Ceph Benchmark

Hyper-converged infrastructure with Proxmox VE virtualization platform and integrated Ceph Storage.

To optimize performance in hyper-converged deployments with Proxmox VE and Ceph storage the appropriate hardware setup can help a lot. This benchmark presents some possible setups and their performance outcomes with the intention to support Proxmox users to make better decisions.

EXECUTIVE SUMMARY

Hyper-converged setups with Proxmox VE can already be deployed on a minimum cluster setup of three nodes, enterprise class SATA SSDs, and with a 10 gigabit network. As long as there is enough CPU power and enough RAM, a decent performance of a three node cluster is possible.

- Since by default Ceph uses a replication of three, the data is still available even after losing one node, thus providing a highly available and distributed storage solution—fully software-defined and 100 % open-source.
- Although it is possible to run virtual machines/containers and Ceph on the same node, a separation does make sense in bigger workloads.
- To match your needs for growing workloads, the Proxmox VE and Ceph server clusters can be extended on the fly with additional nodes without any downtime.
- The Proxmox VE virtualization platform integrates Ceph storage since early 2014 with the release of Proxmox VE 3.2. Since then it has been used on thousands of servers worldwide, which provided an enormous amount of feedback and experience.

TABLE OF CONTENTS

TEST BED CONFIGURATION

All benchmark summarized in this paper has been completed in January and February 2018 on standard server hardware with a default Proxmox VE/Ceph server installation. The following section describes the testbed configuration.

SERVER HARDWARE

For benchmarking we used up to 6 identical servers with the below specifications:

CPU:	Single Intel® Xeon® E5-2620v4 2,1 GHZ 8/16 2133
Mainboard:	Supermicro X10SRi-F S2011-3
Case:	2U Supermicro Chassis 8x Hotswap
Dual 1 Gbit NIC:	Intel I350 (on board)
Dual 10 Gbit NIC:	Intel X550T
Dual 100 Gbit NIC:	Mellanox MCX456A-ECAT ConnectX-4, x16 PCIe 3.0
Memory:	4 x 16 GB DDR4 FSB2400 288-pin REG x4 1R

NETWORK SWITCHES

For the different Ceph networks, we used the following network switches:

1 GbE	Cisco SG300-28	MTU 9000	RJ45
10 GbE	Cicso SG350XG-2F10	MTU 9000	RJ45
100 GbE	Mellanox MSN2100-CB2F	MTU 9000	QSFP+ (DAC cable)

SOFTWARE VERSION (JAN/FEB 2018)

This benchmark has been completed with Proxmox VE 5.1, pve-kernel-4.13.13-5-pve and Ceph Version 12.2.2 (Luminous).

STORAGE/SSD FOR CEPH OSD

It's essential to use very reliable enterprise class SSDs with high endurance and power-loss protection. We also recommend to test the single SSD write performance with the Flexible I/O tester (fio) before you start using them as Ceph OSD devices.

The following table shows fio write results from a traditional spinning disk and from various selected SSDs:

		4K	
	Bandwidth (KB)	IO/s	Latency (ms)
Intel SSD DC P3700 Series 800 GB	300650	75162	0.01
Samsung SM863 240GB 2.5inch SSD	**69942**	**17485**	**0.06**
Intel DC S3510 120GB	50075	12518	0.08
Intel DC S3500 120GB	48398	12099	0.08
Samsung SSD 850 EVO 1TB	1359	339	2.94
Crucial MX100 512GB	1017	254	3.93
Seagate Constellation 7200.2 500 GB	471	117	8.47

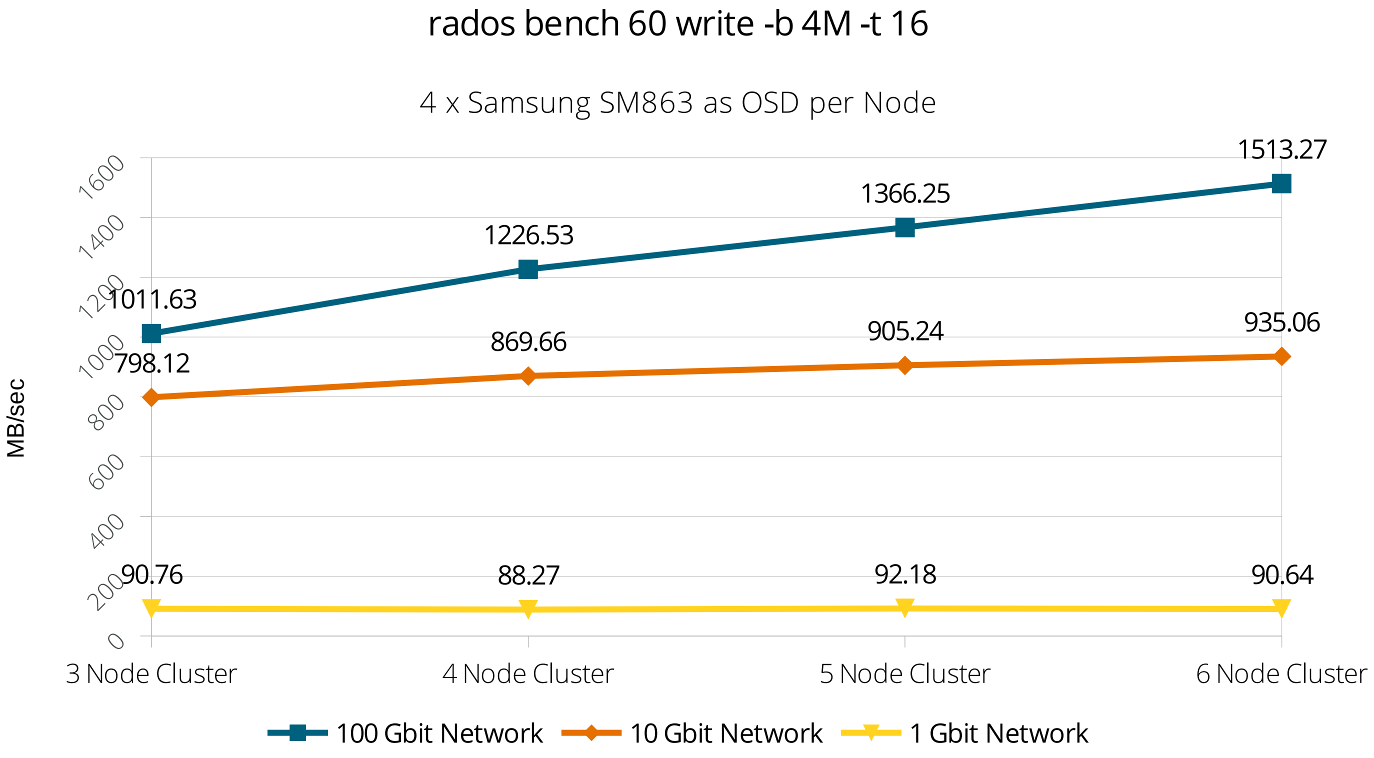
Based on these fio tests, we decided to use 24 x Samsung SM863 Series, 2.5", 240 GB SSD, SATA-3 (6 Gb/s) MLC. We connected 4 SSDs per server, using the on board SATA connectors.

We used the following fio command for the tests:

```
fio --ioengine=libaio –filename=/dev/sdx --direct=1 --sync=1 --rw=write --bs=4K
 --numjobs=1 --iodepth=1 --runtime=60 --time_based --group_reporting --name=fio
       --output-format=terse,json,normal --output=fio.log --bandwidth-log
```

Note: This command will destroy any data on your disk.

RADOS BENCH 60 WRITE -B 4M -T 16

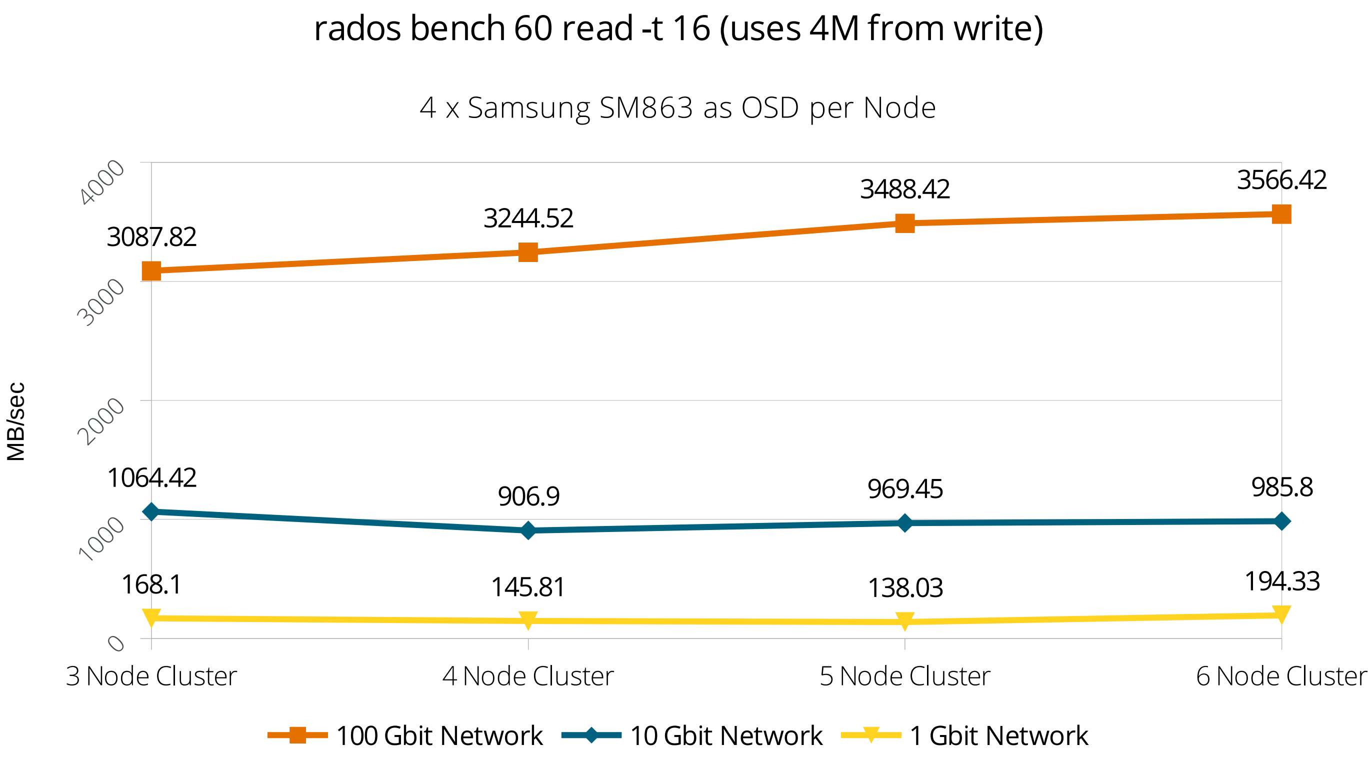


SUMMARY

The Rados write benchmark shows that the 1 Gbit network is a real bottleneck and is in fact too slow for Ceph. The 10 Gbit network already provides acceptable performance, but the clear winner is the 100 Gbit network with the highest speed and lowest latency.

Benchmarking has been done with four BlueStore OSDs per node (4 x Samsung SM863 240GB 2.5 SSD). The SM863 SSDs are currently also available in 1 TB and 2 TB.

RADOS BENCH 60 READ -T 16



SUMMARY

The Rados read benchmark displays that the 1 Gbit network is a real bottleneck and is in fact too slow for Ceph. The 10 Gbit network already provides acceptable performance. But the 100 Gbit network is the clear winner providing the highest speed and the lowest latency, especially visible in the read benchmark.

Benchmarking has been done with four BlueStore OSDs per node (4 x Samsung SM863 240GB 2.5 SSD).

HARDWARE FAQ

Can I use NVMe SSDs, for example M2 or PCI-Express cards?

Yes, this provides the highest disk performance.

Can I create a fast pool with NVMe SSDs, a semi fast pool with SSDs, and a slow pool with spinning disks?

Yes, building several pools can help in situations where budget is limited but big storage is needed.

Can I only use spinning disks in such a small setup (for example 3 nodes)?

No, the performance is very low.

Which CPUs should I prefer: More cores or a higher frequency?

CPUs with both, a lot of cores and a high frequency, are the best choice. This is true for Intel Xeon and AMD Epic CPUs.

How much RAM do I need per server?

Every OSD needs RAM for caches. Also VM workloads will need RAM. So the amount depends on the number of your OSDs and your VM workload. In most cases the best recommendation is getting as much RAM as possible.

Why do you test with 100 Gbit? Isn’t 10 Gbit enough?

The latency of the Ceph network has to be low, therefore the 100 Gbit technology is a good and future-proof choice.

Can I use a mesh network, so that I won’t need an expensive 10 or 100 Gbit switch?

Yes, a mesh network is ok if you have dual NICs and just 3 nodes. This is cost effective and fast.

Why did you use 240 GB SSDs for your tests?

As we did only run benchmark tests on these testlab servers, we didn’t need more space. For production setups, we recommend 1 TB or 2 TB SSDs.

Can I use consumer or pro-sumer SSDs, as these are much cheaper than enterprise class SSDs?

No. Never. These SSDs wont provide the needed performance, nor reliability and endurance. See the fio results from above and/or run your own fio tests.

Can I mix various disk types?

It is possible, but the cluster performance will drop to the performance of the slowest disk.

Can I mix different disk sizes?

No, it’s not recommended to use different disk sizes in small clusters because this will provoke unbalanced data distribution.

APPENDIX

1 - RADOS BENCHMARK ON 100 GBIT NETWORK

rados bench 60 write -b 4M -t 16 --no-cleanup				
	3x PVE server	4x PVE server	5x PVE server	6x PVE server
	4x OSD	4x OSD	4x OSD	4x OSD
Total time run	60,045639	60,022828	60,035793	60,034033
Total writes made	15186	18405	20506	22712
Write size	4194304	4194304	4194304	4194304
Object size	4194304	4194304	4194304	4194304
Bandwidth (MB/sec)	1011,63	1226,53	1366,25	1513,27
Stddev Bandwidth	34,6175	38,7777	39,5123	45,511
Max bandwidth (MB/sec)	1104	1296	1440	1644
Min bandwidth (MB/sec)	944	1124	1268	1416
Average IOPS	252	306	341	378
Stddev IOPS	8	9	9	11
Max IOPS	276	324	360	411
Min IOPS	236	281	317	354
Average Latency(s)	0,0632572	0,0521754	0,0468404	0,0422886
Stddev Latency(s)	0,0263501	0,0225026	0,0210645	0,0196891
Max latency(s)	0,158512	0,181573	0,265788	0,228123
Min latency(s)	0,013722	0,0137502	0,0135812	0,0136967

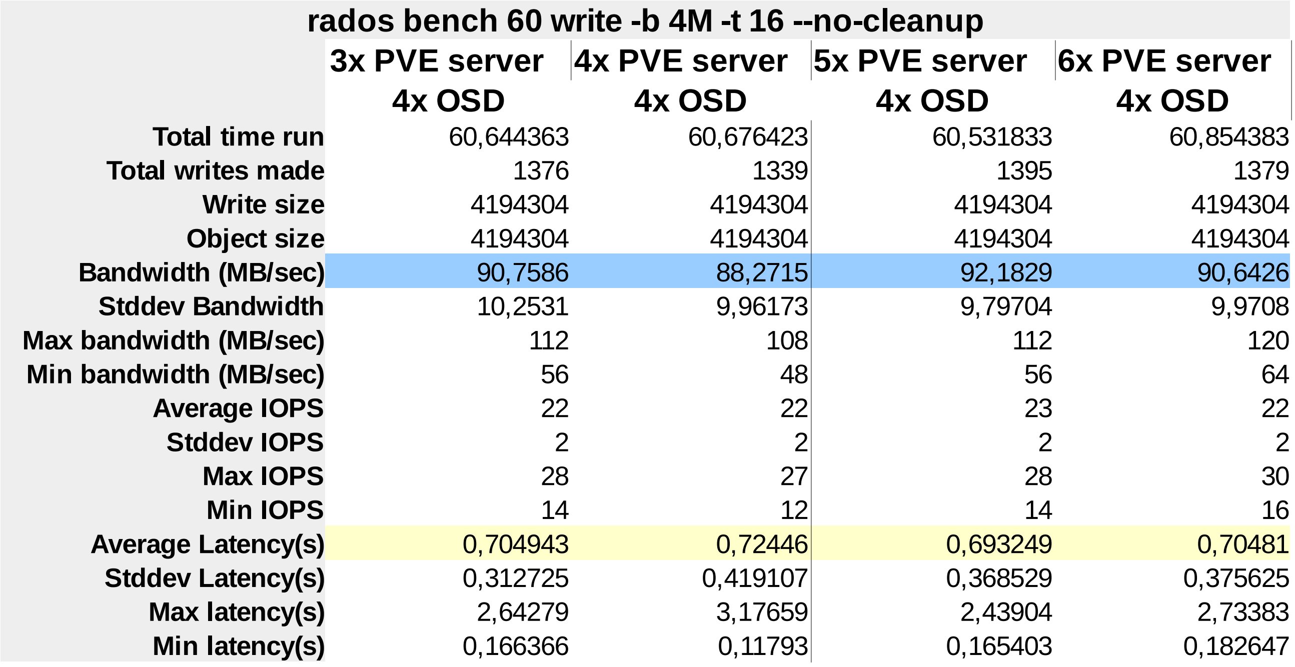
rados bench 60 read -t 16 (uses 4M from write)				
	3x PVE server	4x PVE server	5x PVE server	6x PVE server
	4x OSD	4x OSD	4x OSD	4x OSD
Total time run	19,69677	22,376184	23,336659	23,120112
Total reads made	15205	18150	20352	20614
Read size	4194304	4194304	4194304	4194304
Object size	4194304	4194304	4194304	4194304
Bandwidth (MB/sec)	3087,82	3244,52	3488,42	3566,42
Average IOPS	771	811	872	891
Stddev IOPS	15	17	20	58
Max IOPS	802	847	901	939
Min IOPS	743	778	791	658
Average Latency(s)	0,0199895	0,0189694	0,0176035	0,0171928
Max latency(s)	0,14009	0,128277	0,258353	0,812953
Min latency(s)	0,0110604	0,0111142	0,0112411	0,0108717

2 - RADOS BENCHMARK ON 10 GBIT NETWORK

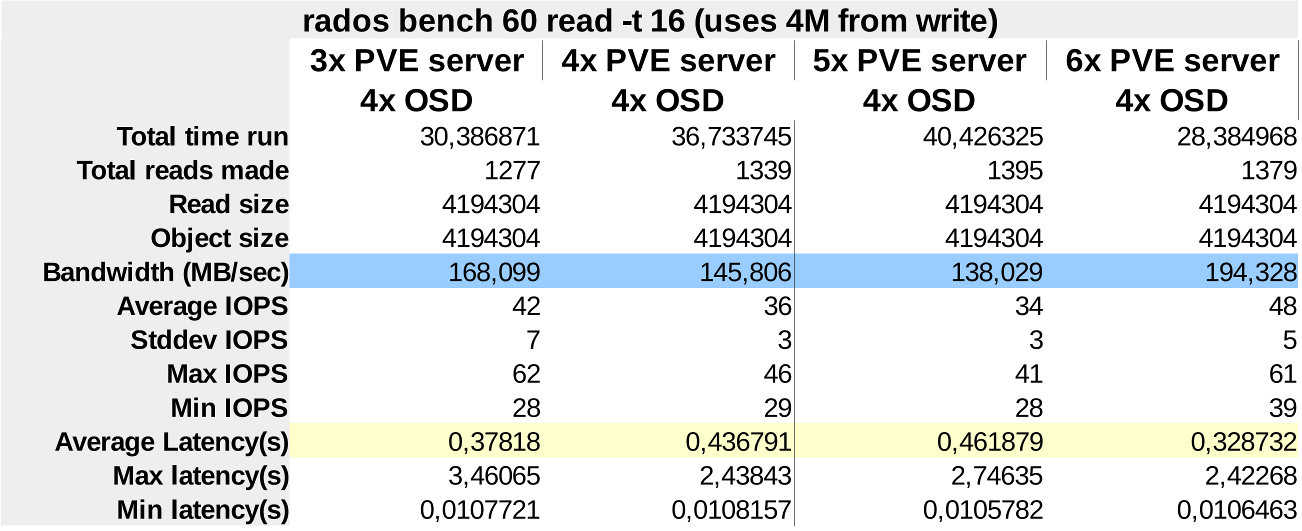
rados bench 60 write -b 4M -t 16 --no-cleanup				
	3x PVE server	4x PVE server	5x PVE server	6x PVE server
	4x OSD	4x OSD	4x OSD	4x OSD
Total time run	60,076578	60,202949	60,050489	60,060058
Total writes made	11987	13089	13590	14040
Write size	4194304	4194304	4194304	4194304
Object size	4194304	4194304	4194304	4194304
Bandwidth (MB/sec)	798,115	869,658	905,238	935,064
Stddev Bandwidth	25,5063	43,3892	22,8583	24,2148
Max bandwidth (MB/sec)	844	944	960	980
Min bandwidth (MB/sec)	732	624	860	848
Average IOPS	199	217	226	233
Stddev IOPS	6	10	5	6
Max IOPS	211	236	240	245
Min IOPS	183	156	215	212
Average Latency(s)	0,0801816	0,0735552	0,0706956	0,0684405
Stddev Latency(s)	0,0352137	0,0413322	0,0380508	0,035766
Max latency(s)	0,444184	0,74549	0,489463	0,404689
Min latency(s)	0,0205181	0,0187928	0,0193819	0,0192305

rados bench 60 read -t 16 (uses 4M from write)				
	3x PVE server	4x PVE server	5x PVE server	6x PVE server
	4x OSD	4x OSD	4x OSD	4x OSD
Total time run	45,046167	58,114332	56,073188	56,968754
Total reads made	11987	13176	13590	14040
Read size	4194304	4194304	4194304	4194304
Object size	4194304	4194304	4194304	4194304
Bandwidth (MB/sec)	1064,42	906,902	969,447	985,804
Average IOPS	266	226	242	246
Stddev IOPS	41	36	36	37
Max IOPS	352	298	324	320
Min IOPS	178	149	130	157
Average Latency(s)	0,0595262	0,069985	0,0654058	0,0643058
Max latency(s)	0,800953	1,16396	1,487	1,49632
Min latency(s)	0,0106962	0,0106668	0,0106488	0,0107538

3 - RADOS BENCHMARK ON 1 GBIT NETWORK



rados bench 60 write -b 4M -t 16 --no-cleanup				
	3x PVE server	4x PVE server	5x PVE server	6x PVE server
	4x OSD	4x OSD	4x OSD	4x OSD
Total time run	60,644363	60,676423	60,531833	60,854383
Total writes made	1376	1339	1395	1379
Write size	4194304	4194304	4194304	4194304
Object size	4194304	4194304	4194304	4194304
Bandwidth (MB/sec)	90,7586	88,2715	92,1829	90,6426
Stddev Bandwidth	10,2531	9,96173	9,79704	9,9708
Max bandwidth (MB/sec)	112	108	112	120
Min bandwidth (MB/sec)	56	48	56	64
Average IOPS	22	22	23	22
Stddev IOPS	2	2	2	2
Max IOPS	28	27	28	30
Min IOPS	14	12	14	16
Average Latency(s)	0,704943	0,72446	0,693249	0,70481
Stddev Latency(s)	0,312725	0,419107	0,368529	0,375625
Max latency(s)	2,64279	3,17659	2,43904	2,73383
Min latency(s)	0,166366	0,11793	0,165403	0,182647



rados bench 60 read -t 16 (uses 4M from write)				
	3x PVE server	4x PVE server	5x PVE server	6x PVE server
	4x OSD	4x OSD	4x OSD	4x OSD
Total time run	30,386871	36,733745	40,426325	28,384968
Total reads made	1277	1339	1395	1379
Read size	4194304	4194304	4194304	4194304
Object size	4194304	4194304	4194304	4194304
Bandwidth (MB/sec)	168,099	145,806	138,029	194,328
Average IOPS	42	36	34	48
Stddev IOPS	7	3	3	5
Max IOPS	62	46	41	61
Min IOPS	28	29	28	39
Average Latency(s)	0,37818	0,436791	0,461879	0,328732
Max latency(s)	3,46065	2,43843	2,74635	2,42268
Min latency(s)	0,0107721	0,0108157	0,0105782	0,0106463

LEARN MORE

Wiki: https://pve.proxmox.com

Community Forums: https://forum.proxmox.com

Bugtracker: https://bugzilla.proxmox.com

Code repository: https://git.proxmox.com

HOW TO BUY

Find an authorised reseller in your area: www.proxmox.com/partners or

Visit the Proxmox Online Shop to purchase a subscription: https://shop.maurer-it.com

SALES AND INQUIRIES

https://www.proxmox.com

Proxmox Customer Portal

https://my.proxmox.com

TRAINING PROXMOX VE

Learn Proxmox VE easily, visit https://www.proxmox.com/training

ABOUT PROXMOX

Proxmox Server Solutions GmbH is a privately held company based in Vienna, Europe.

Proxmox Server Solutions GmbH
Bräuhausgasse 37
1050 Vienna
Austria

office@proxmox.com
https://www.proxmox.com