KVStore Performance Summary

Date: January 6, 2026
Server: Standard_E80ids_v4 (80 vCPUs, 2 NUMA nodes)
Network: 30 Gbps (8 NICs with Accelerated Networking)
Payload: 128 tokens per block (512 bytes per block)
Test Client: Azure Linux VM (same VNet, private gRPC connectivity)
Configuration: Dual NUMA (Port 8085 + 8086), 8 client processes

---

Executive Summary

Metric	Value
Peak Throughput	**~94K tokens/sec** (reads + writes)
Peak TPS	~82 iterations/sec
Optimal Concurrency	32-64 total (8 processes × 4-8 each)
Lookup Latency	p50=4.7ms, p99=12.5ms (at C=32)
Read Latency	p50=26ms, p99=69ms (at C=32)
Bottleneck	Azure Storage latency (~20-40ms per read)

Key Finding: Lookup is fast (~5ms) since it's metadata-only. Read latency dominates (~20-60ms) due to Azure Storage blob retrieval. Network is NOT the bottleneck.

Interactive Charts

Open benchmark_charts.html in a browser for interactive performance charts.

Benchmark Results (8-Process Sweep)

Token Calculation

- **Per iteration:** 7 Lookups + 6 Reads + 3 Writes
- **Tokens transferred:** 6 reads × 128 + 3 writes × 128 = **1,152 tokens/iteration**

Performance by Concurrency

Concurrency	TPS	K Tokens/s	Lookup p50	Lookup p90	Lookup p99	Read p50	Read p90	Read p99
1	3.6	4.1	4.95ms	8.08ms	11.16ms	18.81ms	20.75ms	22.84ms
2	7.4	8.5	4.84ms	7.91ms	11.07ms	19.17ms	21.24ms	25.51ms

Concurrency	TPS	K Tokens/s	Lookup p50	Lookup p90	Lookup p99	Read p50	Read p90	Read p99
4	14.6	16.8	4.72ms	7.67ms	10.80ms	19.60ms	22.06ms	32.24ms
6	22.0	25.3	4.63ms	7.55ms	11.92ms	20.19ms	22.71ms	28.79ms
8	28.8	33.2	4.72ms	7.49ms	12.07ms	20.61ms	23.52ms	35.80ms
16	47.6	54.8	4.56ms	7.33ms	11.44ms	22.29ms	27.95ms	43.36ms
32 ★	**65.8**	**75.8**	4.68ms	6.46ms	12.49ms	25.96ms	39.86ms	69.17ms
64	80.5	92.7	5.23ms	9.02ms	23.54ms	41.97ms	83.60ms	157.82ms
80	**81.5**	**93.9**	5.63ms	11.54ms	28.65ms	57.04ms	104.81ms	166.90ms

Key Observations

1. **Lookup latency is stable**: p50 stays ~4.5-5.5ms across all concurrency levels
2. **Read latency increases with load**: p50 goes from 19ms (C=1) to 57ms (C=80)
3. **Throughput plateaus at C=64-80**: ~82 TPS / ~94K tokens/sec is the ceiling
4. **Sweet spot is C=32**: Best balance of throughput (76K tokens/s) and latency (p50=26ms)

Latency Breakdown

Lookup (Metadata Query)

- **Operation:** Check if KV blocks exist in cache (hash-based)
- **Typical latency:** 4-6ms p50, 7-12ms p99
- **No blob data transferred** - just metadata

Read (Blob Retrieval)

- **Operation:** Stream token blocks from Azure Storage
- **Typical latency:** 20-60ms p50, 70-170ms p99
- **Latency dominated by storage access time**

Why Read >> Lookup?

```
Lookup: Client → gRPC → Server → Table Query → Response
Read:   Client → gRPC → Server → Blob Download → Stream → Client
                                      ↑
                                ~15-30ms per blob
```

Recommended Operating Points

Profile	Total Concurrency	Throughput	Lookup p50	Read p50	Use Case
Low Latency	8 (8×1)	~33K tok/s	4.7ms	21ms	Real-time, interactive
Balanced ⭐	32 (8×4)	~76K tok/s	4.7ms	26ms	Web apps, APIs
High Throughput	64 (8×8)	~93K tok/s	5.2ms	42ms	Batch processing
Max Throughput	80 (8×10)	~94K tok/s	5.6ms	57ms	Background jobs

Capacity Planning

Single Server Capacity

Latency SLA	Recommended Config	Throughput	TPS
Read p50 < 25ms	8 processes × 1	~33K tok/s	29
Read p50 < 30ms	8 processes × 2	~55K tok/s	48
Read p50 < 45ms	8 processes × 8	~93K tok/s	81
Read p50 < 60ms	8 processes × 10	~94K tok/s	82

Scaling Formula

```
Servers Needed = Target_Throughput ÷ Per_Server_Capacity

Example: 1M tokens/sec with p50 < 30ms
- Per server: ~55K tokens/sec (8×2 config)
- Servers: 1,000,000 ÷ 55,000 ≈ 19 servers
```

Summary

Key Metrics

Metric	Value
Peak Throughput	~94K tokens/sec
Optimal Concurrency	32 (8 processes × 4)
Lookup Latency	p50=4.7ms (stable)
Read Latency	p50=26ms at C=32, 57ms at C=80

Metric	Value
Bottleneck	Azure Storage latency

Key Insights

1. **Lookup is fast and stable** (~5ms) - metadata query only
2. **Read latency scales with concurrency** - 20ms→60ms as load increases
3. **Throughput plateaus at ~94K tokens/sec** - storage-bound
4. **Sweet spot is C=32** - best throughput/latency balance

Generated from benchmark sweep on January 6, 2026 See benchmark_charts.html for interactive charts