Comtrade Three-Sixty

Comparison between **CDFS** (Comtrade Distributed FS), **CephFS**, **HDFS** (Hadoop Distributed FS), **GPFS** (IBM Spectrum Scale)

Gregor Molan | **Branko Blagojević** | **Ivan Arizanović**

Comtrade Group / Comtrade 360

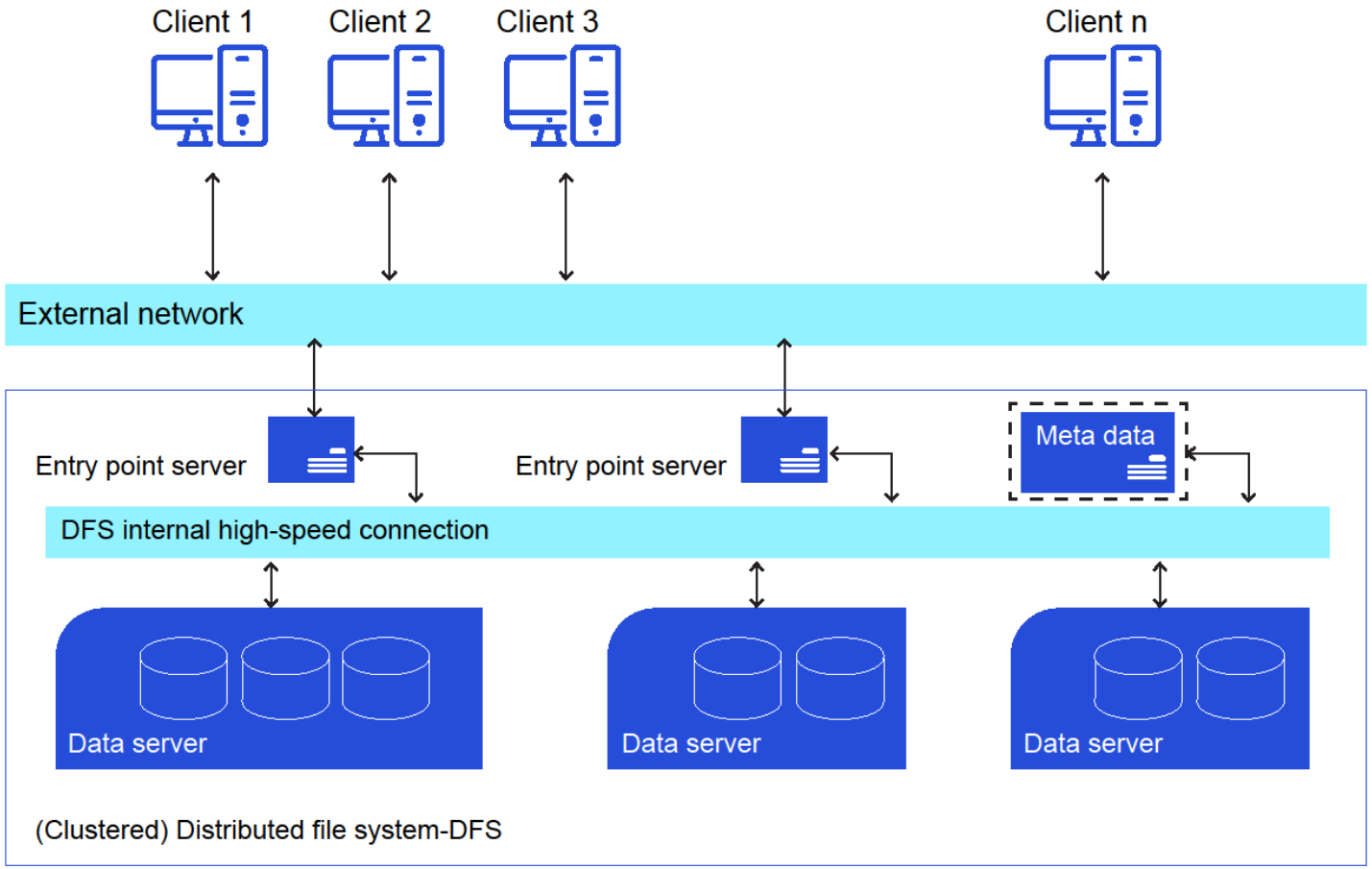
Content

- Introduction
- (Clustered) Distributed file system
- High-performance file systems: **Selection**
- High-performance file systems: **Features**
- High-performance file systems: **Results of comparison**
- Conclusion

Introduction

- **Introduction**
  - **(Clustered) Distributed file system**
- High-performance file systems: **Selection**
- High-performance file systems: **Features**
- High-performance file systems: **Results of comparison**
- Conclusion

(Clustered) **Distributed file system**



Comtrade Three-Sixty

High-performance file systems: **Selection**

- Introduction
- **High-performance file systems: Selection**
  - **CephFS**
  - **HDFS (Hadoop Distributed File System)**
  - **GPFS (IBM Spectrum Scale)**
  - **CDFS (Comtrade Distributed FS)**
- High-performance file systems: Features
- High-performance file systems: Results of comparison
- Conclusion

Comtrade Three-Sixty

CephFS

- Massively scalable
- Part of the Linux kernel since 2010
- RADOS: Ceph's foundation

Interfaces:

- S3-compatible
- Swift-compatible

Comtrade Three-Sixty

HDFS (Hadoop Distributed File System)

Basic features

- Distributed
- Scalable
- Portable
- Written in Java

HDFS services

- Name Node (master)
- Data Node
- Secondary Name Node
- Job tracker
- Task Tracker

GPFS (IBM Spectrum Scale)

- #1 HPC in 2019 (Oak Ridge National Laboratory)
- OS support for server:
  - AIX
  - Linux
  - Windows
- Configuration update on a mounted file system
- Data replication
- Active File Management (AFM)
  - Data share across clusters

CDFS (Comtrade Distributed FS)

CDFS = appliance of CERN EOS at Comtrade

- Comtrade: The industry partner of CERN EOS
- The initial usage of CERN EOS
  - Data collection for CERN LHC experiments
- Current usage of CERN EOS
  - Data collection for all CERN experiments
  - The primary data storage software (including CERN staff data)
- The fastest file system for parallel data collection
- Cluster size: > 700 PB
  - Node resync: < 15 minutes

Comtrade Three-Sixty

High-performance file systems: **Features**

- Introduction
- High-performance file systems: Selection
- **High-performance file systems: Features**
  - **Requirements**
  - **Fault tolerance**
  - **Advantages of RAIN**
- High-performance file systems: Results of comparison
- Conclusion

Comtrade Three-Sixty

Requirements

- High throughput
- Low latency
- Expendability
- High availability
- High reliability

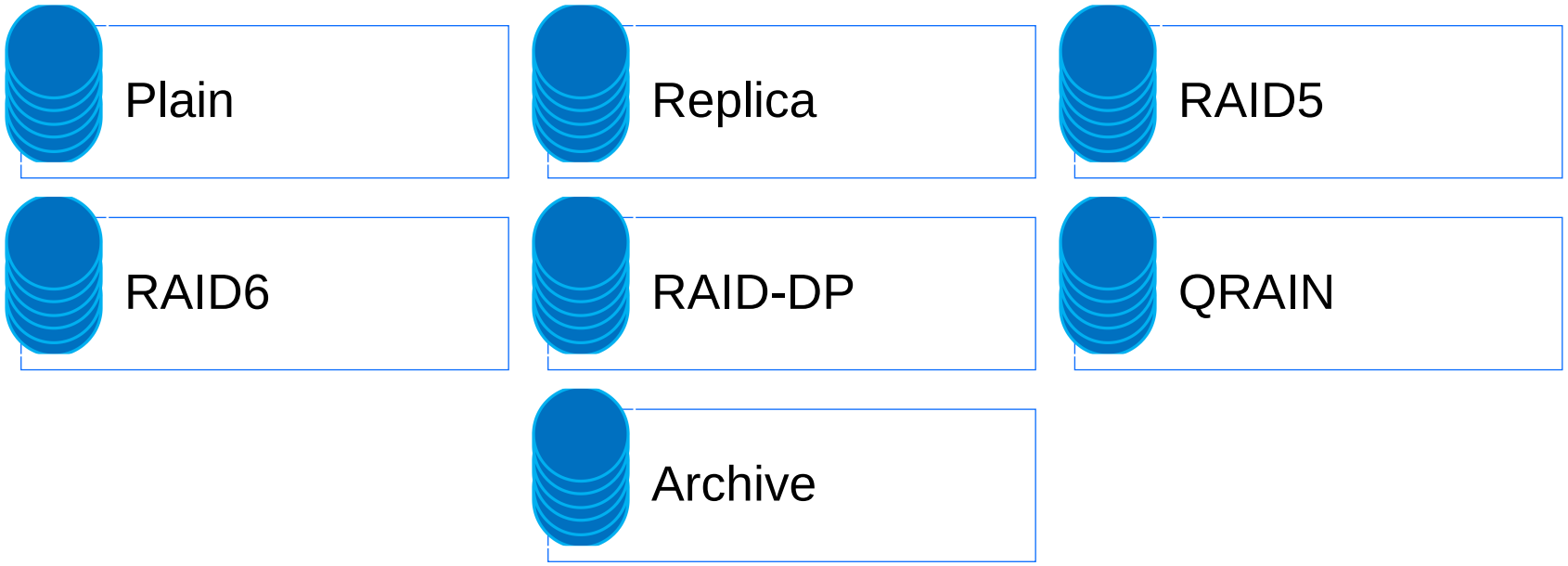
Comtrade Three-Sixty

Fault tolerance

- CephFS
  - Using JBOD instead of RAID
  - Snapshots
  - Replication
  - File and directory layouts
- HDFS
  - 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
  - Using IBM Spectrum Scale RAID
  - Snapshots
  - Synchronous and asynchronous replication
- CDFS (based on EOS)
  - Uses JBOD in the form of RAIN

About the RAIN – **File Layouts**

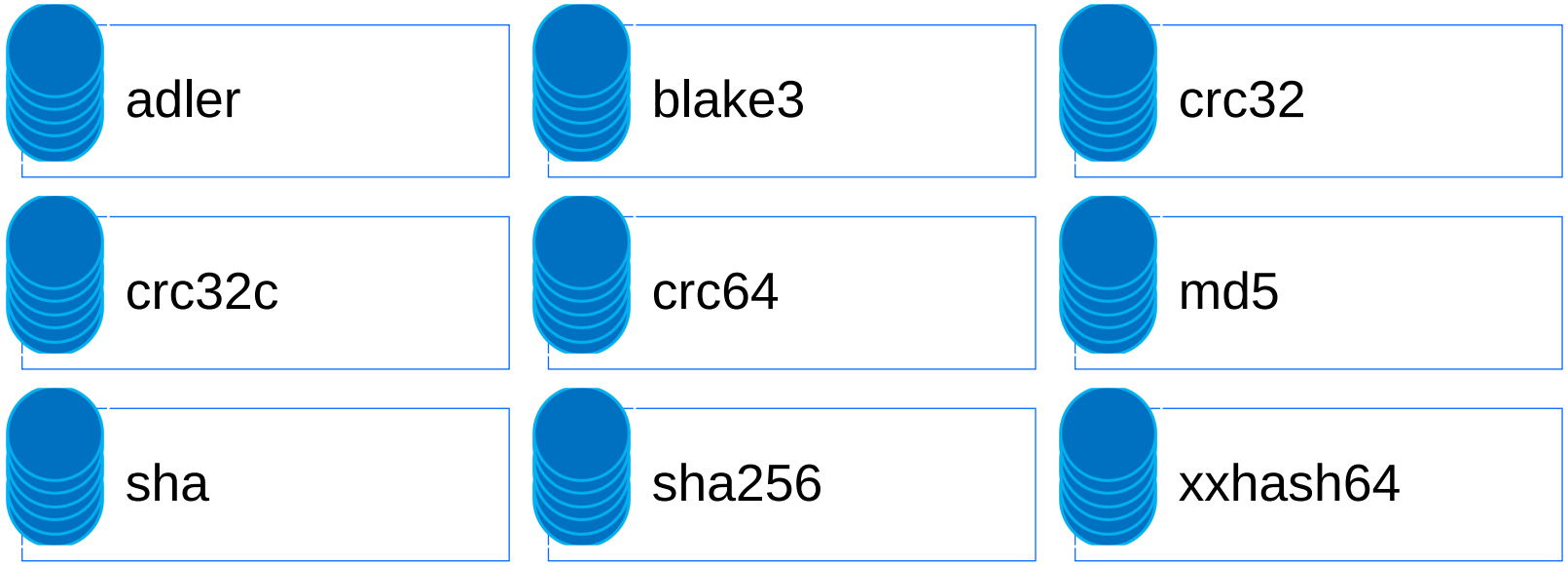
RAIN = Software implementation of the RAID concept across independent servers on the network



Comtrade Three-Sixty

About the RAIN – **Checksums**

RAIN = Checksums calculated and recorded for every file (chunk)



Comtrade Three-Sixty

Advantages of **RAIN**

Advantages

- Scalability
- Reliability
- Cost (JBOD without RAID controller)
- Geotag policies are applied during file placement to improve data loss prevention and IO performance.

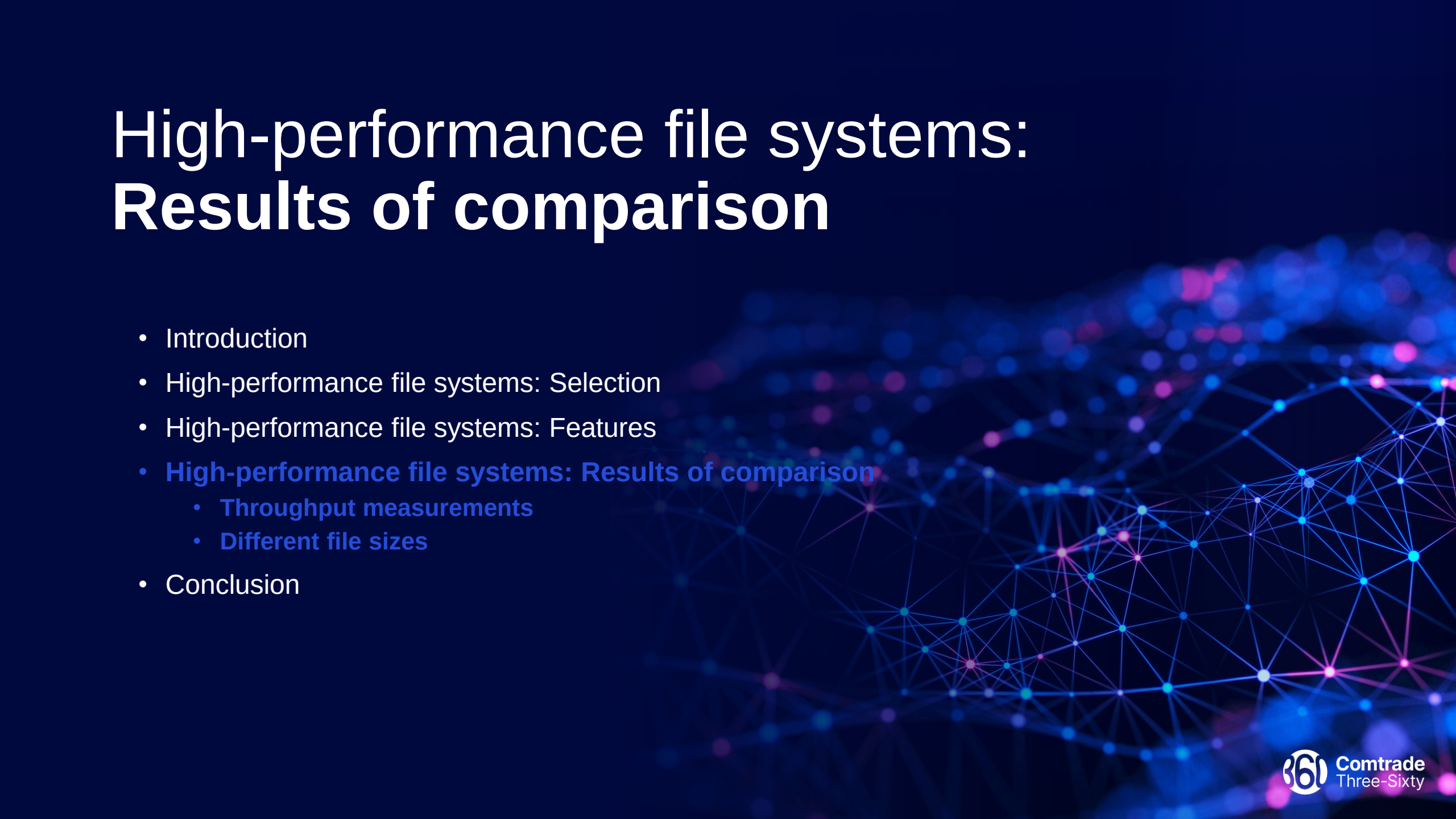
Drawbacks

- All communication is done via the network
- Increased is IO and computational effort for non-sequential writes and server draining

Comtrade Three-Sixty

High-performance file systems: **Results of comparison**

- Introduction
- High-performance file systems: Selection
- High-performance file systems: Features
- **High-performance file systems: Results of comparison**
  - **Throughput measurements**
  - **Different file sizes**
- Conclusion



Comtrade Three-Sixty

Testing environment

- DFS as a single node*
- Clusters on different HDDs
- Identical disk drives
- Identical HW components:
  - Motherboards
  - Network adapters
  - Memory

	Cores	Memory	HDD
Client: Linux	2	4 GB	80 GB
Client: Windows	4	10 GB	80 GB
Servers: Linux	8	20 GB	500 GB

Comtrade Three-Sixty

Different file sizes

Small files

- Size: 1 MB
- Transfer: 100 files at once
- Potential issue:
  - Authentication time overhead

Medium files

- Size: 100 MB
- Transfer: 10 files at once
- Potential issue: -

Larger files

- Size: 2 GB
- Transfer: 2 files at once
- Potential issue:
  - Time out

Comtrade Three-Sixty

Testing description

Download

- Create new test files on the server space
- Clear file cache on the client and the server machines
- Download the files from the server space to the client machine
- Verifying the MD5 hash and calculating the transfer speed from execution time
- Remove created and copied test files

Upload

- Create new test files on the client machine
- Clear file cache on the client and the server machines
- Upload the files from the client machine to the server space
- Verifying the MD5 hash and calculating the transfer speed from execution time
- Remove created and copied test files

Comtrade Three-Sixty

Throughput 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

Throughput 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
		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

Iterations (EOS)	28 (checksums
Iterations (IBM)	28 (checksums
Iterations (Ceph)	52 (checksums
Iterations (Hadoop)	11 (checksums

		Test [MB/s]	min	max	a
Upload	Linux	EOS: xrdcp command	359,71	444,44	41
		EOS Fusex	134,72	192,64	16
		IBM Spectrum Scale	176,62	188,71	18
		Ceph on Linux	131,89	162,39	14
		Hadoop on Linux	9,43	10,17	
	Windows	EOS-wnc	174,21	204,60	18
		EOS-drive ST	186,54	210,24	19
		EOS: Samba	165,56	231,33	19
		Ceph on Win	102,68	141,96	13
		Hadoop on Win	4,50	5,22	
Download	Linux	EOS: xrdcp command	301,20	436,68	41
		EOS Fusex	186,67	217,11	20
		IBM Spectrum Scale	306,75	345,18	32
		Ceph on Linux	20,44	183,49	3
		Hadoop on Linux	8,06	10,51	
	Windows	EOS-wnc	128,10	177,31	15
		EOS-drive ST	148,70	185,92	15
		EOS: Samba	72,97	97,50	8
		Ceph on Win	17,63	81,00	2

Throughput results - Medium Files

s	100
	1

g time [ms]
6,05
20,30
17,86
6,64
281,92
67,78
101,22
42,94
18,68
314,61
5,90
19,74
5,72
10,62
251,71
91,90
54,24
70,02
28,26

Iterations (EOS)	28	(checksums OK)
Iterations (IBM)	28	(checksums OK)
Iterations (Ceph)	52	(checksums OK)
Iterations (Hadoop)	11	(checksums OK)

Number of files	10
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	182,64	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	208,38	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

Iterations (EOS)	27	(checksums
Iterations (IBM)	28	(checksums
Iterations (Ceph)	52	(checksums
Iterations (Hadoop)	11	(checksums

		Test [MB/s]	min	max	
Upload	Linux	EOS: xrdcp command	329,49	405,27	3
		EOS Fusex	187,92	237,63	2
		IBM Spectrum Scale	283,61	318,22	2
		Ceph on Linux	141,00	163,37	1
		Hadoop on Linux	9,74	10,10	
	Windows	EOS-wnc	158,40	331,09	2
		EOS-drive ST	212,22	294,47	2
		EOS: Samba	164,11	229,82	1
		Ceph on Win	128,18	158,19	1
		Hadoop on Win	4,61	4,72	
Download	Linux	EOS: xrdcp command	328,68	365,97	3
		EOS Fusex	218,66	233,36	2
		IBM Spectrum Scale	328,95	364,96	3
		Ceph on Linux	188,80	355,49	2
		Hadoop on Linux	9,28	10,63	
	Windows	EOS-wnc	119,92	213,86	1
		EOS-drive ST	179,86	210,49	1
		EOS: Samba	17,95	35,43	
		Ceph on Win	105,38	141,66	1

Throughput results - Large Files

s	10
	100

Avg time [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
☆	1230,60
	3915,54

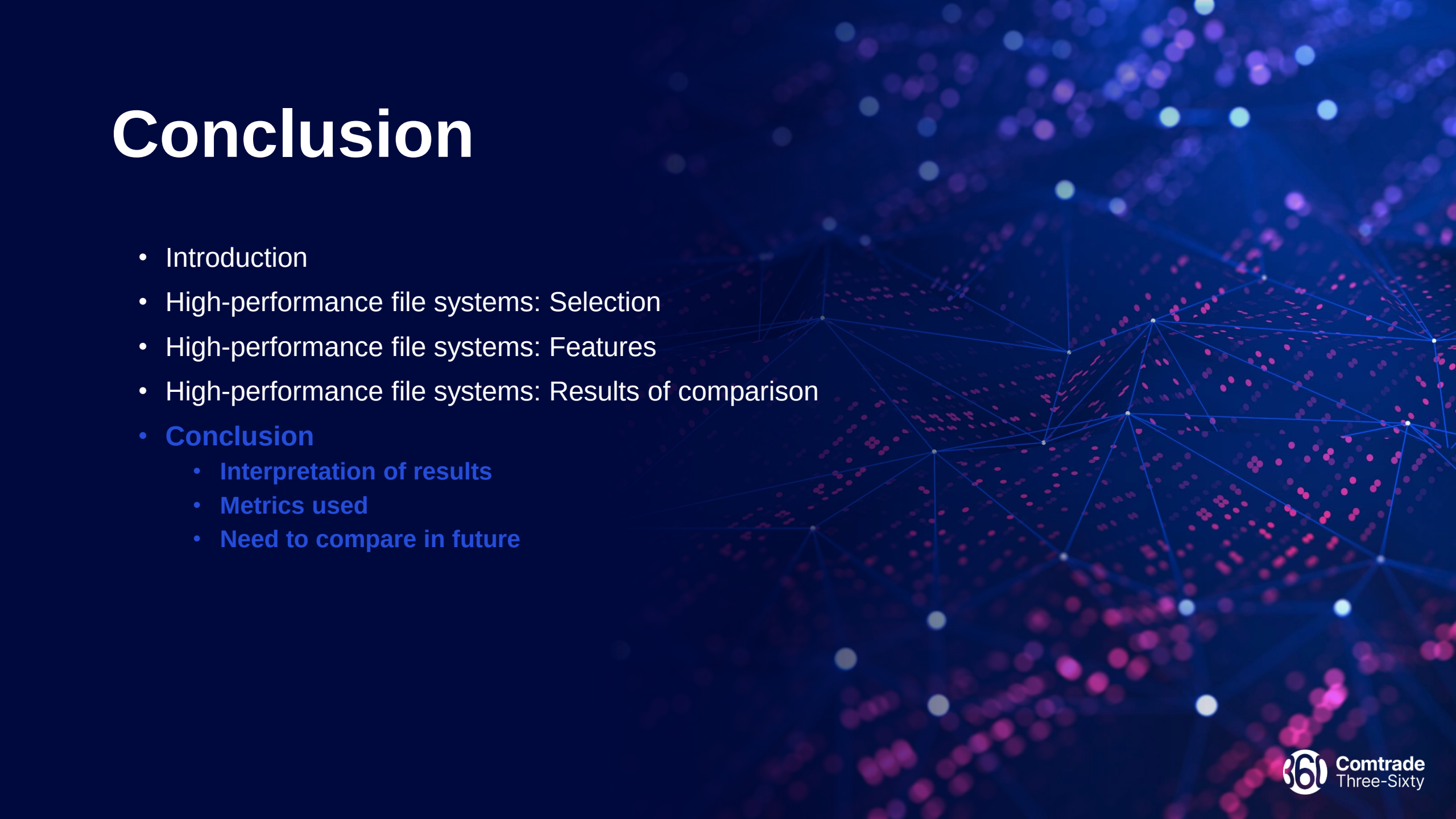
Iterations (EOS)	27	(checksums OK)
Iterations (IBM)	28	(checksums OK)
Iterations (Ceph)	52	(checksums OK)
Iterations (Hadoop)	11	(checksums OK)

Number of files	2
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
	Linux: EOS Fusex	187,92	237,63	210,76	210,51	☆ 9,49
	Linux: IBM Spectrum Scale	283,61	318,22	294,47	293,28	★ 6,79
	Linux: Ceph on Linux	141,00	163,37	157,56	158,17	12,69
	Linux: 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
	Windows: EOS-drive ST	212,22	294,47	237,44	234,72	★ 8,42
	Windows: EOS: Samba	164,11	229,82	181,25	178,59	☆ 11,03
	Windows: Ceph on Win	128,18	158,19	153,32	154,04	13,04
	Windows: 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
	Linux: EOS Fusex	218,86	233,36	227,13	227,15	8,81
	Linux: IBM Spectrum Scale	328,95	364,96	342,54	341,65	★ 5,84
	Linux: Ceph on Linux	188,80	355,49	265,08	264,04	☆ 7,54
	Linux: 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
	Windows: EOS-drive ST	179,86	210,49	190,24	189,72	★ 10,51
	Windows: EOS: Samba	17,95	35,43	25,85	25,54	77,37
	Windows: Ceph on Win	105,38	141,66	122,82	122,90	☆ 16,28

Conclusion

- Introduction
- High-performance file systems: Selection
- High-performance file systems: Features
- High-performance file systems: Results of comparison
- **Conclusion**
  - **Interpretation of results**
  - **Metrics used**
  - **Need to compare in future**



Comtrade Three-Sixty

Interpretation of results

The best

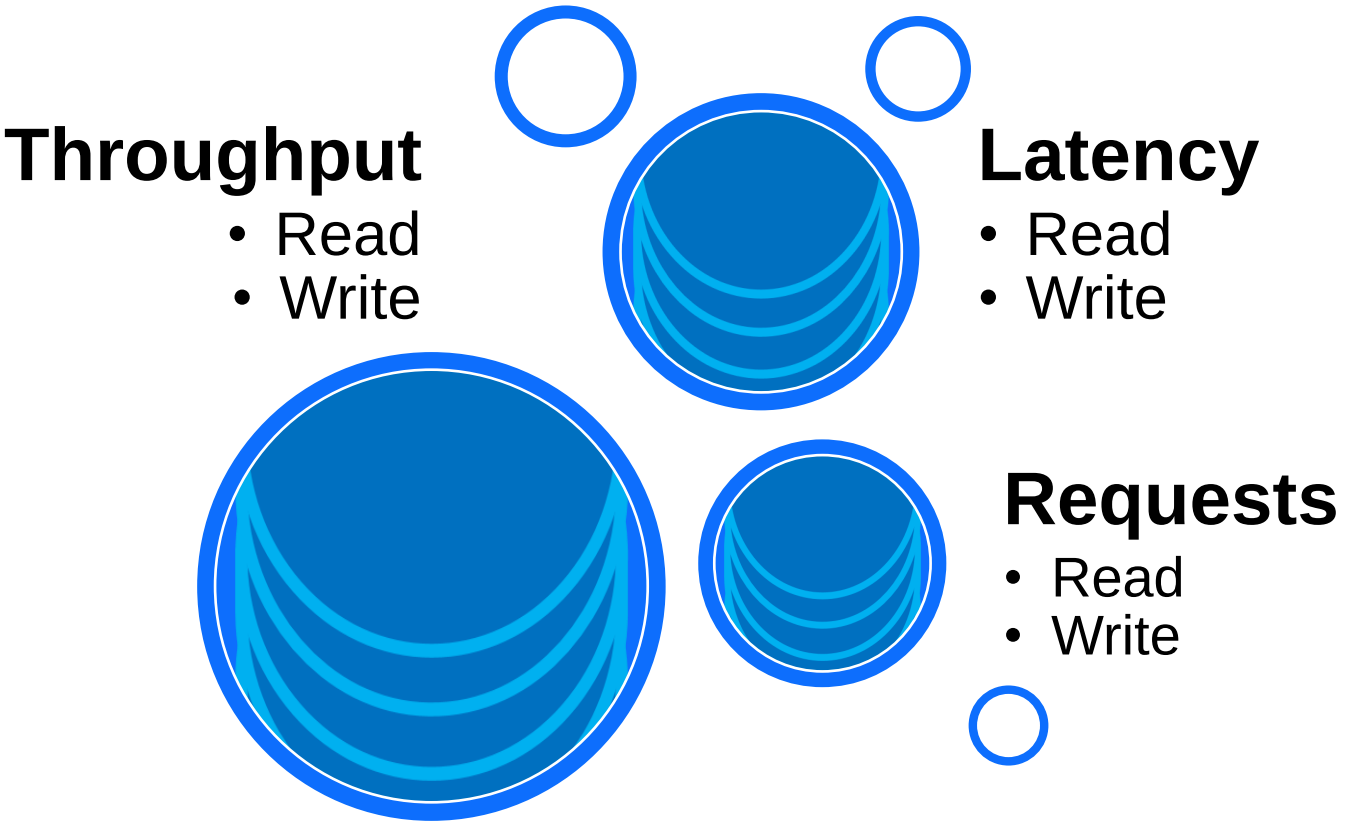
- Small files
  1. EOS on Linux
  1. Ceph on Linux
  1. GPFS on Linux
- Medium files
  1. EOS on Linux
  2. GPFS on Linux
- Large files
  1. EOS on Linux
  2. GPFS on Linux

Not the best

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

Comtrade Three-Sixty

Metrics used (data and metadata)



Need to compare in future

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

Comtrade Three-Sixty

Thank you

Comparison between **CDFS** (Comtrade Distributed FS), **CephFS**, **HDFS** (Hadoop Distributed FS), **GPFS** (IBM Spectrum Scale)

Gregor Molan
gregor@comtrade.com

Branko Blagojević

Ivan Arizanović

Comtrade Group / Comtrade 360