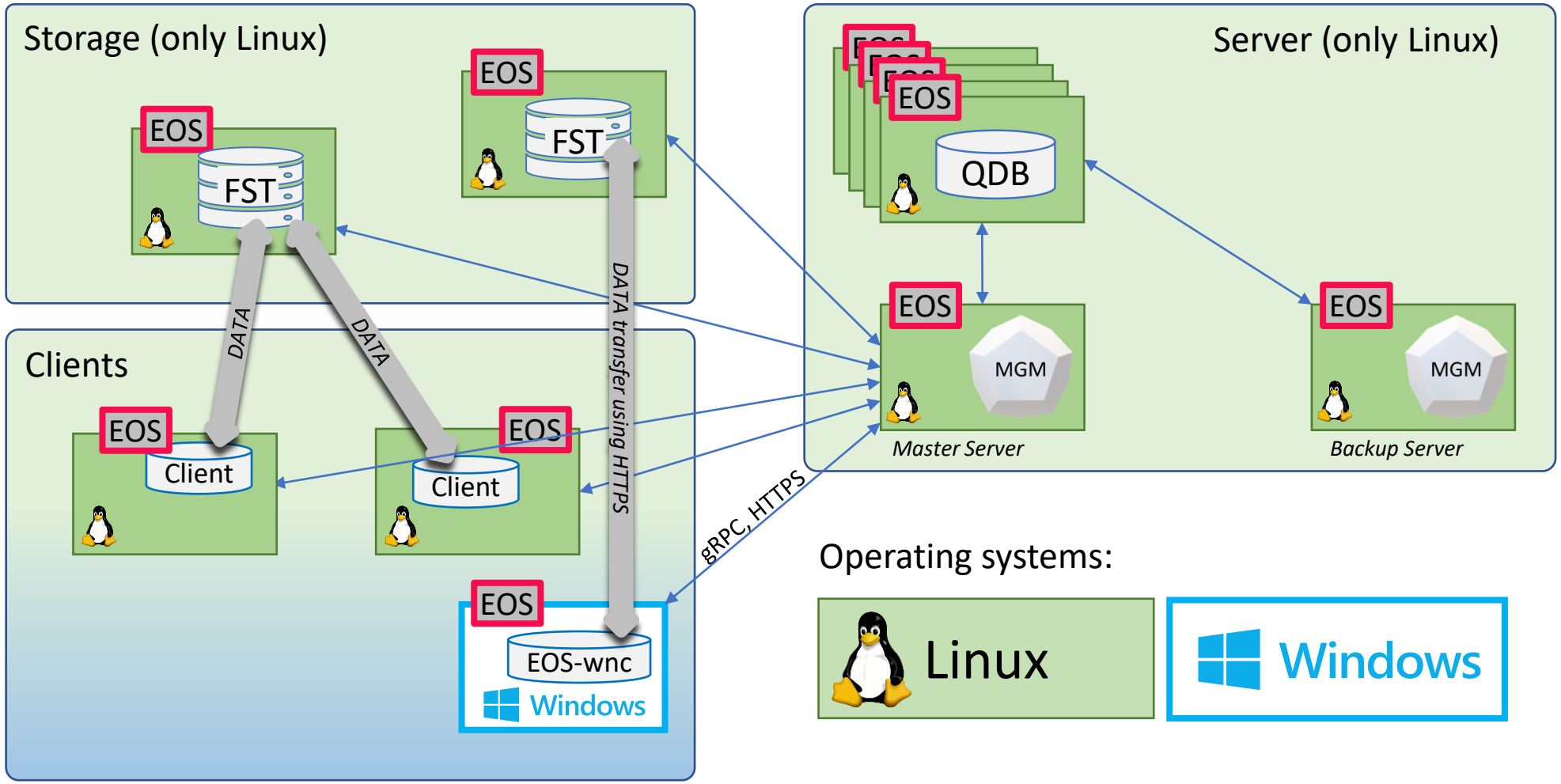
ii. EOS-drive on Windows

- General info about EOS-wnc: presented last year
  - Windows Driver Frameworks (WDF)
    - User-Mode Driver Framework (UMDF)
    - Preferred: Kernel-Mode Driver Framework (KMDF)
- File system filter driver filters I/O operations for a file system
- Proto buffer (started with gRPC, tightly connected with Proto buffer)
- Communication channels: Proto buffer, gRPC, https
- Dokan

The bridge: Windows through Samba



The bridge: Windows drive



Comtrade Three-Sixty

iii. Distributed FS Comparison

- CephFS
  - Using JBOD instead of RAID
  - Snapshots
  - Replication
  - File and directory layouts
- HDFS (Hadoop Distributed FS)
  - Using JBOD instead of RAID
  - Stores each file as a sequence of blocks which are replicated for fault tolerance
  - The block size and replication factor are configurable per file.
- GPFS (IBM Spectrum Scale)
  - Using IBM Spectrum Scale RAID
  - Snapshots
  - Synchronous and asynchronous replication
- CDFS (Comtrade Distributed FS)
  - Based on CERN EOS
  - Uses JBOD in the form of RAIN

Comparison testing: Download

1. Create new test files on the server space
2. Clear file cache on the clients and the servers
3. Download the files from the server space to the client
4. Verifying the MD5 hash and calculating the transfer speed from execution time
5. Remove created and copied test files

Comparison testing: Upload

1. Create new test files on the client.
2. Clear file cache on the client and the server.
3. Upload the files from the client to the server space
4. Verifying the MD5 hash and calculating the transfer speed from execution time
5. Remove created and copied test files

Comparison results

Iterations (EOS)	21	(checksums OK)						
Iterations (IBM)	21	(checksums OK)						
Iterations (Ceph)	23	(checksums OK)	Number of files	100				
Iterations (Hadoop)	14	(checksums OK)	File size [MB]	1				

		Test [MB/s]	min	max	avg	trim25%	Avg time [ms]
Upload	Linux	EOS: xrdcp command	142,86	181,82	165,24	165,87	6,05
		EOS Fusex	45,81	52,03	49,27	49,32	20,30
		IBM Spectrum Scale	54,08	58,82	55,99	55,92	17,86
		Ceph on Linux	145,99	156,25	150,62	150,50	6,64
		Hadoop on Linux	3,46	3,64	3,55	3,55	281,92
	Windows	EOS-wnc	14,25	15,18	14,75	14,76	67,78
		EOS-drive ST	9,70	10,00	9,88	9,89	101,22
		EOS: Samba	22,68	24,13	23,29	23,28	42,94
		Ceph on Win	50,28	56,50	53,52	53,51	18,68
		Hadoop on Win	3,13	3,21	3,18	3,18	314,61
Download	Linux	EOS: xrdcp command	95,24	196,08	169,56	174,52	5,90
		EOS Fusex	48,45	53,05	50,67	50,63	19,74
		IBM Spectrum Scale	158,73	187,27	174,76	175,16	5,72
		Ceph on Linux	7,47	110,13	94,19	97,40	10,62
		Hadoop on Linux	3,68	4,30	3,97	3,97	251,71
	Windows	EOS-wnc	10,66	11,17	10,88	10,88	91,90
		EOS-drive ST	17,95	19,03	18,44	18,43	54,24
		EOS: Samba	13,19	15,82	14,28	14,24	70,02
		Ceph on Win	1,93	47,25	35,38	37,60	28,26
		Hadoop on Win	1,69	2,60	2,20	2,21	454,84

Legend:	[MB/s]		
Red	0	-	10
Orange	10	-	20
Yellow	20	-	30
Light green	30	-	40
Green	40	-	∞

Iterations (EOS)	28	(checksums OK)						
Iterations (IBM)	28	(checksums OK)						
Iterations (Ceph)	52	(checksums OK)	Number of files	10				
Iterations (Hadoop)	11	(checksums OK)	File size [MB]	100				

		Test [MB/s]	min	max	avg	trim25%	Avg time [ms]
Upload	Linux	EOS: xrdcp command	359,71	444,44	411,14	412,67	243,22
		EOS Fusex	134,72	192,64	160,16	160,07	624,38
		IBM Spectrum Scale	176,62	188,71	181,08	181,01	552,25
		Ceph on Linux	131,89	162,39	140,86	139,97	709,95
		Hadoop on Linux	9,43	10,17	9,94	9,97	10064,96
	Windows	EOS-wnc	174,21	204,60	186,28	185,66	536,82
		EOS-drive ST	186,54	210,24	197,67	197,40	505,89
		EOS: Samba	165,56	231,33	196,82	196,65	508,08
		Ceph on Win	102,68	141,96	136,28	137,07	733,77
		Hadoop on Win	4,50	5,22	4,61	4,56	21670,61
Download	Linux	EOS: xrdcp command	301,20	436,68	412,94	417,50	242,16
		EOS Fusex	186,67	217,11	206,36	207,14	484,59
		IBM Spectrum Scale	306,75	345,18	322,89	322,24	309,70
		Ceph on Linux	20,44	183,49	31,49	28,27	3175,40
		Hadoop on Linux	8,06	10,51	9,37	9,39	10668,22
	Windows	EOS-wnc	128,10	177,31	151,39	151,01	660,54
		EOS-drive ST	148,70	185,92	157,76	156,40	633,87
		EOS: Samba	72,97	97,50	81,26	80,39	1230,60
		Ceph on Win	17,63	81,00	25,54	23,66	3915,54
		Hadoop on Win	4,31	4,62	4,50	4,51	22217,73

Legend:	[MB/s]		
Red	0	-	100
Orange	100	-	150
Yellow	150	-	200
Light green	200	-	250
Green	250	-	∞

Iterations (EOS)	27	(checksums OK)						
Iterations (IBM)	28	(checksums OK)						
Iterations (Ceph)	52	(checksums OK)	Number of files	2				
Iterations (Hadoop)	11	(checksums OK)	File size [MB]	2000				

		Test [MB/s]	min	max	avg	trim25%	Avg time [s]
Upload	Linux	EOS: xrdcp command	329,49	405,27	371,03	371,17	5,39
		EOS Fusex	187,92	237,63	210,76	210,51	9,49
		IBM Spectrum Scale	283,61	318,22	294,47	293,28	6,79
		Ceph on Linux	141,00	163,37	157,56	158,17	12,69
		Hadoop on Linux	9,74	10,10	9,91	9,91	201,83
	Windows	EOS-wnc	158,40	331,09	231,25	227,75	8,65
		EOS-drive ST	212,22	294,47	237,44	234,72	8,42
		EOS: Samba	164,11	229,82	181,25	178,59	11,03
		Ceph on Win	128,18	158,19	153,32	154,04	13,04
		Hadoop on Win	4,61	4,72	4,66	4,66	428,85
Download	Linux	EOS: xrdcp command	328,68	365,97	353,00	354,47	5,67
		EOS Fusex	218,66	233,36	227,13	227,15	8,81
		IBM Spectrum Scale	328,95	364,96	342,54	341,65	5,84
		Ceph on Linux	188,80	355,49	265,08	264,04	7,54
		Hadoop on Linux	9,28	10,63	10,12	10,15	197,66
	Windows	EOS-wnc	119,92	213,86	170,17	169,49	11,75
		EOS-drive ST	179,86	210,49	190,24	189,72	10,51
		EOS: Samba	17,95	35,43	25,85	25,54	77,37
		Ceph on Win	105,38	141,66	122,82	122,90	16,28
		Hadoop on Win	4,30	4,73	4,55	4,55	440,00

Legend:	[MB/s]		
Red	0	-	100
Orange	100	-	150
Yellow	150	-	200
Light green	200	-	250
Green	250	-	∞

^ST - Single-thread
^MT - Multi-thread

Comtrade Three-Sixty

Comparison results - Small Files

Iterations (EOS)	21	(checksums OK)
Iterations (IBM)	21	(checksums OK)
Iterations (Ceph)	23	(checksums OK)
Iterations (Hadoop)	14	(checksums OK)

Number of files	100
File size [MB]	1

		Test [MB/s]	min	max	avg	trim25%	Avg time [ms]
Upload	Linux	EOS: xrdcp command	142,86	181,82	165,24	165,87	6,05
		EOS Fusex	45,81	52,03	49,27	49,32	20,30
		IBM Spectrum Scale	54,08	58,82	55,99	55,92	17,86
		Ceph on Linux	145,99	156,25	150,62	150,50	6,64
		Hadoop on Linux	3,46	3,64	3,55	3,55	281,92
	Windows	EOS-wnc	14,25	15,18	14,75	14,76	67,78
		EOS-drive ST	9,70	10,00	9,88	9,89	101,22
		EOS: Samba	22,68	24,13	23,29	23,28	42,94
		Ceph on Win	50,28	56,50	53,52	53,51	18,68
		Hadoop on Win	3,13	3,21	3,18	3,18	314,61
Download	Linux	EOS: xrdcp command	95,24	196,08	169,56	174,52	5,90
		EOS Fusex	48,45	53,05	50,67	50,63	19,74
		IBM Spectrum Scale	158,73	187,27	174,76	175,16	5,72
		Ceph on Linux	7,47	110,13	94,19	97,40	10,62
		Hadoop on Linux	3,68	4,30	3,97	3,97	251,71
	Windows	EOS-wnc	10,66	11,17	10,88	10,88	91,90
		EOS-drive ST	17,95	19,03	18,44	18,43	54,24

Iterations (EOS)	
Iterations (IBM)	
Iterations (Ceph)	
Iterations (Hadoop)	

		Test [MB/s]	
Upload	Linux	EOS: xrdcp command	3
		EOS Fusex	1
		IBM Spectrum Scale	1
		Ceph on Linux	1
		Hadoop on Linux	
	Windows	EOS-wnc	1
		EOS-drive ST	1
		EOS: Samba	1
		Ceph on Win	1
		Hadoop on Win	
Download	Linux	EOS: xrdcp command	3
		EOS Fusex	1
		IBM Spectrum Scale	3
		Ceph on Linux	
		Hadoop on Linux	
	Windows	EOS-wnc	1
		EOS-drive ST	1

Comtrade Three-Sixty

Comparison results - Medium Files

Iterations (EOS)	28	(checksums OK)					
Iterations (IBM)	28	(checksums OK)					
Iterations (Ceph)	52	(checksums OK)	Number of files	10			
Iterations (Hadoop)	11	(checksums OK)	File size [MB]	100			

		Test [MB/s]	min	max	avg	trim25%		Avg time [ms]
Upload	Linux	EOS: xrdcp command	359,71	444,44	411,14	412,67	★	243,22
		EOS Fusex	134,72	192,64	160,16	160,07	★	624,38
		IBM Spectrum Scale	176,62	188,71	181,08	181,01	★	552,25
		Ceph on Linux	131,89	162,39	140,86	139,97		709,95
		Hadoop on Linux	9,43	10,17	9,94	9,97		10064,96
	Windows	EOS-wnc	174,21	204,60	186,28	185,66	★	536,82
		EOS-drive ST	186,54	210,24	197,67	197,40	★	505,89
		EOS: Samba	165,56	231,33	196,82	196,65	★	508,08
		Ceph on Win	102,68	141,96	136,28	137,07		733,77
		Hadoop on Win	4,50	5,22	4,61	4,56		21670,61
Download	Linux	EOS: xrdcp command	301,20	436,68	412,94	417,50	★	242,16
		EOS Fusex	188,87	217,11	208,36	207,14	★	484,59
		IBM Spectrum Scale	306,75	345,18	322,89	322,24	★	309,70
		Ceph on Linux	20,44	183,49	31,49	28,27		3175,40
		Hadoop on Linux	8,06	10,51	9,37	9,39		10668,22
	Windows	EOS-wnc	128,10	177,31	151,39	151,01	★	660,54
		EOS-drive ST	148,70	185,92	157,76	156,40	★	633,87
		EOS: Samba	72,97	97,50	81,26	80,39	★	1230,60

		Test [MB/s]
Upload	Linux	EOS: xrdcp command
		EOS Fusex
		IBM Spectrum Scale
		Ceph on Linux
		Hadoop on Linux
	Windows	EOS-wnc
		EOS-drive ST
		EOS: Samba
		Ceph on Win
		Hadoop on Win
Download	Linux	EOS: xrdcp command
		EOS Fusex
		IBM Spectrum Scale
		Ceph on Linux
		Hadoop on Linux
	Windows	EOS-wnc
		EOS-drive ST
		EOS: Samba

Comtrade Three-Sixty

Comparison results - Large Files

10	
100	

ime [ms]
243,22
624,38
552,25
709,95
10064,96
536,82
505,89
508,08
733,77
21670,61
242,16
484,59
309,70
3175,40
10668,22
660,54
633,87

Iterations (EOS)	27	(checksums OK)	
Iterations (IBM)	28	(checksums OK)	
Iterations (Ceph)	52	(checksums OK)	Number of files 2
Iterations (Hadoop)	11	(checksums OK)	File size [MB] 2000

		Test [MB/s]	min	max	avg	trim25%	Avg time [s]
Upload	Linux	EOS: xrdcp command	329,49	405,27	371,03	371,17	5,39
		EOS Fusex	187,92	237,63	210,76	210,51	9,49
		IBM Spectrum Scale	283,61	318,22	294,47	293,28	6,79
		Ceph on Linux	141,00	163,37	157,56	158,17	12,69
		Hadoop on Linux	9,74	10,10	9,91	9,91	201,83
	Windows	EOS-wnc	158,40	331,09	231,25	227,75	8,65
		EOS-drive ST	212,22	294,47	237,44	234,72	8,42
		EOS: Samba	164,11	229,82	181,25	178,59	11,03
		Ceph on Win	128,18	158,19	153,32	154,04	13,04
		Hadoop on Win	4,61	4,72	4,66	4,66	428,85
Download	Linux	EOS: xrdcp command	328,68	365,97	353,00	354,47	5,67
		EOS Fusex	218,86	233,36	227,13	227,15	8,81
		IBM Spectrum Scale	328,95	364,96	342,54	341,65	5,84
		Ceph on Linux	188,80	355,49	265,08	264,04	7,54
		Hadoop on Linux	9,28	10,63	10,12	10,15	197,66
	Windows	EOS-wnc	119,92	213,86	170,17	169,49	11,75
		EOS-drive ST	179,86	210,49	190,24	189,72	10,51

Comtrade Three-Sixty

Comparison results – Windows vs Linux

10
100

ime [ms]
243,22
624,38
552,25
709,95
10064,96
536,82
505,89
508,08
733,77
21670,61
242,16
484,59
309,70
3175,40
10668,22
660,54
633,87

Iterations (EOS)	27	(checksums OK)		
Iterations (IBM)	28	(checksums OK)		
Iterations (Ceph)	52	(checksums OK)	Number of files	2
Iterations (Hadoop)	11	(checksums OK)	File size [MB]	2000

		Test [MB/s]	min	max	avg	trim25%		Avg time [s]
Upload	Linux	EOS: xrdcp command	329,49	405,27	371,03	371,17	★	5,39
		EOS Fusex	187,92	237,63	210,76	210,51	★	9,49
		IBM Spectrum Scale	283,61	318,22	294,47	293,28	★	6,79
		Ceph on Linux	141,00	163,37	157,56	158,17		12,69
		Hadoop on Linux	9,74	10,10	9,91	9,91		201,83
	Windows	EOS-wnc	158,40	331,09	231,25	227,75	★	8,65
		EOS-drive ST	212,22	294,47	237,44	234,72	★	8,42
		EOS: Samba	164,11	229,62	181,25	178,99	★	11,03
		Ceph on Win	128,18	158,19	153,32	154,04		13,04
		Hadoop on Win	4,61	4,72	4,66	4,66		428,85
Download	Linux	EOS: xrdcp command	328,68	365,97	353,00	354,47	★	5,67
		EOS Fusex	218,66	233,36	227,13	227,15		8,81
		IBM Spectrum Scale	328,95	364,96	342,54	341,65	★	5,84
		Ceph on Linux	188,80	355,49	265,08	264,04	★	7,54
		Hadoop on Linux	9,28	10,63	10,12	10,15		197,66
	Windows	EOS-wnc	119,92	213,86	170,17	169,49	★	11,75
		EOS-drive ST	179,86	210,49	190,24	189,72	★	10,51

Comtrade Three-Sixty

Interpretation of results

The best

- Small files
  1. EOS on Linux
  1. Ceph on Linux
  1. GPFS on Linux
- Medium files
  1. EOS on Linux
  2. GPFS on Linux
- Large files
  1. EOS on Linux
  2. GPFS on Linux

Not the best

- All file sizes
  - Hadoop on Win
  - Hadoop on Linux
  - Samba

Plans for comparison

High Availability metrics

- MTBF
  - Mean time between failures
- Failover resync time
- Resync of replaced disk

High Availability requirements

- Load balancing
- Data scalability
- Geographical diversity
- Backup to tape

Comtrade Three-Sixty

Thank You!

Update on EOS productisation - Comtrade 360's results

Luca Mascetti
Elvin Sindrilaru
CERN IT Storage and Data Management

luca.mascetti@cern.ch
elvin.alin.sindrilaru@cern.ch

Gregor Molan
Comtrade 360's AI Lab

gregor.molan@comtrade.com