Astra DB PCU performance benchmarks

Throughput results for Small, Medium, and Standard PCUs on AWS, Azure, and GCP

Introduction

This document presents our most recent benchmarking of Astra DB Small, Medium, and Standard PCUs. Testing of additional workload types is ongoing, and we will update this document as those results become available.

Because every workload is distinct, performance in your production environment may differ from the results shown here, influenced by factors such as data model, payload size, read/write mix, data volume, and access patterns. These results are provided for comparative purposes only and do not constitute a guarantee. Their purpose is to give you an informed starting point for selecting the PCU type best suited to your expected workload**.**

Please note that Datasets that exceed available cache, may cause added latency.

Find your workload

Select a workload to jump to its results.

Workload	Payload profile	Read/write mix	Clouds
Key-value (KV)	4K/1K	20% reads / 80% writes	AWS, Azure, GCP
Key-value (KV)	4K/1K	60% reads / 40% writes	AWS, Azure, GCP
Key-value (KV)	4K/1K	80% reads / 20% writes	AWS, Azure, GCP
IoT	4K/1K	20% reads / 80% writes	AWS, Azure, GCP
IoT	4K/1K	60% reads / 40% writes	AWS, Azure, GCP
IoT	4K/1K	80% reads / 20% writes	AWS, Azure, GCP

Results

Key-value (KV), 4K/1K payloads

20% reads / 80% writes

Cloud	PCU size	Reads (ops/s)		Writes (ops/s)	
		p99 < 10 ms	p99 < 25 ms	p99 < 10 ms	p99 < 25 ms
AWS	Small	900	1,500	5,200	7,600
	Medium	2,700	3,300	13,200	18,000
	Standard	3,500	5,900	≥ 23,600	≥ 23,600
Azure	Small	1,100	1,500	8,400	8,400
	Medium	1,700	2,900	16,400	18,800
	Standard	≥ 5,900	≥ 5,900	≥ 23,600	≥ 23,600
GCP	Small	1,100	1,500	6,000	9,200

Cloud	PCU size	Reads (ops/s)		Writes (ops/s)	
		p99 < 10 ms	p99 < 25 ms	p99 < 10 ms	p99 < 25 ms
	Medium	3,500	5,500	20,400	≥ 23,650
	Standard	5,500	≥ 5,900	≥ 23,600	≥ 23,600

Standard reached the highest write load tested on all three clouds. Medium sustained 18,000 to 23,650+ writes/s at 25 ms.

60% reads / 40% writes

Cloud	PCU size	Reads (ops/s)		Writes (ops/s)	
		p99 < 10 ms	p99 < 25 ms	p99 < 10 ms	p99 < 25 ms
AWS	Small	2,100	2,100	3,000	3,800
	Medium	14,100	≥ 16,600	7,800	≥ 11,400
	Standard	≥ 17,700	≥ 17,700	≥ 11,800	≥ 11,800
Azure	Small	1,500	4,500	4,600	≥ 6,120
	Medium	3,300	6,900	≥ 11,830	≥ 11,830
	Standard	14,100	≥ 17,700	≥ 11,820	≥ 11,820
GCP	Small	2,100	3,900	2,600	≥ 4,675
	Medium	7,500	9,900	9,370	≥ 10,950
	Standard	10,500	14,700	≥ 11,800	≥ 11,800

Medium read throughput at 25 ms varied by cloud, from 6,900 to 16,600+ ops/s. Small sustained 2,100 to 4,500 reads/s.

80% reads / 20% writes

Cloud	PCU size	Reads (ops/s)		Writes (ops/s)	
		p99 < 10 ms	p99 < 25 ms	p99 < 10 ms	p99 < 25 ms
AWS	Small	2,800	5,200	1,500	2,070
	Medium	9,200	11,600	4,300	≥ 5,125
	Standard	14,800	≥ 23,600	≥ 5,600	≥ 5,600
Azure	Small	2,000	5,200	2,500	≥ 2,670
	Medium	2,800	9,200	≥ 5,310	≥ 5,310
	Standard	21,200	≥ 23,600	≥ 5,900	≥ 5,900
GCP	Small	3,600	5,200	1,500	≥ 2,060
	Medium	10,000	13,200	≥ 4,410	≥ 4,410
	Standard	14,000	17,200	≥ 5,910	≥ 5,910

At 25 ms, Medium sustained 9,200 to 13,200 reads/s, compared with 17,200 to 23,600+ for Standard.

IoT, 4K/1K payloads

This workload simulates the access patterns of a time-series IoT application. Each read accesses approximately 4 KB of data, and each write inserts approximately 1 KB of data.

20% reads / 80% writes

Cloud	PCU size	Reads (ops/s)		Writes (ops/s)	
		p99 < 10 ms	p99 < 25 ms	p99 < 10 ms	p99 < 25 ms
AWS	Small	305	1,110	4,400	4,400
	Medium	830	1,880	7,600	11,600
	Standard	1,700	3,510	20,400	≥ 23,655

Cloud	PCU size	Reads (ops/s)		Writes (ops/s)	
		p99 < 10 ms	p99 < 25 ms	p99 < 10 ms	p99 < 25 ms
Azure	Small	300	500	5,190	5,190
	Medium	1,000	1,000	7,600	7,600
	Standard	1,710	2,655	15,700	16,400
GCP	Small	100	700	4,385	4,385
	Medium	1,090	2,280	11,600	16,400
	Standard	1,280	2,430	≥ 23,550	≥ 23,550

60% reads / 40% writes

Cloud	PCU size	Reads (ops/s)		Writes (ops/s)	
		p99 < 10 ms	p99 < 25 ms	p99 < 10 ms	p99 < 25 ms
AWS	Small	1,480	2,115	1,420	2,600
	Medium	4,500	7,520	5,000	10,030
	Standard	7,440	9,975	≥ 11,830	≥ 11,830
Azure	Small	885	1,490	2,990	≥ 4,850
	Medium	3,300	5,700	5,400	≥ 10,600
	Standard	7,380	7,380	≥ 11,800	≥ 11,800
GCP	Small	1,490	1,490	2,200	≥ 2,200
	Medium	3,890	6,290	6,990	≥ 8,480
	Standard	5,100	6,970	11,100	≥ 11,400

80% reads / 20% writes

Cloud	PCU size	Reads (ops/s)		Writes (ops/s)	
		p99 < 10 ms	p99 < 25 ms	p99 < 10 ms	p99 < 25 ms
AWS	Small	2,800	3,610	1,100	1,480
	Medium	6,000	10,000	3,300	4,310
	Standard	7,600	10,040	≥ 5,990	≥ 5,990
Azure	Small	2,820	≥ 2,260	1,700	4,400
	Medium	6,800	10,000	≥ 4,600	≥ 4,600
	Standard	11,800	14,100	4,600	5,100
GCP	Small	1,990	3,600	1,160	1,720
	Medium	5,190	9,200	≥ 3,490	≥ 3,490
	Standard	10,000	11,600	≥ 3,470	≥ 3,470

Not yet benchmarked (Coming Q4 2026)

The following use cases are still being tested, and will be included in this document as soon as they become available.

- **LWT:** Lightweight transactions.
- **SAI:** Storage-Attached Indexing queries.
- **Vector:** Vector search.