

Как мы разворачивали своё распределённое S3 хранилище на базе MinIO

Алексей Плетнёв

Базис-Центр

zix@bazissoft.ru



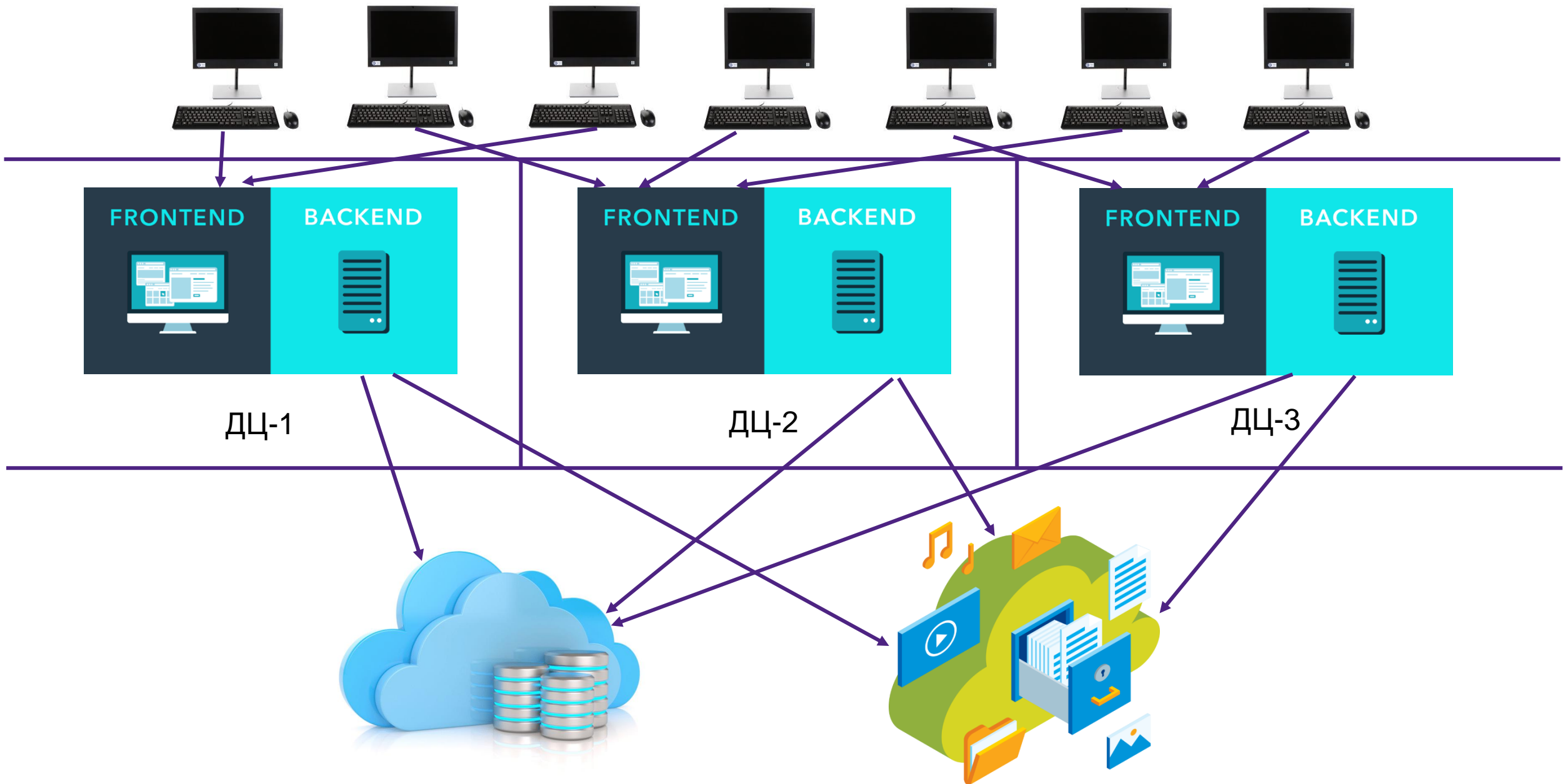
Зачем нам хранилище, почему распределённое и почему S3

Немного о компании

БАЗИС – комплексная система автоматизации проектирования и технологической подготовки производства корпусной мебели.

- 1988 г. – первая версия АС БАЗИС для СМ ЭВМ 3 поколения
- 1997 г. – первая версия для ОС Windows
- 2005 г. – первая версия с трёхмерным математическим ядром
- 2010 г. – флагманы российской мебельной промышленности полностью автоматизировали циклы производства с помощью САПР БАЗИС
- 2013 г. – полная совместимость программных решений с оборудованием ведущих производителей
- 2016 г. – выход на международный рынок

P.S. Команда разработчиков 15 человек



Какие есть варианты

ОБЛАЧНЫЕ



Google
Cloud
Storage



Yandex Cloud

Yandex Object
Storage



ЛОКАЛЬНЫЕ



MINIO



OpenIO



SCALITY



ceph

Сравнение стоимости

Сравнение стоимости (с поддержкой)

Почему MinIO

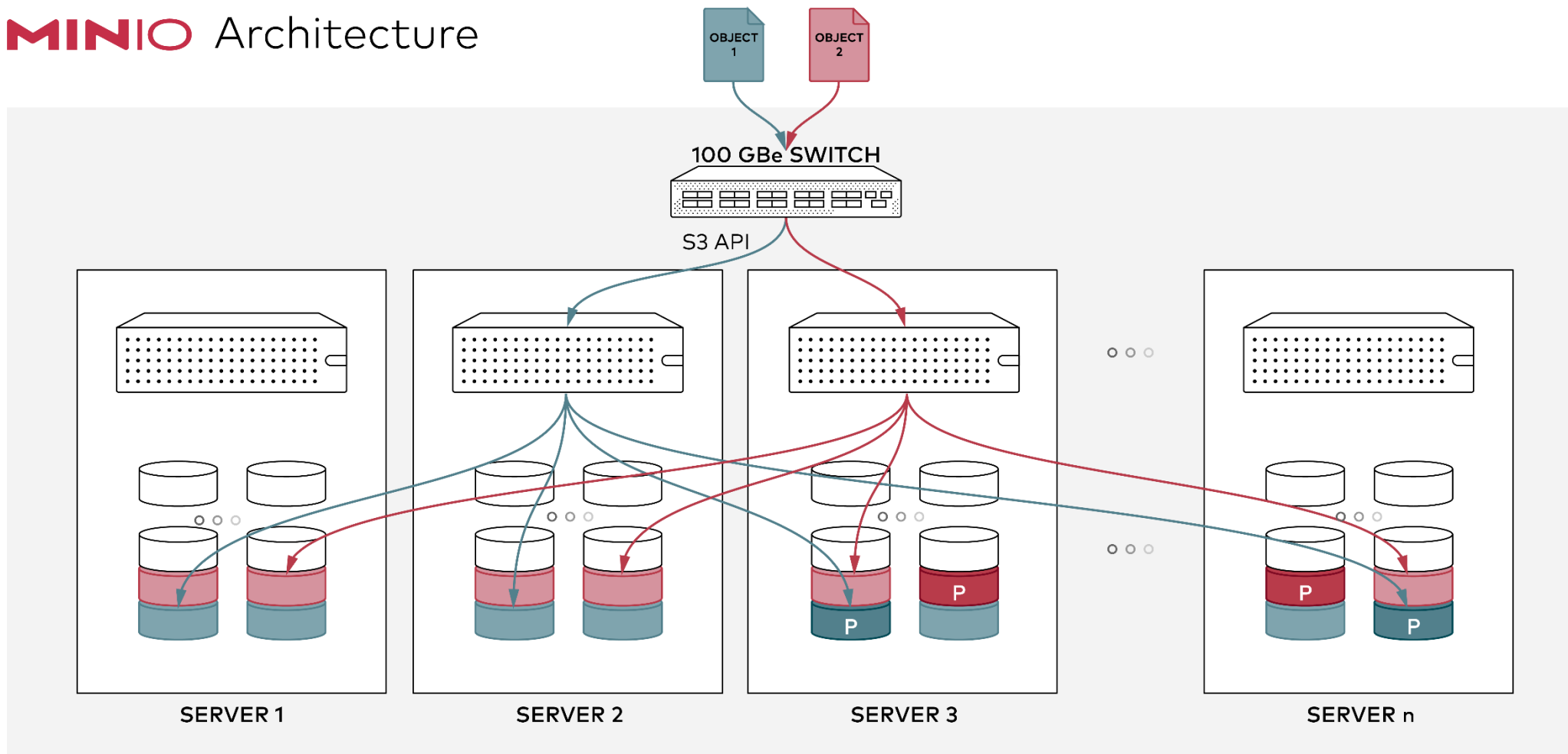


```
λ mc support perf minio-aws
(With 64 MiB object size, 32 concurrency) PUT: 156 GiB/s GET: 304 GiB/s
(With 64 MiB object size, 48 concurrency) PUT: 166 GiB/s GET: 318 GiB/s
(With 64 MiB object size, 72 concurrency) PUT: 167 GiB/s GET: 325 GiB/s
```

```
MinIO 2022-01-08T03:11:54Z, 32 servers, 256 drives
PUT: 167 GiB/s, 2,670 objs/s
GET: 325 GiB/s, 5,197 objs/s
```

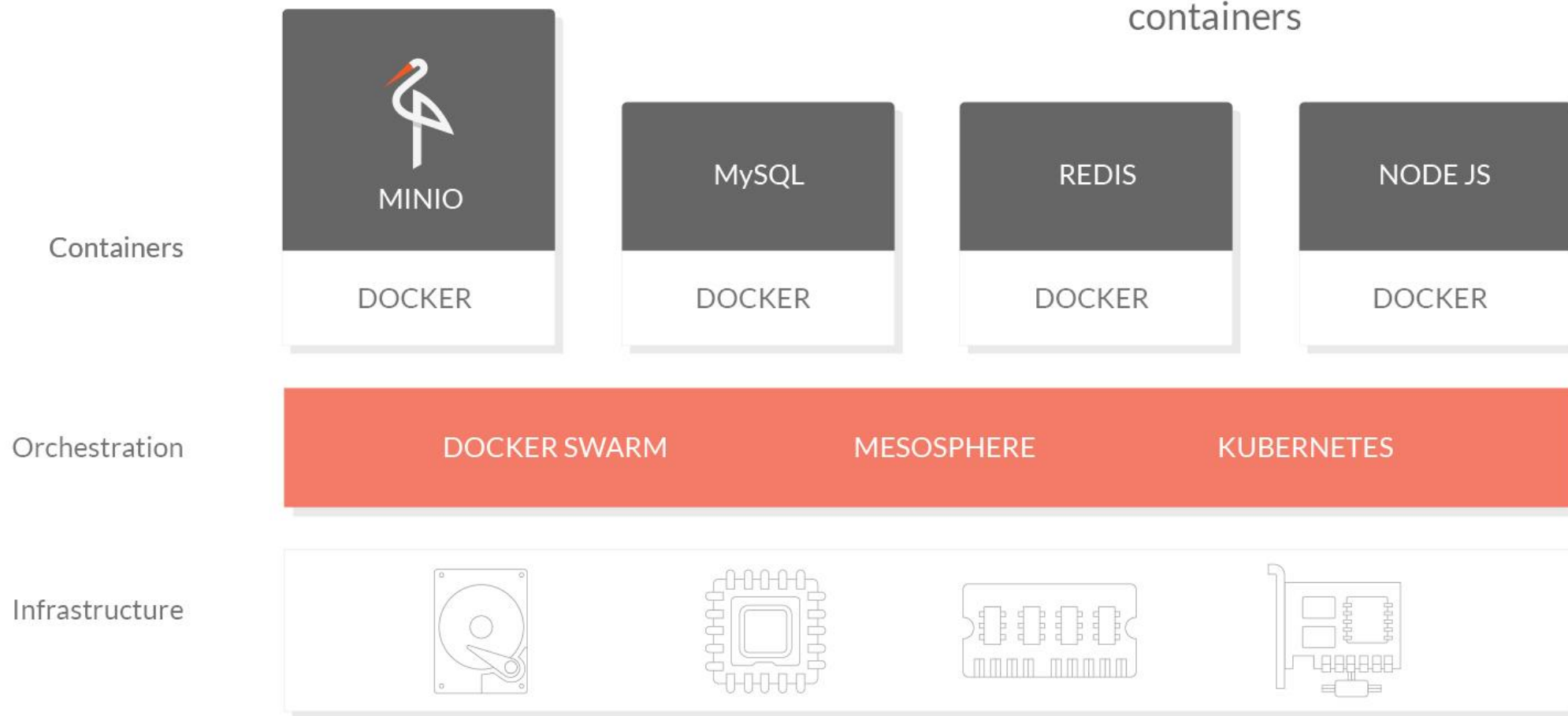
```
λ
```

MINIO Architecture



```
minio server http://host{1...n}/export{1...m}
```

Provision object storage as containers





50 KB




20 KB

Compression Guide

slack channel 18817

MinIO server allows streaming compression to ensure efficient disk space usage. Compression happens inflight, i.e objects are compressed before being written to disk(s). MinIO uses `klauspost/compress/s2` streaming compression due to its stability and performance.

This algorithm is specifically optimized for machine generated content. Write throughput is typically at least 500MB/s per CPU core, and scales with the number of available CPU cores. Decompression speed is typically at least 1GB/s.



[Pull requests](#) [Issues](#) [Marketplace](#) [Explore](#)

minio / minio Public

[Edit Pins](#) [Watch](#) 600


[Code](#) [Issues](#) 16 [Pull requests](#) 9 [Discussions](#) [Actions](#) [Security](#) 7 [Insights](#)












master

5 branches

325 tags

[Go to file](#) [Add file](#) [Code](#)

 klauspost Make isIndexedMetaV2 return errors (#15012) ... ✓ f7cecf0 7 hours ago 🕒 9,552 commits

	.github	Add tests for Access Management Plugin (#14909)	19 days ago
	buildscripts	feat: Single drive XL implementation (#14970)	2 days ago
	cmd	Make isIndexedMetaV2 return errors (#15012)	7 hours ago
	dockerscripts	add code to parse secrets natively instead of shell scripts (#13883)	6 months ago
	docs	fix: s3-check-md5 to not panic for incomplete md5	yesterday
	helm-releases	update helm v4.0.2	23 days ago
	helm/minio	Add PVC annotations to StatefulSet PVC templates (#14915)	16 days ago
	internal	handle IPv6 sourceIPs properly (#15005)	20 hours ago
	.dockerignore	simplify dockerfiles and remove duplication (#12419)	12 months ago
	.gitignore	add healing for invalid shards by skipping the blocks (#13978)	5 months ago
	.golangci.yml	start using t.SetEnv instead of os.Setenv (#14787)	last month

About

Multi-Cloud Object Storage

[min.io/download](#)

go

kubernetes

cloud

k8s


cloudnative


objectst


multi-cloud


cloudstorage


multi-cloud-kubernetes


 Readme

 AGPL-3.0 license


 Code of conduct

 33.4k stars

 600 watching

 3.9k forks

Releases 324

 Bugfix Release Latest

Проблемы при внедрении



MinIO fault tolerance with big drives and nodes counts #15026



Answered by klauspost

AlexZIX asked this question in Q&A



AlexZIX 13 days ago



Erasure code (EC)

OBJECT ERASURE-CODED OVER 16 DRIVES
Tolerates up to any 8 disk failures

DATA BLOCK



PARITY BLOCK



EC:N – доступно об отказоустойчивости

Erasure Set Size	Default Parity (EC:N) <small>EC – группа дисков размером от 4 до 16 M – всего дисков на всех нодах</small>
5 or fewer	EC:2 <small>N – диски с информацией для восстановления</small>
6-7	EC:3 <ul style="list-style-type: none">• $N \leq \text{TRUNC}(M / 2)$• $N \leq 8$• Потеря $N/2$ дисков = readonly
8 or more	EC:4 <ul style="list-style-type: none">• Потеря $N/2-1$ дисков = readwrite

EC:N – доступно об отказоустойчивости

For example, an 16-server distributed setup with 200 disks per node would continue serving files, up to 4 servers can be offline in default configuration i.e. around 800 disks down MinIO would continue to read and write objects.

- $N \leq \text{TRUNC}(M / 2)$
- $N \leq 8$

ЕС:N – доступно об отказоустойчивости

`http://minio-{1...16}.example.net/disk{1...200}`



ЕС:N – доступно об отказоустойчивости

`http://minio-{1...16}.example.net/disk{1...200}`

`http://minio-{1...16}/disk1 http://minio-{1...16}/disk2 ... http://minio-{1...16}/disk200`

Наигрались – давайте денег



harshavardhana 13 days ago Maintainer

@AlexZIX these discussions are best had on a call with our sales involved feel free to reach out to us at sales@min.io

We can recommend hardware and size your cluster appropriately.



ravindk89 13 days ago Collaborator



@AlexZIX MinIO automatically handles generating erasure code sets - and in fact your first syntax would create multiple single-disk pools, which would be very inadvisable.

We have general documentation on running MinIO in distributed mode [here](#). As noted, more specific sizing and hardware recommendations are handled through professional engagement.

ЕС:N – доступно об отказоустойчивости

Total Drives (N)	Data Drives (D)	Parity Drives (P)	Storage Usage Ratio
16	8	8	2.00
16	9	7	1.79
16	10	6	1.60
16	11	5	1.45
16	12	4	1.34
16	13	3	1.23
16	14	2	1.14

Классы хранения

```
export MINIO_STORAGE_CLASS_STANDARD=EC:3  
export MINIO_STORAGE_CLASS_RRS=EC:2
```

EECAMC²

Тестовый стенд

«For example, an 16-server distributed setup with 200 disks per node...»
(<https://docs.min.io/docs/distributed-minio-quickstart-guide.html>)



Выбор оборудования

MINIO MINIMUM REQUIREMENTS



Processor

Dual Intel® Xeon® Scalable Gold CPUs (minimum 8 cores per socket).



Drives

SATA/SAS HDDs for capacity and NVMe SSDs for high-performance (minimum of 8 drives per server).



Network

25GbE for capacity and 100GbE NICs for high-performance.



Memory

128GB RAM

Выбор оборудования

CPU: 8 ядер Intel Xeon E5-2650 v4

RAM: 16G

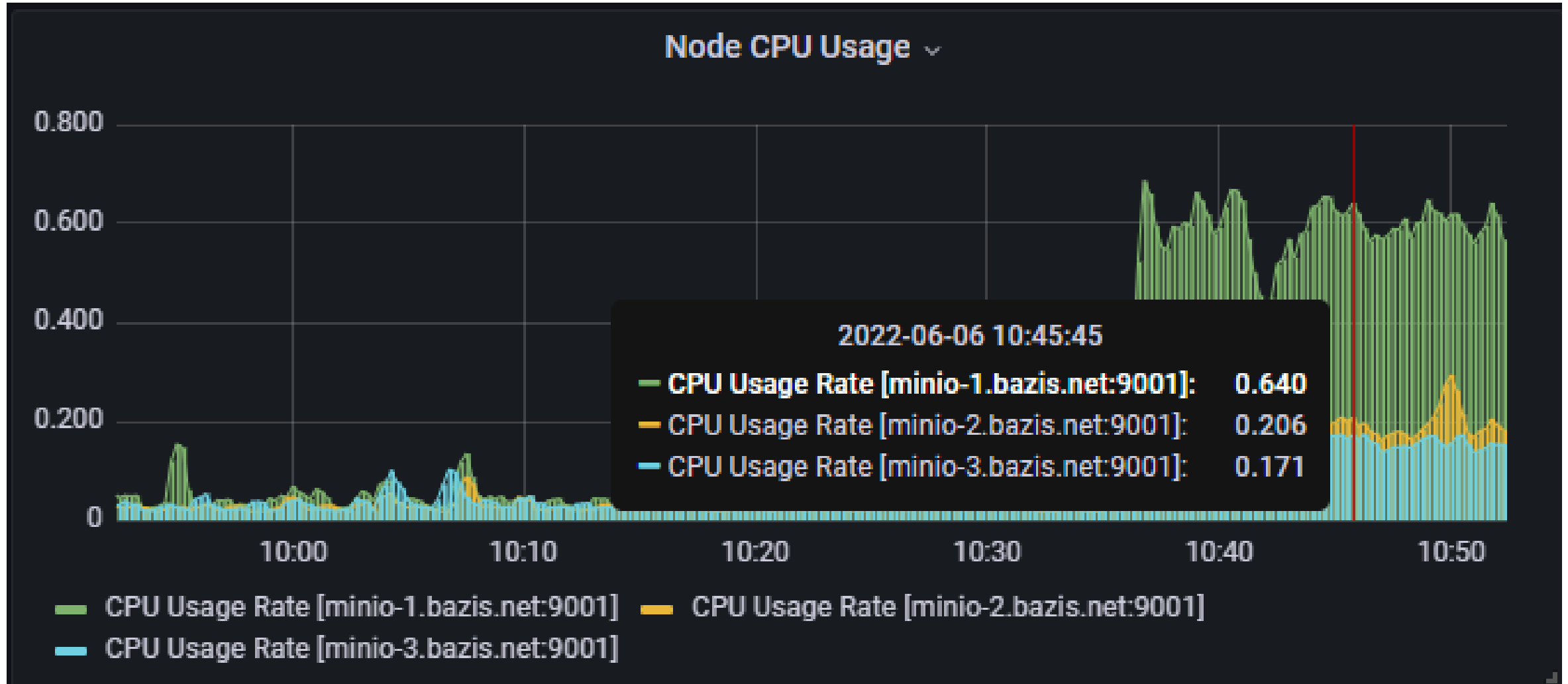
SSD: NVME Intel P4608i 6.4 TB

HDD: SATA Seagate Barracuda
ST5000LM003 5 TB

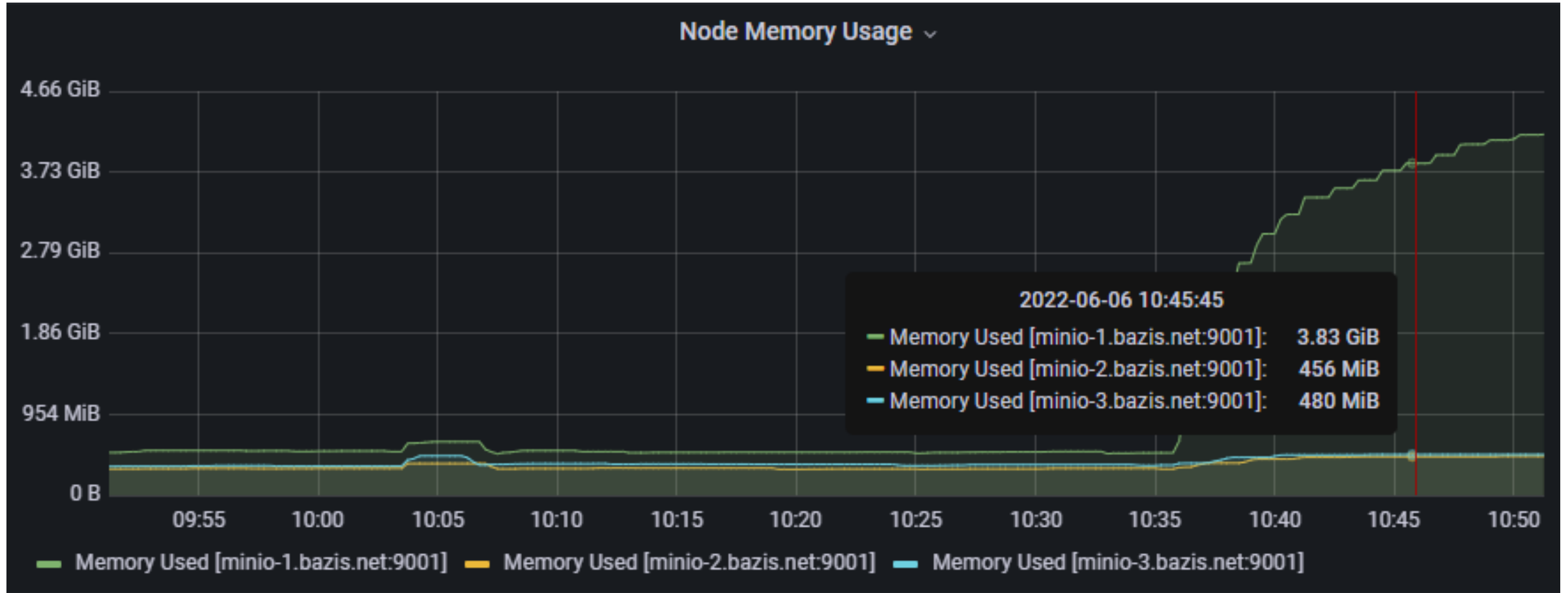
NODES: 3

root@minio-1: ~														
ATOP - minio-1 2022/06/02 11:12:09 ----- 10s elapsed														
PRC	sys	0.35s	user	0.41s	#proc	224	#tslpi	319	#tslpu	0	#zombie	0	#exit	4
CPU	sys	2%	user	3%	irq	0%	idle	782%	wait	14%	ipc notavail	curscal	2%	
cpu	sys	0%	user	0%	irq	0%	idle	96%	cpu004 w	3%	ipc notavail	curscal	2%	
cpu	sys	0%	user	0%	irq	0%	idle	99%	cpu001 w	1%	ipc notavail	curscal	2%	
cpu	sys	0%	user	1%	irq	0%	idle	96%	cpu006 w	3%	ipc notavail	curscal	2%	
cpu	sys	0%	user	0%	irq	0%	idle	96%	cpu000 w	3%	ipc notavail	curscal	2%	
cpu	sys	0%	user	0%	irq	0%	idle	98%	cpu002 w	2%	ipc notavail	curscal	2%	
cpu	sys	0%	user	0%	irq	0%	idle	99%	cpu005 w	0%	ipc notavail	curscal	2%	
cpu	sys	0%	user	0%	irq	0%	idle	97%	cpu007 w	2%	ipc notavail	curscal	2%	
cpu	sys	0%	user	0%	irq	0%	idle	99%	cpu003 w	0%	ipc notavail	curscal	2%	
CPL	avg1	0.47	avg5	0.46	avg15	0.51	cs	26133	intr	4152		numcpu	8	
MEM	tot	15.6G	free	1.6G	cache	3.9G	buff	13.3M	slab	6.0G	vmwal	0.0M	hptot	0.0M
SWP	tot	0.0M	free	0.0M							vmcom	8.8G	vmli	7.8G
PSI	cs	0/0/0	ms	0/0/0	mf	0/0/0	is	3/4/5	if	3/4/5				
LVM	dd--vg8-hdd8	busy	14%	read	7	write	46	MBr/s	0.0	MBw/s	0.2	avio	27.1 ms	
LVM	dd--vg9-hdd9	busy	14%	read	78	write	0	MBr/s	0.2	MBw/s	0.0	avio	17.3 ms	
LVM	dd--vg6-hdd6	busy	7%	read	41	write	0	MBr/s	0.2	MBw/s	0.0	avio	16.4 ms	
LVM	dd--vg7-hdd7	busy	5%	read	35	write	0	MBr/s	0.3	MBw/s	0.0	avio	13.1 ms	
LVM	ssd--vg-ssd1	busy	0%	read	10	write	0	MBr/s	0.0	MBw/s	0.0	avio	4.80 ms	
LVM	sd--vg2-ssd2	busy	0%	read	9	write	0	MBr/s	0.0	MBw/s	0.0	avio	4.89 ms	
LVM	g-ubuntu--lv	busy	0%	read	0	write	5	MBr/s	0.0	MBw/s	0.0	avio	0.80 ms	
DSK	sdh	busy	14%	read	7	write	48	MBr/s	0.0	MBw/s	0.1	avio	26.1 ms	
DSK	sdi	busy	14%	read	81	write	0	MBr/s	0.2	MBw/s	0.0	avio	16.7 ms	
DSK	sdf	busy	7%	read	43	write	0	MBr/s	0.2	MBw/s	0.0	avio	15.6 ms	
DSK	sdg	busy	5%	read	39	write	0	MBr/s	0.3	MBw/s	0.0	avio	11.8 ms	
DSK	sdb	busy	0%	read	10	write	0	MBr/s	0.0	MBw/s	0.0	avio	4.40 ms	
DSK	sd	busy	0%	read	9	write	0	MBr/s	0.0	MBw/s	0.0	avio	4.89 ms	
DSK	sda	busy	0%	read	0	write	3	MBr/s	0.0	MBw/s	0.0	avio	1.33 ms	
NET	transport	tcp	3489	tcpo	4619	udpi	0	udpo	0	tcpao	0	tcpo	3	
NET	network	ipi	3490	ipo	2796	ipfrw	0	deliv	3490	icmpi	0	icmpo	0	
NET	eth0	0%	pcki	3531	pcko	2797	sp	2000 Mbps	si	1531 Kbps	so	3483 Kbps	erro	0
PID	SYSCPU	USRCPU	VGROW	RGROW	RDDSK	WRDSK	RUID	EUID	ST	EXC	THR	S	CPUNR	CPU CMD
102350	0.20s	0.29s	OK	OK	6100K	1336K	minio	minio	--	-	139	S	1	5% minio
1258	0.07s	0.10s	OK	OK	96K	OK	minio	minio	--	-	40	S	6	2% minio
142672	0.03s	0.02s	920K	1080K	OK	OK	root	root	--	-	1	R	5	1% atop
10	0.01s	0.00s	OK	OK	OK	OK	root	root	--	-	1	I	4	0% rcu_sched
553	0.01s	0.00s	OK	OK	OK	OK	root	root	--	-	1	I	2	0% kworker/2:1H-x
1086	0.01s	0.00s	OK	OK	OK	OK	root	root	--	-	1	S	0	0% xfsaild/dm-8
142445	0.01s	0.00s	OK	OK	OK	OK	root	root	--	-	1	I	5	0% kworker/5:0-ev
142447	0.01s	0.00s	OK	OK	OK	OK	root	root	--	-	1	I	0	0% kworker/0:1-xf
142589	0.00s	0.00s	OK	OK	OK	OK	root	root	--	-	1	S	1	0% sshd
568	0.00s	0.00s	OK	OK	OK	OK	root	root	--	-	1	S	6	0% systemd-udev
1215	0.00s	0.00s	OK	OK	OK	OK	root	root	--	-	1	S	6	0% cron
69	0.00s	0.00s	OK	OK	OK	OK	root	root	--	-	1	S	4	0% khugepaged

Выбор оборудования - CPU



Выбор оборудования - RAM



Масштабирование

`http://minio-{1...16}.example.net/disk{1...200}`

`http://minio-{1...16}/disk{1...200} http://minio-{17...32} /disk{1...200}`

MinIO generally recommends planning capacity such that server pool expansion is only required after **2+ years** of deployment uptime.

Do **not** perform “rolling” (e.g. one node at a time) restarts.

Выбор архитектуры

РАСПРЕДЕЛЁННАЯ

- Больше объём хранения с тем же количеством дисков
- Не нужно включать версионирование
- Можно развернуть узлы в большем количестве локаций
- Объект гарантированно доступен на всех узлах сразу после загрузки
- Пользователи и политики применены сразу ко всему кластеру

РЕПЛИКАЦИЯ

- Выше скорость если каждый из кластеров размещён в своём ДЦ
- Лучше изолированность и отказоустойчивость

Проблемы при эксплуатации

Московская область, г. Коломна
2 SSD, 4 HDD

2 линка по 1000 Mbps
Средний пинг 14 мс.
Расстояние 750 км.

Санкт-Петербург
2 SSD, 4 HDD



SSD – EC: 3
HDD – EC: 4

2 линка по 500 Mbps
Средний пинг 19 мс.
Расстояние 600 км.

2 линка по 500 Mbps
Средний пинг 24 мс.
Расстояние 1200 км.



Казань
2 SSD, 4 HDD

Критические ошибки

#14224 - пропали все пользователи и политики

#13907 - неверно работало кэширование

#14381 - побились сжатые файлы

Created

Assigned

Mentioned

is:issue author:AlexZIX archived:false is:closed

2 Open

11 Closed

Visibility

Organization

Sort

minio/mc

Speedtest for network fnd objects doesn't work in distributed setup

community

triage

#4103 by AlexZIX was closed 19 hours ago

9

minio/minio

Unable to remove bucket

community

working as intended

#14957 by AlexZIX was closed 11 days ago

2

minio/minio

s2: corrupt input - compression

community

triage

#14381 by AlexZIX was closed on Feb 23

11

minio/minio

Lost users and IAM policies

community

fixed

#14224 by AlexZIX was closed on Feb 1

1

minio/minio

Very slow healing process

community

question

working as intended

#14198 by AlexZIX was closed on Jan 27

9

minio/mc

Wrong healing status

community

#3937 by AlexZIX was closed on Jan 26

4

minio/minio

Distributed instance degraded in case of 1 drive have issues

community

question

working as intended

#14174 by AlexZIX was closed on Jan 25

1

minio/console

Wrong button style

#1405 by AlexZIX was closed on Jan 17

1

minio/console

Undefined value in Access Rules

1

1

1

minio/minio

Can't upload file when caching enabled

community

1

12

minio/console

Total objects count doesn't fit its component layout

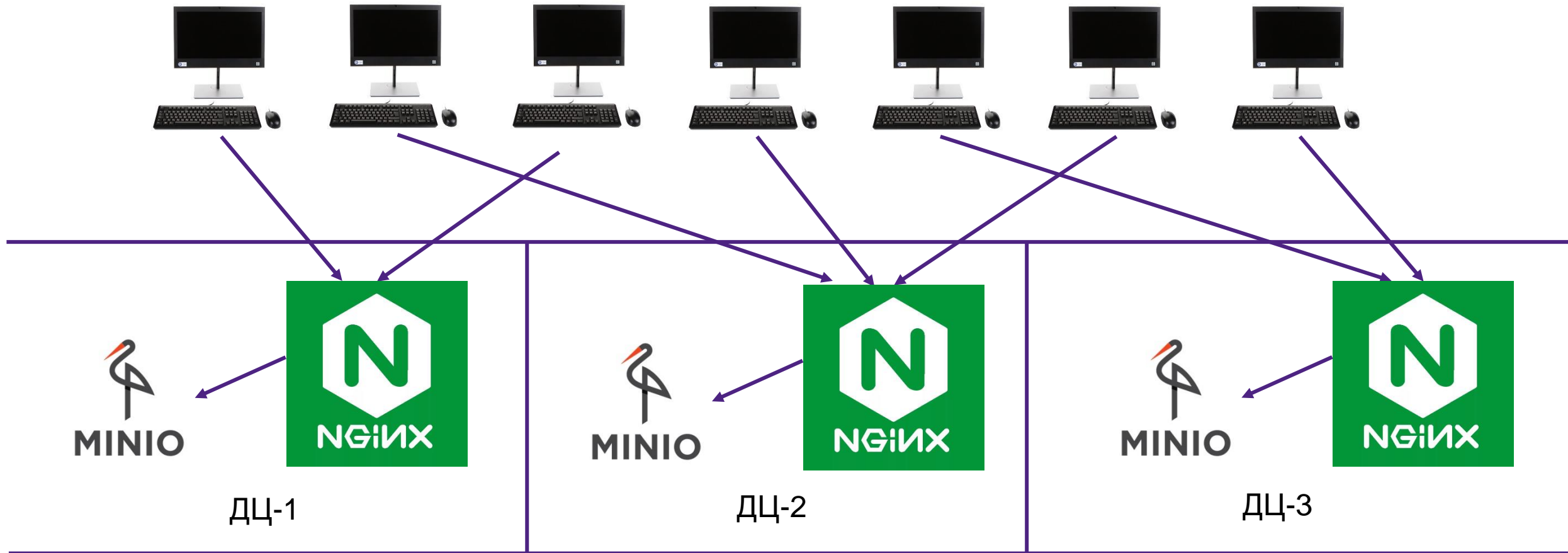
bug

UI

1

1

CDN на базе MinIO + NGINX



Критические ошибки

#14224 - пропали все пользователи и политики

#13907 - неверно работало кэширование

#14381 - побились сжатые файлы

Created

Assigned

Mentioned

is:issue author:AlexZIX archived:false is:closed

2 Open

11 Closed

Visibility

Organization

Sort

minio/mc

Speedtest for network fnd objects doesn't work in distributed setup

community

trriage

#4103 by AlexZIX was closed 19 hours ago

9

minio/minio

Unable to remove bucket

community

working as intended

#14957 by AlexZIX was closed 11 days ago

2

minio/minio

s2: corrupt input - compression

community

trriage

#14381 by AlexZIX was closed on Feb 23

11

minio/minio

Lost users and IAM policies

community

fixed

#14224 by AlexZIX was closed on Feb 1

1

minio/minio

Very slow healing process

community

question

working as intended

#14198 by AlexZIX was closed on Jan 27

9

minio/mc

Wrong healing status

community

#3937 by AlexZIX was closed on Jan 26

4

minio/minio

Distributed instance degraded in case of 1 drive have issues

community

question

working as intended

#14174 by AlexZIX was closed on Jan 25

1

minio/console

Wrong button style

#1405 by AlexZIX was closed on Jan 17

1

minio/console

Undefined value in Access Rules

#1403 by AlexZIX was closed on Jan 17

1

minio/minio

Can't upload file when caching enabled

community

#13907 by AlexZIX was closed on Dec 14, 2021

12

minio/console

Total objects count doesn't fit its component layout

bug

UI

#1322 by AlexZIX was closed on Dec 18, 2021

1

Восстановление после сбоев

- #3937 - неверное отображение хода восстановления
- #14198 - восстановление занимает оооочень долгое время
- #14174 - деградация кластера в случае выхода из строя одного из дисков
- Невозможность доступа к консоли управления в случае выхода из строя N дисков.

Created

Assigned

Mentioned

is:issue author:AlexZIX archived:false is:closed

2 Open

11 Closed

Visibility

Organization

Sort

minio/mc

Speedtest for network fnd objects doesn't work in distributed setup

community

trriage

#4103 by AlexZIX was closed 19 hours ago

9

minio/minio

Unable to remove bucket

community

working as intended

#14957 by AlexZIX was closed 11 days ago

2

minio/minio

s2: corrupt input - compression

community

trriage

#14381 by AlexZIX was closed on Feb 23

11

minio/minio

Lost users and IAM policies

community

fixed

#14224 by AlexZIX was closed on Feb 1

1

minio/minio

Very slow healing process

community

question

working as intended

#14198 by AlexZIX was closed on Jan 27

9

minio/mc

Wrong healing status

community

#3937 by AlexZIX was closed on Jan 26

4

minio/minio

Distributed instance degraded in case of 1 drive have issues

community

question

working as intended

#14174 by AlexZIX was closed on Jan 25

1

minio/console

Wrong button style

#1405 by AlexZIX was closed on Jan 17

1

minio/console

Undefined value in Access Rules

1

#1403 by AlexZIX was closed on Jan 17

1

minio/minio

Can't upload file when caching enabled

community

1

#13907 by AlexZIX was closed on Dec 14, 2021

12

minio/console

Total objects count doesn't fit its component layout

bug

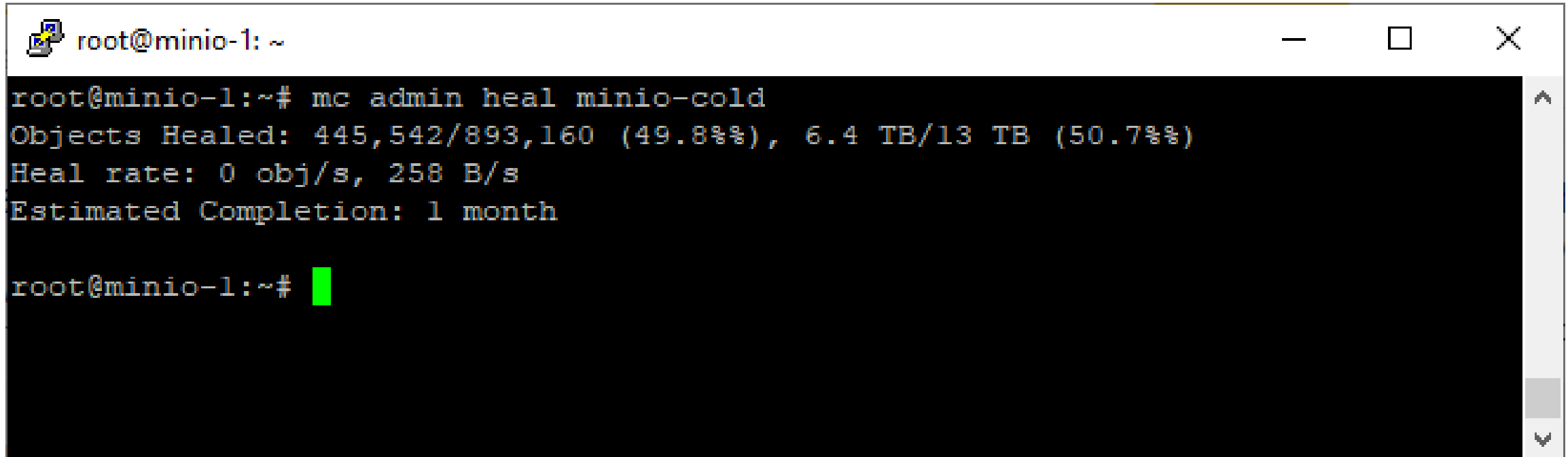
UI

1

#1322 by AlexZIX was closed on Dec 18, 2021

1

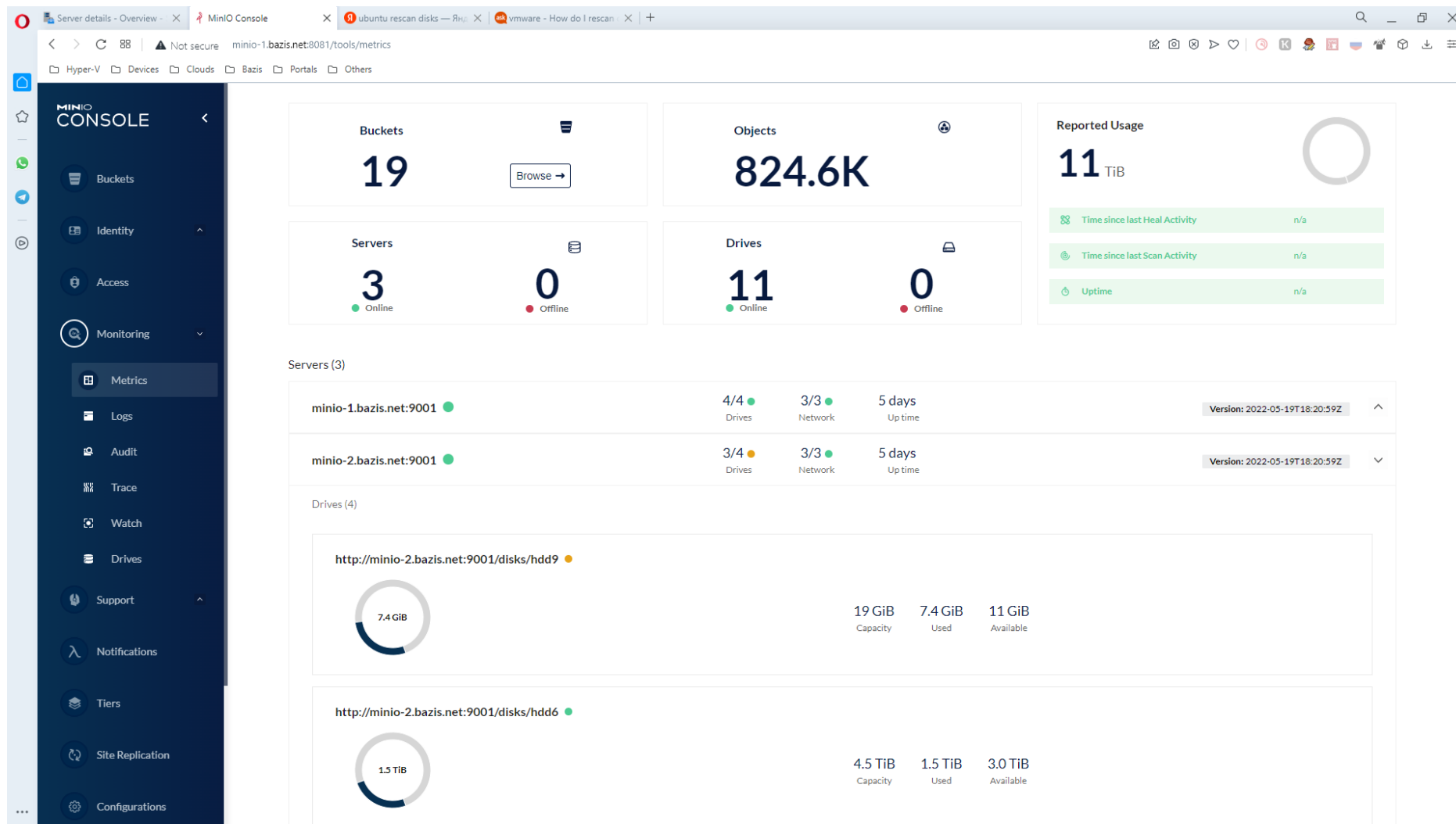
Реальный срок восстановления непрогнозируем

A terminal window titled 'root@minio-1: ~' with standard window controls (minimize, maximize, close). The terminal shows the execution of the 'mc admin heal minio-cold' command. The output indicates that 445,542 out of 893,160 objects have been healed, representing 49.8% of the total. The total capacity is 13 TB, with 6.4 TB currently used, representing 50.7% of the capacity. The heal rate is 0 objects per second and 258 bytes per second. The estimated completion time is 1 month. The prompt 'root@minio-1:~#' is followed by a red cursor.

```
root@minio-1:~# mc admin heal minio-cold
Objects Healed: 445,542/893,160 (49.8%%), 6.4 TB/13 TB (50.7%%)
Heal rate: 0 obj/s, 258 B/s
Estimated Completion: 1 month

root@minio-1:~#
```

Деградация производительности



Отсутствие мониторинга дисков

 AlexZIX added **community** **triage** labels on Jan 25



klauspost commented on Jan 25

Contributor ...

Unmount a partially failed drive:

For nodes with one or more drives that are either partially failed or operating in a degraded state (increasing disk errors, SMART warnings, timeouts in MinIO logs, etc.), you can safely unmount the drive if the cluster has sufficient remaining healthy drives to maintain read and write quorum. **Missing drives are less disruptive to the deployment than drives that are reliably producing read and write errors.**

<https://docs.min.io/minio/baremetal/installation/restore-minio.html>

MinIO поверх ZFS

```
root@minio-2:~# zpool list
```

NAME	SIZE	ALLOC	FREE	CKPOINT	EXPANDSZ	FRAG	CAP	DEDUP	HEALTH	ALTROOT
hdd6	4.55T	1.28T	3.26T	—	—	5%	28%	1.00x	ONLINE	—
hdd7	4.55T	1.28T	3.26T	—	—	5%	28%	1.00x	ONLINE	—
hdd8	4.48T	1.28T	3.20T	—	—	5%	28%	1.00x	ONLINE	—

Нет точных данных об используемом месте

The screenshot displays the MinIO Console interface for a cluster of three servers. The left sidebar contains navigation options: Buckets, Identity, Access, Monitoring, Metrics (selected), Logs, Audit, Trace, Watch, Drives, Support, License, Settings, and Documentation. The main content area shows the following data:

Server	Drives	Network	Up Time	Version
minio-1.bazis.net:9001	4/4	3/3	1 Month 1 Day	2022-03-17T06:34:49Z
Drives (4)				
http://minio-1.bazis.net:9001/disks/hdd6				
Capacity: 4.5 TiB	Used: 1.2 TiB	Available: 3.3 TiB		
http://minio-1.bazis.net:9001/disks/hdd7				
Capacity: 4.5 TiB	Used: 1.2 TiB	Available: 3.3 TiB		
http://minio-1.bazis.net:9001/disks/hdd8				
Capacity: 4.5 TiB	Used: 799 GiB	Available: 3.7 TiB		
http://minio-1.bazis.net:9001/disks/hdd9				
Capacity: 4.5 TiB	Used: 1.2 TiB	Available: 3.3 TiB		
minio-2.bazis.net:9001	4/4	3/3	4 Days	2022-03-17T06:34:49Z
minio-3.bazis.net:9001	4/4	3/3	2 Weeks	2022-03-17T06:34:49Z

И что в итоге

Вопросы

Алексей Плетнёв

Базис-Центр

zix@bазисsoft.ru

