

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 1/39

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

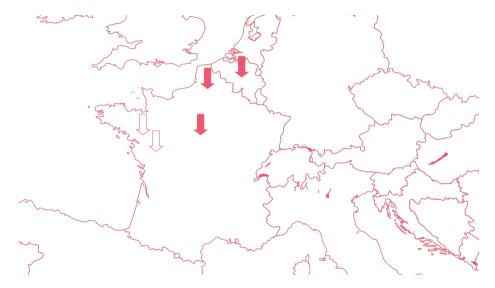
(can be unavailable randomly)

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

Consensus can be implemented reasonably well in practice, so why avoid it?

Software complexity

- Software complexity
- Performance issues:

- Software complexity
- Performance issues:
 - The leader is a **bottleneck** for all requests

- Software complexity
- Performance issues:
 - The leader is a **bottleneck** for all requests
 - Sensitive to higher latency between nodes

- 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)

Consensus can be implemented reasonably well in practice, so why avoid it?

- 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)

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

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:

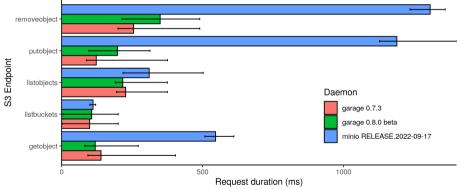
Key: file path + name	Value: file data + metadata					
index.html	Content-Type: text/html; charset=utf-8					
	Content-Length: 24929					
	 binary blob>					
img/logo.svg	Content-Type: text/svg+xml					
	Content-Length: 13429					
	 binary blob>					
download/index.html	Content-Type: text/html; charset=utf-8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

▶ Maps well to CRDT data types

Performance gains in practice

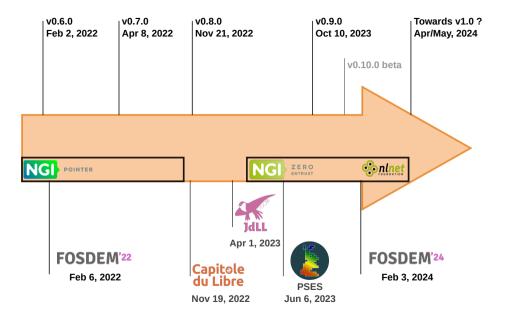
S3 endpoint latency in a simulated geo-distributed cluster

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

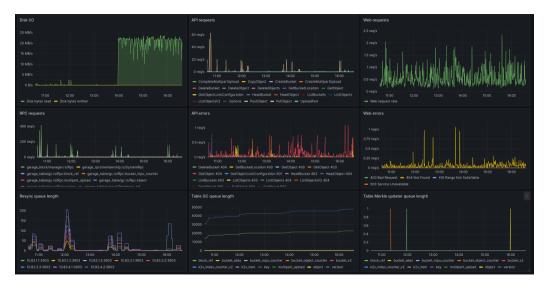


April 2022 - Garage v0.7.0

Focus on observability and ecosystem integration

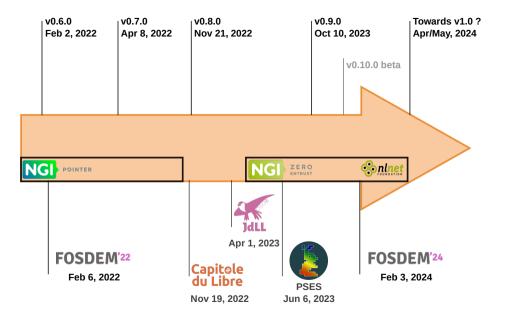
- **Monitoring:** metrics and traces, using OpenTelemetry
- Replication modes with 1 or 2 copies / weaker consistency
- Kubernetes integration
- Admin API (v0.7.2)
- Experimental K2V API (v0.7.2)

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	2.57ms				
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

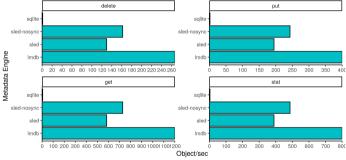
LMDB: very stable, good performance, reasonably small files on disk

Sled will be removed in Garage v1.0

DB engine performance comparison

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

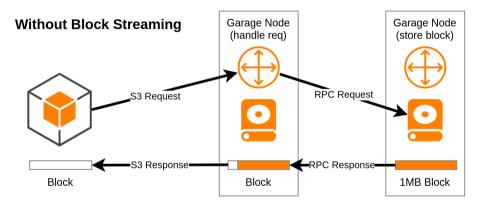
Daemon: Garage v0.8 no-fsync to avoid being impacted by block manager Benchmark: warp, mixed mode, 5min bench, 2568 objects, initialized with 200 objects. Environment mknet (Ryzen 5 1400, 1668 RAM, SSD). DC topo (3 nodes, 1Gb/s, 1ms latency).



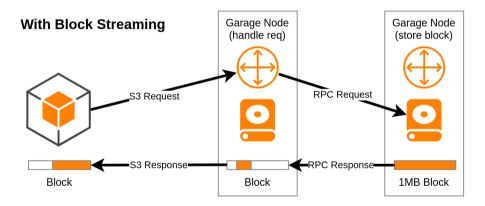
Get the code to reproduce this graph at https://git.deuxfleurs.fr/Deuxfleurs/mknet

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

Block streaming



Block streaming

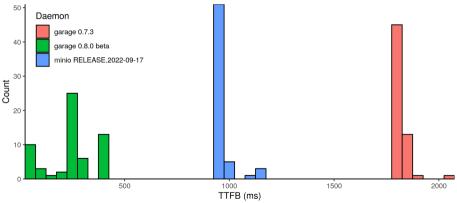


TTFB benchmark

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.

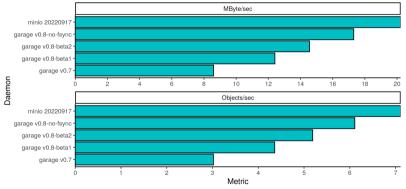
Except for Minio, the queried node does not store any data (gateway) to force net. communications.



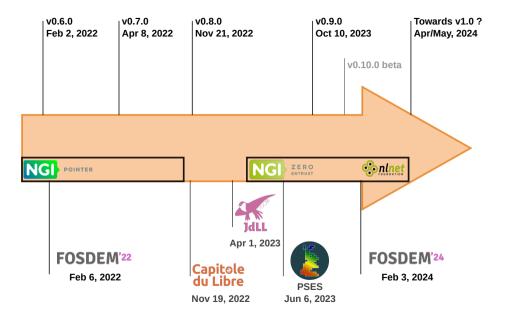
Get the code to reproduce this graph at https://git.deuxfleurs.fr/Deuxfleurs/mknet

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

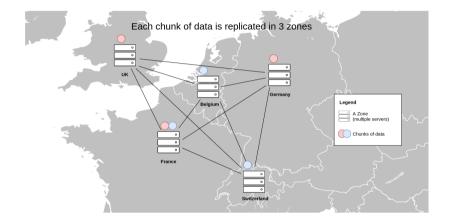
- Support multiple HDDs per node
- ► S3 compatibility:
 - support basic lifecycle configurations
 - allow for multipart upload part retries
- LMDB by default, deprecation of Sled
- New layout computation algorithm

Layout computation

<pre>[root@celeri:/home/lx]# docker exec -ti e338 /garage status</pre>						
==== 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	lo (jupiter)	Drosera (atuin)	Courgette (neptune)
Partition 1	Datura (atuin)	Courgette (neptune)	lo (jupiter)
Partition 2	lo(jupiter)	Celeri (neptune)	Drosera (atuin)
:	:	:	:
Partition 255	Concombre (neptune)	lo (jupiter)	Drosera (atuin)

What a "layout" is

A layout is a precomputed index table:

Partition	Node 1	Node 2	Node 3	
Partition 0	lo (jupiter)	Drosera (atuin)	Courgette (neptune)	
Partition 1	Datura (atuin)	Courgette (neptune)	lo (jupiter)	
Partition 2	lo(jupiter)	Celeri (neptune)	Drosera (atuin)	
:	:	:	:	
Partition 255	Concombre (neptune)	lo (jupiter)	Drosera (atuin)	

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

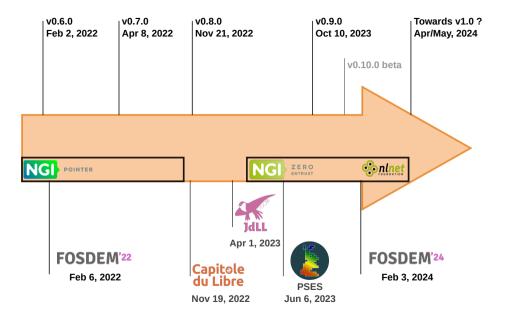
What a "layout" is

A layout is a precomputed index table:

Partition	Node 1	Node 2	Node 3	
Partition 0	lo (jupiter)	Drosera (atuin)	Courgette (neptune)	
Partition 1	Datura (atuin)	Courgette (neptune)	lo (jupiter)	
Partition 2	lo(jupiter)	Celeri (neptune)	Drosera (atuin)	
:	:	:	:	
Partition 255	Concombre (neptune)	lo (jupiter)	Drosera (atuin)	

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

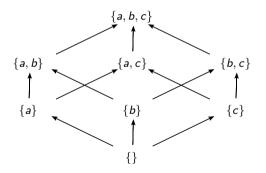
Not using consensus limits us to the following:

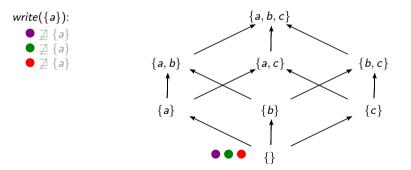
Conflict-free replicated data types (CRDT)

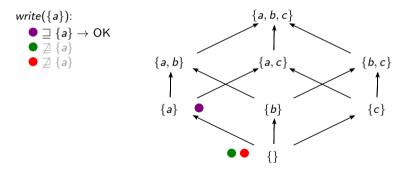
Non-transactional key-value stores such as S3 are equivalent to a simple CRDT: a map of **last-writer-wins registers** (each key is its own CRDT)

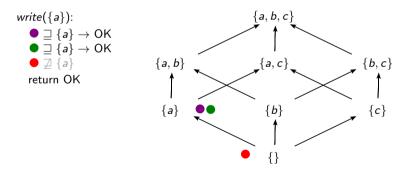
Read-after-write consistency

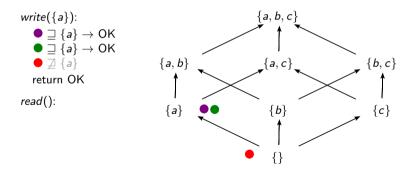
Can be implemented using quorums on read and write operations

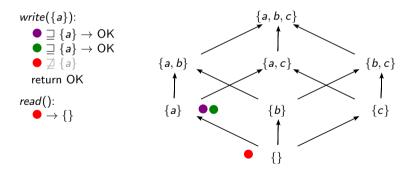


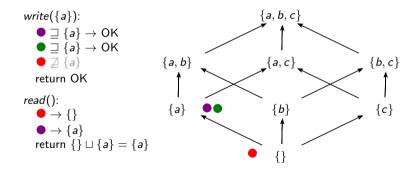


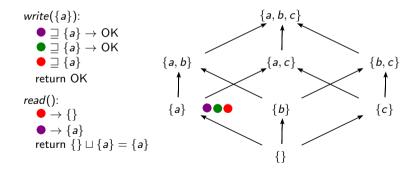












Property: If node A did an operation write(x) and received an OK response, and node B starts an operation read() after A received OK, then B will read a value $x' \supseteq x$.

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
- 2. Wait for k > n/2 nodes to reply with values x_1, \ldots, x_k
- **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:

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

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

$$\{A, B, C\} \to \{A, D, E\}$$

A hard problem: layout changes

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

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

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

$$\{A, B, C\} \to \{A, D, E\}$$

During the rebalancing, D and E don't yet have the data, and B and C 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:

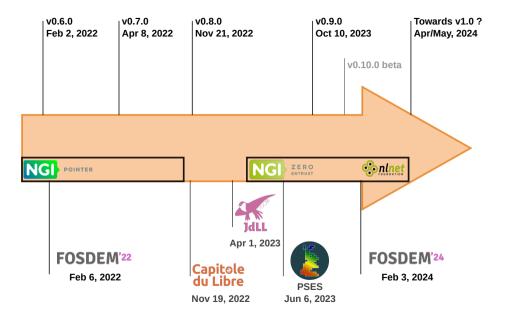
- keep track of data transfer to new nodes
- use multiple write quorums (new nodes + old nodes while data transfer is in progress)
- switching reads to new nodes only once copy is finished
- Implemented in v0.10
- Validated with Jepsen testing

Name means and doctated	Time	NUMP	Results of		Log	
menor real for talka	202212 10111-14 1	in the second se	succession and a	historylati	Annual lan	z
parage art2 dpt to fac2	20231310153341		punits eds		Advertision	44
menon nol dor telal menon nol dor telal	2023131111508.30	1754	posita eda posita eda	bidden tat	Adver. 300	-
menne selà der infaù parage ingli die tellali	20221211112044	faine .	produceds	history (a)	Advertised	
paraja ar2 dje tele2	10021210111100.0	faine .	punkt of	history of	Address inco	2
manage and 2 Apr indea	2023121711-0634	faint	mails etc.	history (a)	Appendix	÷
parage avg2 dys tafa2 parage avf2 dat tafa2			provides and a			
parage and disc infect means with the infect			prosite eds			
pience multiple tallal	200212101141414	Long Long	prinks rela			
parage and day todad	20031317134531	fairs	annalis ada	kintery fast	agent.240	-
parapa Joși dos tafai	202313101343.14	1764	punits eds			
menos nel de ndel	2022121111458.50	104	main etc.			
parage real day talkal parage ant2 day talkal	PRATA LITTLE N. J	Land .	produceds produceds	book and take	*****	-
manage and our Mini	200212 LITLASIC	100	punkt, eds		And and a local division of the local divisi	2
menes wi2 dor infe?	201233333111145629	faine				
			minits role	history Cali	Address Text	÷
parage and operated	20221213143434	fairs	annika ada	history fact	39441.349	**
paraga negl car tafai menan arti car tafai			pundta ada			
parage regilings halfed		failer .	secolds and			2
purson and cost tafa?	2003-13-17114-54-68	fains	annalis ada	kidon tat	aqual.bo	-
menors to \$2 cpc helps]	20231311114.46.11	faine	punits.eds	bidan tat	aquet.loc	-
menes artiz spectation		faine	produceds.	history (at	Appen.log	-
paraga mgi car bilai paraga anii car tafai	2022121714-0614	Lana	produceda	historytati	Address of the	
BARROW MALL OF LANS	100313 LITLAND	Colora	provide and a			2
parage regiliger talked	2012/12/1714 4141	fains	anisht rds	history (a)	Address (Sec)	-
parage and operating	2012/12/1714/4141	faite	annika eda	history Call	3911.30	**
marage real car tain) marage and car tain)	2003-13-11114-38-41	faise	punits.eds	biday (at	39445.200	-
manage and optically		Calles .			Appendiate	2
marnes real or talkal	2003131014354	faine	position and a	hiddreter	1000.00	2
manage and 2 spectation 2	2003.13.11114.35.44 2003.13.11114.34.44 2003.13.11114.34.44	Local Contract	produceds.	history(a)	Appendix.	-
paraga angi pa lalia) paraga anti pe talai	2012/12/1714/22/0	Entited	erada.eda	history(#	Address (principal)	*
menon real or talks)	2003-13-11114-31-04	Calles .	provide and a	history (a)	Apress Sec	2
maraon anti ar infait	20031317143311	false .	and the sale	history (a)		2
parage real or fulled						
pursps and pr todal	2012-12-12114-28-12	faine	punits.eds	hiday (a)	Appending	-
menes real or talkal	200333101423.00 20033310142344	faine	punks.eds	bidante	Adver-pop	-
parage and prinds] parage avoid on being		Calles .			Appendiate	2
mence art2 pr tofa2						
menon real or talk)						
manage and 2 periods 2	20231317162434	false	enader.ede	history (a)	pyson.log	
parage regil pr fulled	2012/12/1714/2241	Extine	produceds.	historytes	Address (principal)	**
paraga anti primfal marana mali primfal	20231310162168	Calles .	proside and a	baday (a)	Apres 100	2
marmore and 2 as indeal	20031317162140	false.	and the sale	history (a)	tion in	2
manues and 2 or 16 day	2012/12/1714 08:11	fains	produceds.	MANAGER	aquat.bo	-
mence well or tellad	202312-10114-08-11	faine	punks.eds	hidayte	Aquat.log	ŵ
menen setä peindaä	202212.1014.17.0	faise	scalta.eds	bidantati	Aquen.log	ń
Called a Cave marine Called a Checkman	2023-12 1014-15 M	Dillor .	modu.rds		pynni.loj	2
parage and a table parage and a table	20231310141549	faine				2
	200333310143410	false	souths ads	hidowise	agent.ber	÷
Collect & Correspondence	2023131714141410	false	enadereds.	historylati	jepson.log	
parage art2 x tafa2	202313131141333	faine	mahanda	hidaytat	39441.300	49
Datage average average	202313111141231	false .	punita.eda		Apren.log	-
menon set2 rists2 menon rec2 rists2	202313.11114.06.11	Colora	sconti. tels		Appendial log	2
mence art2 s telez	2022-12-1714-08-12	false	pundta.eda	MANAGER	agent.ho	-
menon rout / Mini	2022-12-1014-05-11	false .	punks.eds		Annual lines	4

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

Operating big Garage clusters

Operating Garage

\$ garage status						
==== HEALTHY NODE	S ====					
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%)

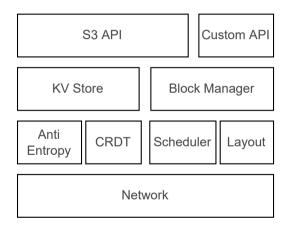
Operating Garage

\$ garage status	c					
==== HEALTHY NODE ID	S ==== Hostname	Address	T	7	C	
			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%)

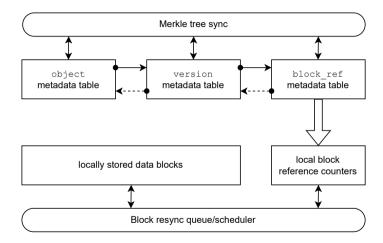
\$ garage status ==== HEALTHY NODE	S ====					
ID	Hostname	Address	Tags	Zone	Capacity	DataAvail
76797283f6c7e162	carcajou	[2001:470:ca43::22]:3991	[carcajou]	neptune	200.0 GB	166.3 GB (73.5%)
8073f25ffb7d6944	piranha	[2a01:cb05:911e:ec00:223:24ff:feb0:ea82]:3991	[piranha]	corrin	500.0 GB	457.3 GB (94.0%)
3aed398eec82972b	origan	[2a01:e0a:5e4:1d0:223:24ff:feaf:fdec]:3991	[origan]	jupiter	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%)
==== FAILED NODES						
ID	Hostname	Address				ist 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

\$ 08	S garage worker list							
TID	State	Name		Done	Queue		Consec	
1	Idle	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								
17	Idle	bucket_alias queue						
18	Idle	key Merkle						
19		key sync						
20	Idle	key GC						
21		key queue						
22	Idle	object Merkle						
23	Idle							17 hours ago
24		object GC						
25								
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		bucket_mpu_counter queue						
38		version Merkle						
39								17 hours ago
40		version GC						
41								
42								
43	Idle							17 hours ago
44								
45								
46	Idle	object lifecycle worker	-	-	-	-	-	

Digging deeper

-completed	2024-01-23
lity	
count	4
ons_detected	
oleted	2023-12-27T13:49:33.234Z
	2024-01-31T03:23:02.234Z
ity	4
anquility	
lity 1	
	lity count ons_detected oleted ity anquility lity 1 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

Current deployments: $\,<$ 10 TB, we don't have much experience with more

Where to find us



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

