

#### Alex Auvolat, Deuxfleurs Association

https://garagehq.deuxfleurs.fr/
Matrix channel: #garage:deuxfleurs.fr

Alex Auvolat, Deuxfleurs

Garage, the low-tech storage platform for geo-distributed clusters FOSDEM'24, 2024-02-03 1 / 40

### Who I am



Alex Auvolat PhD; co-founder of Deuxfleurs



#### Deuxfleurs

A non-profit self-hosting collective, member of the CHATONS network



### Our objective at Deuxfleurs

Promote self-hosting and small-scale hosting as an alternative to large cloud providers

Our objective at Deuxfleurs

# Promote self-hosting and small-scale hosting as an alternative to large cloud providers

Why is it hard?

Our objective at Deuxfleurs

#### Promote self-hosting and small-scale hosting as an alternative to large cloud providers

Why is it hard?

#### **Resilience**

we want good uptime/availability with low supervision

Commodity hardware (e.g. old desktop PCs)





Commodity hardware (e.g. old desktop PCs)

(can die at any time)

► Commodity hardware (e.g. old desktop PCs)

(can die at any time)

#### ▶ Regular Internet (e.g. FTTB, FTTH) and power grid connections

Commodity hardware (e.g. old desktop PCs)
 (can die at any time)

▶ Regular Internet (e.g. FTTB, FTTH) and power grid connections

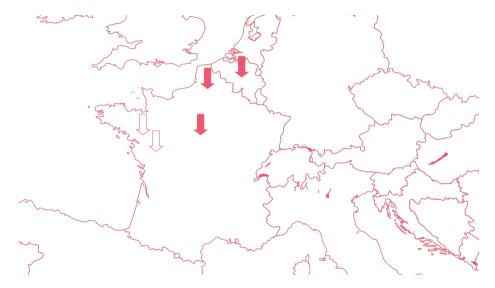
(can be unavailable randomly)

Commodity hardware (e.g. old desktop PCs)

(can die at any time)

 Regular Internet (e.g. FTTB, FTTH) and power grid connections (can be unavailable randomly)

**Geographical redundancy** (multi-site replication)



Object storage: a crucial component



S3: a de-facto standard, many compatible applications

Object storage: a crucial component



S3: a de-facto standard, many compatible applications

MinIO is self-hostable but not suited for geo-distributed deployments

## Object storage: a crucial component



S3: a de-facto standard, many compatible applications

MinIO is self-hostable but not suited for geo-distributed deployments

Garage is a self-hosted drop-in replacement for the Amazon S3 object store

Internally, Garage uses only CRDTs (conflict-free replicated data types)

Why not Raft, Paxos, ...? Issues of consensus algorithms:

Internally, Garage uses only CRDTs (conflict-free replicated data types)

Why not Raft, Paxos, ...? Issues of consensus algorithms:

#### Software complexity

Internally, Garage uses only CRDTs (conflict-free replicated data types)

Why not Raft, Paxos, ...? Issues of consensus algorithms:

Software complexity

Performance issues:

Internally, Garage uses only CRDTs (conflict-free replicated data types)

Why not Raft, Paxos, ...? Issues of consensus algorithms:

Software complexity

Performance issues:

► The leader is a **bottleneck** for all requests

Internally, Garage uses only CRDTs (conflict-free replicated data types)

Why not Raft, Paxos, ...? Issues of consensus algorithms:

Software complexity

Performance issues:

- ► The leader is a **bottleneck** for all requests
- Sensitive to higher latency between nodes

Internally, Garage uses only CRDTs (conflict-free replicated data types)

Why not Raft, Paxos, ...? Issues of consensus algorithms:

Software complexity

Performance issues:

- The leader is a **bottleneck** for all requests
- Sensitive to higher latency between nodes
- **Takes time to reconverge** when disrupted (e.g. node going down)

# The data model of object storage

Object storage is basically a **key-value store**:

Key: file path + name	Value: file data + metadata					
index.html	Content-Type: text/html; charset=utf-8					
	Content-Length: 24929					
	 binary blob>					
img/logo.svg	Content-Type: text/svg+xml					
	Content-Length: 13429					
	 binary blob>					
download/index.html	Content-Type: text/html; charset=utf-8					
	Content-Length: 26563					
	  binary blob>					

# The data model of object storage

Object storage is basically a key-value store:

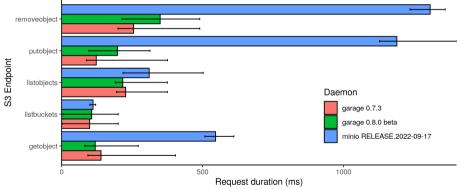
Key: file path + name	Value: file data + metadata					
index.html	Content-Type: text/html; charset=utf-8					
	Content-Length: 24929					
	 binary blob>					
img/logo.svg	Content-Type: text/svg+xml					
	Content-Length: 13429					
	 binary blob>					
download/index.html	Content-Type: text/html; charset=utf-8					
	Content-Length: 26563					
	 binary blob>					

▶ Maps well to CRDT data types

#### Performance gains in practice

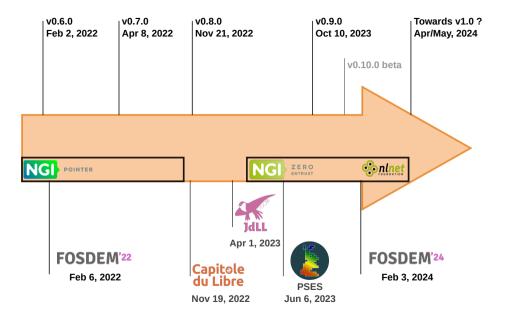
S3 endpoint latency in a simulated geo-distributed cluster

100 measurements, 5 nodes, 50ms RTT + 10ms jitter between nodes no contention: latency is due to intra-cluster communications colored bar = mean latency, error bar = min and max latency



Get the code to reproduce this graph at https://git.deuxfleurs.fr/Deuxfleurs/mknet

#### Recent developments



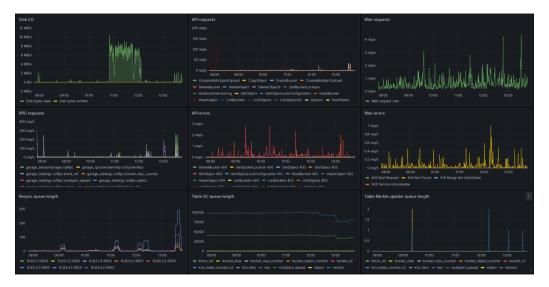
# April 2022 - Garage v0.7.0

Focus on observability and ecosystem integration

**Monitoring:** metrics and traces, using OpenTelemetry

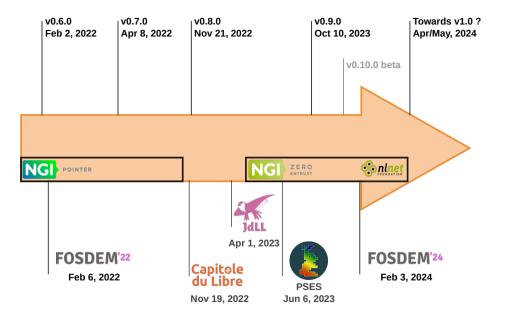
- ▶ Replication modes with 1 or 2 copies / weaker consistency
- ▶ Kubernetes integration for node discovery
- Admin API (v0.7.2)

# Metrics (Prometheus + Grafana)



# Traces (Jaeger)

JAEGER UI Search Compare S	ystem Architecture Monitor		Q Lo	okup by Trace ID	About Jaeger 🗸
← ✓ garage: S3 API ListOb	ojects 1f6c3ec	Find		◎ ^ ∨ × ¥	Trace Timeline v
Trace Start January 22 2024, 17:11:13.164   Dura	ation 119.75ms   Services 1   Depth 6	Total Spans 18			
Оµв	29.94ms	59.88ms		89.81ms	119.75ms
Service & Operation $\lor$ > $\lor$ »	0µs	29.94ms	59.88ms	89.81ms	119.75ms
✓ garage S3 API ListObjects					
✓ garage key get	102µs				
✓ garage RPC garage_table/table.rs/Rpc:ke	69µs				
garage RPC to 76797283/6c7e162	1 39µs				
v garage bucket_alias get	l 51µs				
✓ garage RPC garage_table/table.rs/Rpc:bu	I 40µs				
garage RPC to 76797283/6c7e162	l 20µs				
✓ garage bucket_v2 get	l 59µs				
V garage RPC garage_table/table.rs/Rpc:bu	l 34µs				
garage RPC to 76797283f8c7e162	l 17µs				
✓ garage object get_range					
✓ garage RPC garage_table/table.rs/Rpc:obj					114.2
✓ garage RPC to 967786691f20bb79	6.85ms				
✓ garage RPC >> garage_table/ta	<b>2.57ms</b>				
garage >> RPC garage_tabl	. 🔲 1.93ms				
✓ garage RPC to 3aed398eec82972b					114.2
✓ garage RPC >> garage_table/ta		42.92ms			
garage >> RPC garage_tabl		🔲 1.89ms			



November 2022 - Garage v0.8.0

Focus on performance

- Alternative metadata DB engines (LMDB, Sqlite)
- Performance improvements: block streaming, various optimizations...
- Bucket quotas (max size, max #objects)
- Quality of life improvements, observability, etc.

### About metadata DB engines

Issues with Sled:

- ► Huge files on disk
- Unpredictable performance, especially on HDD
- API limitations
- Not actively maintained

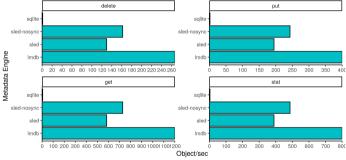
**LMDB:** very stable, good performance, file size is reasonable **Sqlite** also available as a second choice

Sled will be removed in Garage v1.0

# DB engine performance comparison

Comparison of Garage's metadata engines with "minio/warp"

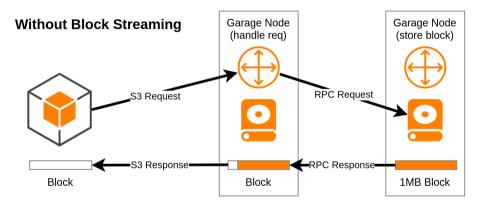
Daemon: Garage v0.8 no-fsync to avoid being impacted by block manager Benchmark: warp, mixed mode, 5min bench, 2568 objects, initialized with 200 objects. Environment mknet (Ryzen 5 1400, 1668 RAM, SSD). DC topo (3 nodes, 1Gb/s, 1ms latency).



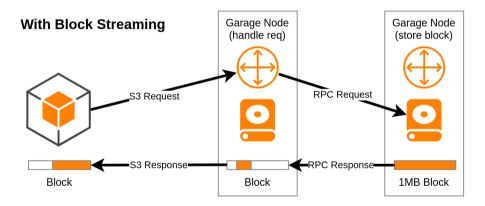
Get the code to reproduce this graph at https://git.deuxfleurs.fr/Deuxfleurs/mknet

NB: Sqlite was slow due to synchronous mode, now configurable

### Block streaming



#### Block streaming

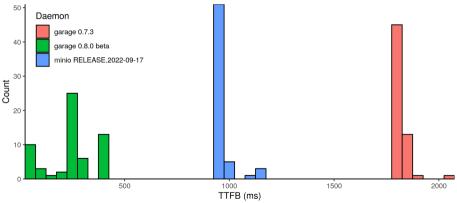


#### TTFB benchmark

TTFB (Time To First Byte) on GetObject over a slow network (5 Mbps, 500 µs)

A 1MB file is uploaded and then fetched 60 times.

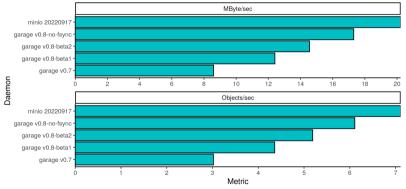
Except for Minio, the queried node does not store any data (gateway) to force net. communications.



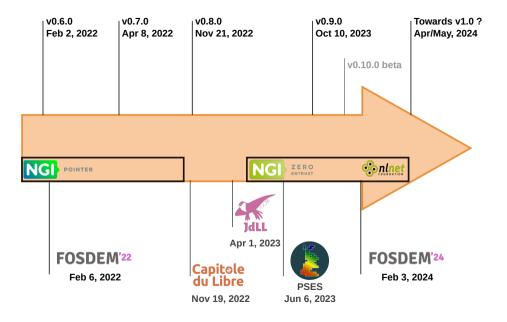
Get the code to reproduce this graph at https://git.deuxfleurs.fr/Deuxfleurs/mknet

# Throughput benchmark

"minio/warp" benchmark, "cluster total" result Ran on a local machine (Ryzen 5 1400, 16GB RAM, SSD) with mknet DC topology (3 nodes, 1GB/s, 1ms lat) warp in mixed mode, 5min bench, 5MB objects, initialized with 200 objects



Get the code to reproduce this graph at https://git.deuxfleurs.fr/Deuxfleurs/mknet



October 2023 - Garage v0.9.0

Focus on streamlining & usability

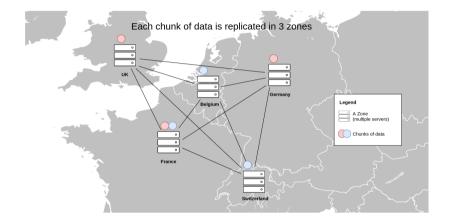
- Support multiple HDDs per node
- ► S3 compatibility:
  - support basic lifecycle configurations
  - allow for multipart upload part retries
- LMDB by default, deprecation of Sled
- New layout computation algorithm

#### Layout computation

[root@celeri:/home/lx]# docker exec -ti e338 /garage status							
==== HEALTHY NODE	S ====						
ID	Hostname	Address	Tags	Zone	Capacity		
5fcb3b6e39db3dcb	concombre	[2001:470:ca43::31]:3901	[concombre,neptune,france,alex]	neptune	500.0 GB		
942dd71ea95f4904	df-ymf	[2a02:a03f:6510:5102:6e4b:90ff:fe3a:6174]:3901	[df-ymf,bespin,belgium,max]	bespin	500.0 GB		
fdfaf7832d8359e0	df-ymk	[2a02:a03f:6510:5102:6e4b:90ff:fe3b:e939]:3901	[df-ymk,bespin,belgium,max]	bespin	500.0 GB		
0a03ab7c082ad929	ananas	[2a01:e0a:e4:2dd0::42]:3901	[ananas,scorpio,france,adrien]	scorpio	2.0 TB		
a717e5b618267806	courgette	[2001:470:ca43::32]:3901	<pre>[courgette,neptune,france,alex]</pre>	neptune	500.0 GB		
2032d0a37f249c4a	abricot	[2a01:e0a:e4:2dd0::41]:3901	[abricot,scopio,france,adrien]	scorpio	2.0 TB		
8cf284e7df17d0fd	celeri	[2001:470:ca43::33]:3901	[celeri,neptune,france,alex]	neptune	2.0 TB		
17ee03c6b81d9235	df-ykl	[2a02:a03f:6510:5102:6e4b:90ff:fe3b:e86c]:3901	[df-ykl,bespin,belgium,max]	bespin	500.0 GB		

Garage stores replicas on different zones when possible

#### Layout computation



Garage stores replicas on different zones when possible

## What a "layout" is

#### A layout is a precomputed index table:

Partition	Node 1	Node 2	Node 3
Partition 0	df-ymk (bespin)	Abricot (scorpio)	Courgette (neptune)
Partition 1	Ananas (scorpio)	Courgette (neptune)	df-ykl (bespin)
Partition 2	df-ymf (bespin)	Celeri (neptune)	Abricot (scorpio)
:	:	:	:
Partition 255	Concombre (neptune)	df-ykl (bespin)	Abricot (scorpio)

## What a "layout" is

#### A layout is a precomputed index table:

Partition	Node 1	Node 2	Node 3
Partition 0	df-ymk (bespin)	Abricot (scorpio)	Courgette (neptune)
Partition 1	Ananas (scorpio)	Courgette (neptune)	df-ykl (bespin)
Partition 2	df-ymf (bespin)	Celeri (neptune)	Abricot (scorpio)
:	:	:	:
Partition 255	Concombre (neptune)	df-ykl (bespin)	Abricot (scorpio)

The index table is built centrally using an optimal algorithm, then propagated to all nodes

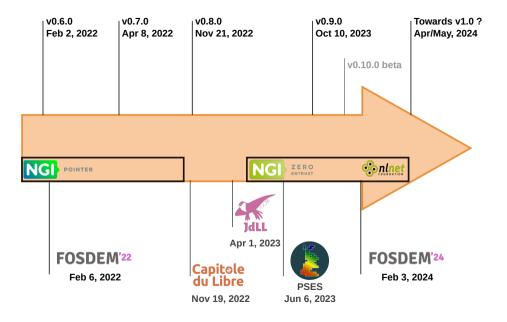
# What a "layout" is

#### A layout is a precomputed index table:

Partition	Node 1	Node 2	Node 3
Partition 0	df-ymk (bespin)	Abricot (scorpio)	Courgette (neptune)
Partition 1	Ananas (scorpio)	Courgette (neptune)	df-ykl (bespin)
Partition 2	df-ymf (bespin)	Celeri (neptune)	Abricot (scorpio)
:	:	:	:
Partition 255	Concombre (neptune)	df-ykl (bespin)	Abricot (scorpio)

The index table is built centrally using an optimal algorithm, then propagated to all nodes

Oulamara, M., & Auvolat, A. (2023). An algorithm for geo-distributed and redundant storage in Garage. arXiv preprint arXiv:2302.13798.



October 2023 - Garage v0.10.0 beta

Focus on consistency

▶ Fix consistency issues when reshuffling data

# Working with weak consistency

Not using consensus limits us to the following:

## Working with weak consistency

Not using consensus limits us to the following:

#### Conflict-free replicated data types (CRDT)

Non-transactional key-value stores such as S3 are equivalent to a simple CRDT: a map of **last-writer-wins registers** (each key is its own CRDT)

## Working with weak consistency

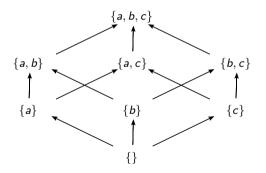
Not using consensus limits us to the following:

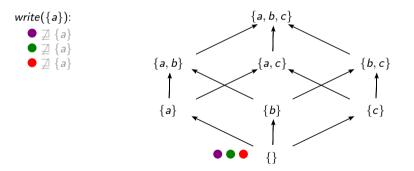
#### Conflict-free replicated data types (CRDT)

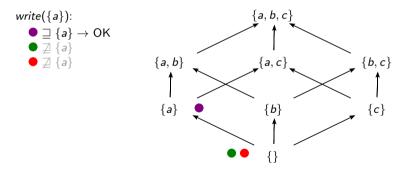
Non-transactional key-value stores such as S3 are equivalent to a simple CRDT: a map of **last-writer-wins registers** (each key is its own CRDT)

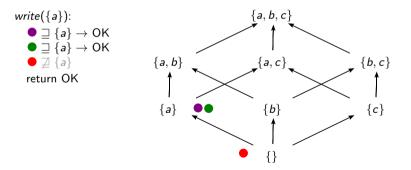
#### Read-after-write consistency

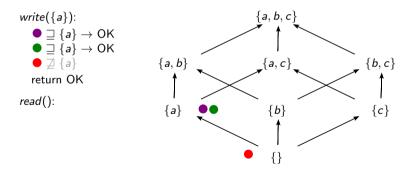
Can be implemented using quorums on read and write operations

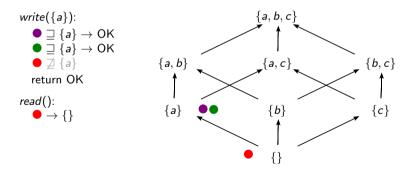


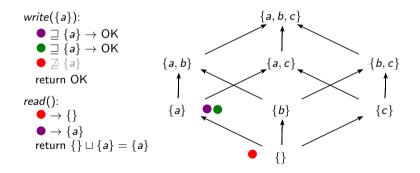




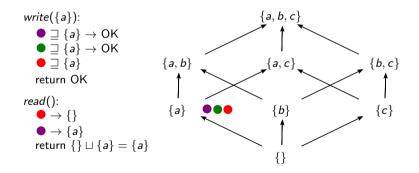








**Property:** If client 1 did an operation write(x) and received an OK response, and client 2 starts an operation read() after client 1 received OK, then client 2 will read a value  $x' \supseteq x$ .



**Property:** If client 1 did an operation write(x) and received an OK response, and client 2 starts an operation read() after client 1 received OK, then client 2 will read a value  $x' \supseteq x$ .

**Algorithm** write(x):

- 1. Broadcast write(x) to all nodes
- 2. Wait for k > n/2 nodes to reply OK
- 3. Return OK

**Algorithm** *read*():

- 1. Broadcast read() to all nodes
- Wait for k > n/2 nodes to reply with values x<sub>1</sub>,..., x<sub>k</sub>
- **3**. Return  $x_1 \sqcup \ldots \sqcup x_k$

# A hard problem: layout changes

▶ We rely on quorums k > n/2 within each partition:

$$n=3, k\geq 2$$

# A hard problem: layout changes

▶ We rely on quorums k > n/2 within each partition:

 $n=3, \qquad k\geq 2$ 

▶ When rebalancing, the set of nodes responsible for a partition can change:

Partition	Node 1	Node 2	Node 3
Partition 0	df-ymk	Abricot	Courgette
Partition 1	Ananas	Courgette	df-ykl
Partition 2	df-ymf	Celeri	Abricot

	Partition	Node 1	Node 2	Node 3
	Partition 0	Dahlia	Abricot	Eucalyptus
$\rightarrow$	Partition 1	Ananas	Euphorbe	Doradille
	Partition 2	Dahlia	Echinops	Abricot

# A hard problem: layout changes

▶ We rely on quorums k > n/2 within each partition:

 $n=3, \qquad k\geq 2$ 

▶ When rebalancing, the set of nodes responsible for a partition can change:

Partition	Node 1	Node 2	Node 3
Partition 0	df-ymk	Abricot	Courgette
Partition 1	Ananas	Courgette	df-ykl
Partition 2	df-ymf	Celeri	Abricot

	Partition	Node 1	Node 2	Node 3
	Partition 0	Dahlia	Abricot	Eucalyptus
$\rightarrow$	Partition 1	Ananas	Euphorbe	Doradille
E	Partition 2	Dahlia	Echinops	Abricot

During the rebalancing, new nodes don't yet have the data, and old nodes want to get rid of the data to free up space

 $\rightarrow$  risk of inconsistency, how to coordinate?

# Handling layout changes without losing consistency

#### Solution:

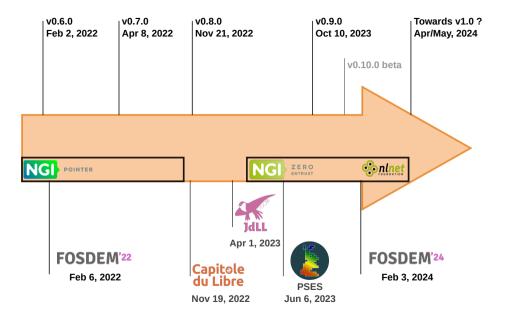
- keep track of data transfer to new nodes
- use multiple write quorums (new nodes + old nodes while data transfer is in progress)
- switching reads to new nodes only once copy is finished
- Implemented in v0.10
- Validated with Jepsen testing

	Asse	Time	NUMP				
	Datage legit dist talk?	202212171212.20	faller	anisht rds	<b>history</b> (a)	10000.200	-
	paraga anti dist tafat	20231310153341					
	parage regider tails ]	2023131111508.00	1754	posita ede	bidden tat	Adver. 300	-
	Denie Marine Mari	PRATA LITTLE AND	Carlot .	SCARL OR	SUBJECT OF	3300.00	
	parage repuised to be	100212111111	faine .	second a sub-	history of	Address inco	2
	mennes real due tella l	202233331113546.41	faine	posits etc.	bidden Lat	Appendix	-
	menor sel 2 der infe2	2003131711.06.34	faint	produceds	history (a)	Appendix.	-
	parage regil dija talah	20221217134434	Later .	produceds	kishary Tati	39441.340	**
	paraga anti dar tafat	2023-13-11115-04-51	faite	punks eds	Links	3944.30	-
	parage and dis later	200313 LITTLATO	Tarse .				
	parage manager takes	MALL LITTLE	Color.	second a rate			
	metada Josi dis telaj	2022-12-1012-01-14	1214	punks eds		Apren las	2
	parage replice tails.	2023131714.59.30	Loss .				
	parage art2 dpc tabe2	2022/12/11111-04-08	faite	an india seda	history Call	39441.340	**
				Brought and		Apres 200	2
	particle and 1 and 1 after	MACHINE CONTRACTOR NO.	faller .	and the sale			-
	manage logit car head	20221210145141	faine	pumba ede	hidante	Annual loc	z
	parage implicate balant	20031317143630	faint	produceds	history (a)	Appendix.	**
	parage art2 opciate2	20221213134564	faite	an india sala	kinang tat	3911.30	**
	BRADE DELL'OCARDO	20031311114.06.11	Faller	Nondra add	bashary car	*****.Tec	**
	BARNER AND THE LARKS		Carlos .				2
	menon and can being	20221212124.01	100	sunda ede	history (a)	Annual Sec	2
	parage and operated	2003333310144330	fains	scala.eds	bistory (at)	Appendix.	-
	parage regiliger taked	2012/12/1114414	fains	south a sele	history (a)	Address yes	**
	paraga anti opi talai	2012/12/1714/4141	Faller	produceds	kodury (at	*9++1.50	2
	BARBOR HOLL OP AMAL	PROPERTY AND ADDRESS OF ADDRESS O	Contract Inc.	Prosta and	tana si a		2
		20021210714-73.0	Talan .	second a solution			
	ments and can be a	200313121143747	faller	analysis and	MANTE	Annual lar	-
	ternes real or tellad	2012/13/11114:35:44	faine	punks.eds	bidden (a)	aquat.loc	÷
	manage and 2 spectation 2	200333337714344	Local Contract	peaks.eds	history(a)	Appendix.	-
		2012/12/1714.33.00					
		10231311111110	1000	Produced a	Social State	*****	-
	Sanata and an infail		Color.				
	mence and 2 pc toda 2						
		2012/13/1114/23/04	faine	positivado	hidoxtet	Appen.log	ń
	service and or being	MACHINE STATE OF ME	Colors.	and the sale			2
	marines not 2 pe indu 2						
remaining and a shall strategy and a shall be acting a much of a strategy and a shall be acting a shal	parage regil or falled						
sement production (Constraint)     sement prod	paraga anti perindat	2102-13-11114-22-41	false	punduada	hidan ter	Aprel 20	-
Anti-Antonio (1990)         Anti-Antonio (1990)         Anti-Antonio (1990)           Anti-Antonio (1990)         Anti-Antonio (1990) <td< td=""><td>earnes real or talk)</td><td>2002333310142140</td><td>false</td><td>pendo.ede</td><td>bidoste</td><td>About you</td><td>-</td></td<>	earnes real or talk)	2002333310142140	false	pendo.ede	bidoste	About you	-
Sector         Sector<		20231310142140	Calles .	secola ede	historytan	Address of the	2
and a second	parage and plants	100313 UTLAUS I	false				2
and entering Control (1997) and an example to the second s	mennes set2 pe inda3						
angenetational 2012.2012/2012/2013/2014 (even remained integrating parakaping parakap	Called + Lans marine	20231317161530	false	anada.eda	history (a)	agant. be	
Dessential and Statistical Statistical Network methods before impairing a specific set of the specific							
sense maj Linki. 2023.1.1111.0.1.20 fello modulosh biological papen.log isi popular jet	parage regit / Mini	2022-12-1014-15-49	false	punita.eda	hidostat	3000.307	-
property and a state of the sta	evenes web a table?	2023131014141	false .	scolts.eds	binoste	Appen.bo	2
paraje sveji z dola 2023-35.1711.4.22.11 Miler – posiskovila lastovjata posecijov stje naraje sveji z dola 2023-33.1711.6.26.11 Miler – posiskovila bolovjata posiskov do paraje sveji z dola 2023-33.1711.6.26.20 Miler – modkovila bolovjata pojesnikov do paraje sveji z dola 2023-33.1711.6.26.20 Miler – modkovila bolovjata pojesnikov do		encana all'hidd H	and	monds. Adds		Address of the	2
mengenetizetaki 2023.53.11114.08.13 falas mendianda kolenteta japannian da mengenetizetaki. 2023.53.11114.08.00 falas mendianda kolenteta japannian da mengenetizetati 2023.53.11114.08.10 falas mendianda kolenteta japannian da mengenetizetati 2023.53.11114.08.10 falas mendianda kolenteta japannian da mengenetizetati 2023.53.1114.08.10 falas mendianda kolenteta japannian da mengenetizetati 2023.53.1114.08.10 falas mendianda kolenteta japannian da mengenetizeta japan da mendianda kolenteta japan da mengenetizeta japan da mengenetizeta japan da mendianda kolenteta japan da mengenetizeta japan da mengenetizeta japan da mendianda kolenteta japan da mengenetizeta japan da mengenetizet							
parage mg2 / hdis2 2023 53 1271 6.56.00 film meads.eds historytai jepen.iko ap manon orf2 / tafa2 2023 53 1271 6.05.11 film meads.eds historytai jepen.iko ap	evence set2 r tofa3	2003333371140443	false	scalta ele	history (at)	Appendix	÷
parapa anti a tabar 2023-12-1111-0.00.12 Pelos mendia anti indensiar peneniny ap- parapa mela anti 2023-12-1111-0.00.12 Pelos mendia anti indensiar peneniny ap-	Called A Dama manage						
manon-nonlin-Mini 2023-13-1111-6-8-11 False pumbuloh hadaysan noosiloy mp	parage art2 x tafa2	2023-12-11114-09-12	false	pundta.eda	hidaytas	adenty of	49

Garage v0.9.0

Iepsen

Garage v0.10 beta

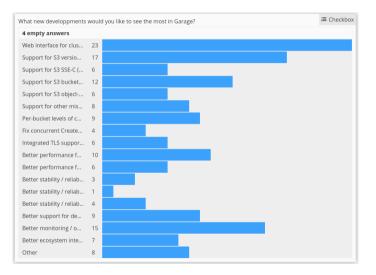


### Towards v1.0...

Focus on security & stability

- **Security audit** in progress by Radically Open Security
- ▶ Misc. S3 features (SSE-C, ...) and compatibility fixes
- ► Improve UX
- ► Fix bugs

# ...and beyond!



# Operating big Garage clusters

# **Operating Garage**

\$ garage status						
==== HEALTHY NODES						
ID	Hostname	Address	Tags	Zone	Capacity	DataAvail
ec5753c546756825	df-pw5	[2a02:a03f:6510:5102:223:24ff:feb0:e8a7]:3991	[df-pw5]	bespin	500.0 GB	429.1 GB (89.0%)
76797283f6c7e162	carcajou	[2001:470:ca43::22]:3991	[carcajou]	neptune	200.0 GB	166.3 GB (73.5%)
8073f25ffb7d6944	piranha	[2a01:cb05:911e:ec00:223:24ff:feb0:ea82]:3991	[piranha]	corrin	500.0 GB	457.3 GB (94.0%)
3aed398eec82972b	origan	[2a01:e0a:5e4:1d0:223:24ff:feaf:fdec]:3991	[origan]	jupiter	500.0 GB	457.1 GB (93.1%)
967786691f20bb79	caribou	[2001:470:ca43::23]:3991	[caribou]	neptune	500.0 GB	453.1 GB (92.3%)

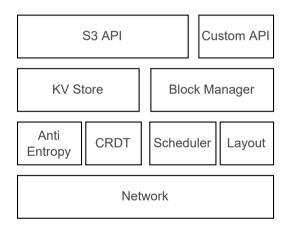
# **Operating Garage**

\$ garage status	c					
==== HEALTHY NODE ID	S ==== Hostname	Address	<b>T</b>	7	C	D-+-4
			Tags	Zone		DataAvail
ec5753c546756825	df-pw5	[2a02:a03f:6510:5102:223:24ff:feb0:e8a7]:3991	[df-pw5]	bespin	500.0 GB	429.1 GB (89.0%)
76797283f6c7e162	carcajou	[2001:470:ca43::22]:3991	[carcajou]	neptune	200.0 GB	166.3 GB (73.5%)
8073f25ffb7d6944	piranha	[2a01:cb05:911e:ec00:223:24ff:feb0:ea82]:3991	[piranha]	corrin	500.0 GB	457.3 GB (94.0%)
3aed398eec82972b	origan	[2a01:e0a:5e4:1d0:223:24ff:feaf:fdec]:3991	[origan]	jupiter	500.0 GB	457.1 GB (93.1%)
967786691f20bb79	caribou	[2001:470:ca43::23]:3991	[caribou]	neptune	500.0 GB	453.1 GB (92.3%)

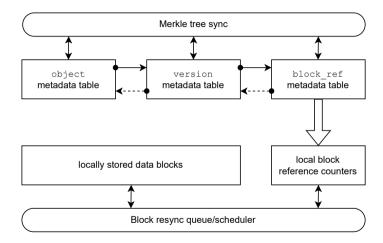
\$ garage status ==== HEALTHY NODES	5 ====					
ID	Hostname	Address	Tags	Zone	Capacity	DataAvail
76797283f6c7e162	carcajou	[2001:470:ca43::22]:3991	[carcajou]	neptune	200.0 GB	166.3 GB (73.5%)
8073f25ffb7d6944	piranha	[2a01:cb05:911e:ec00:223:24ff:feb0:ea82]:3991	[piranha]	corrin	500.0 GB	457.3 GB (94.0%)
3aed398eec82972b	origan	[2a01:e0a:5e4:1d0:223:24ff:feaf:fdec]:3991	[origan]	jupiter		457.1 GB (93.1%)
967786691f20bb79	caribou	[2001:470:ca43::23]:3991	[caribou]	neptune	500.0 GB	453.1 GB (92.3%)
==== FAILED NODES						
ID	Hostname	Address				st seen
ec5753c546756825	df-pw5	[2a02:a03f:6510:5102:223:24ff:feb0:e8a7]:3991	[df-pw5]	bespin 5	00.0 GB 5	minutes ago

# Garage's architecture

#### Garage as a set of components



# Garage's architecture



# Digging deeper

garage stats

Garage version: 20240116133343 [features: k2v, sled, lmdb, sqlite, consul-discoverv, kubernetes-discoverv, metrics, telemetrv-otlp, bundled-libs] Rust compiler version: 1.68.0 Database engine: LMDB (using Heed crate) Table stats: Ttems MklItems MklTodo GcTodo bucket v2 19 Θ 80964 block ref 334735 370927 Block manager stats: number of RC entries (~= number of blocks): 42376 resync queue length: 0 blocks with resync errors: 0 If values are missing above (marked as NC), consider adding the --detailed flag (this will be slow). Storage nodes: Capacity Part. DataAvail Hostname Zone MetaAvail ec5753c546756825 df-pw5 500.0 GB 175 429.1 GB/482.1 GB (89.0%) 429.1 GB/482.1 GB (89.0%) 76797283f6c7e162 carcajou neptune 200.0 GB 70 166.3 GB/226.2 GB (73.5%) 166.3 GB/226.2 GB (73.5%) 8073f25ffb7d6944 piranha 500.0 GB 173 457.3 GB/486.4 GB (94.0%) 457.3 GB/486.4 GB (94.0%) jupiter 500.0 GB 175 3aed398eec82972b origan 457.1 GB/490.7 GB (93.1%) 457.1 GB/490.7 GB (93.1%) 967786691f20bb79 caribou neptune 500.0 GB 175 453.1 GB/490.8 GB (92.3%) 453.1 GB/490.8 GB (92.3%) Estimated available storage space cluster-wide (might be lower in practice): data: 608.3 GB

metadata: 608.3 GB

# Digging deeper

\$ garage worker list									
TID	State	Name		Done	Queue		Consec		
1		Block resync worker #1							
2		Block resync worker #2							
3		Block resync worker #3							
4		Block resync worker #4							
5		Block resync worker #5							
6		Block resync worker #6							
7		Block resync worker #7							
8		Block resync worker #8							
9		Block scrub worker							
10		bucket v2 Merkle							
11		bucket v2 sync						17 hours ago	
12		bucket v2 GC							
13		bucket v2 queue							
14		bucket alias Merkle							
15		bucket alias sync						17 hours ago	
16		bucket alias GC							
17		bucket alias queue							
18		key Merkle							
19		key sync						17 hours ago	
20		key GC							
21		key queue							
22		object Merkle							
23		object sync						17 hours ago	
24		object GC							
25		object queue							
26		bucket_object_counter Merkle							
27		bucket_object_counter_sync						17 hours ago	
28		bucket_object_counter GC							
29		bucket_object_counter queue							
30		multipart_upload Merkle							
31		multipart_upload sync							
32	Idle	multipart_upload GC							
33		multipart_upload queue							
34	Idle	bucket_mpu_counter Merkle							
35		bucket_mpu_counter sync							
36		bucket_mpu_counter GC							
37									
38		version Merkle							
39								17 hours ago	
40									
41									
42									
43	Idle							17 hours ago	
44									
45									
46	Idle	object lifecycle worker	-	-	-	-	-		

# Digging deeper

\$ garage worker get									
completed	2024-01-23								
lity									
count	4								
ons_detected									
oleted	2023-12-27T13:49:33.234Z								
	2024-01-31T03:23:02.234Z								
ity	4								
\$ garage worker get -a resync-tranquility									
lity 1									
lity 1									
lity 1									
lity 1									
lity 1									
	Lity count ons_detected oleted Lity Lity 1 Lity 1 Lity 1 Lity 1								

# Potential limitations and bottlenecks

▶ Global:

▶ Max. ~100 nodes per cluster (excluding gateways)

Metadata:

One big bucket = bottleneck, object list on 3 nodes only

► Block manager:

- Lots of small files on disk
- Processing the resync queue can be slow

# Deployment advice for very large clusters

- Metadata storage:
  - ZFS mirror (x2) on fast NVMe
  - Use LMDB storage engine
- Data block storage:
  - Use Garage's native multi-HDD support
  - XFS on individual drives
  - Increase block size (1MB  $\rightarrow$  10MB, requires more RAM and good networking)
  - Tune resync-tranquility and resync-worker-count dynamically

Other :

- Split data over several buckets
- Use less than 100 storage nodes
- Use gateway nodes

Our deployments: < 10 TB. Some people have done more!

### Where to find us



https://garagehq.deuxfleurs.fr/
mailto:garagehq@deuxfleurs.fr
#garage:deuxfleurs.fr on Matrix

