Micron 6600 ION SSD vs. 16 HDDs: ETL throughput that actually scales

At scale, AI pipeline performance is gated by data movement.[1]

Modern AI and analytics pipelines are increasingly constrained by data movement, where storage throughput and latency materially affect pipeline efficiency and time-to-results.[2]

The extract, transform, load (ETL) process for these systems involves two common storage-intensive stages:

- Preprocessing reads raw datasets (images, text, sensor data) with random access patterns and writes transformed shards sequentially.
- Ingest ingests streams of prepared data into training frameworks with heavy sequential writes and lighter verification reads.

Both stages generate small metadata/index updates throughout. Storage throughput and latency strongly influence GPU utilization and time-to-train.[3]

This technical brief compares ETL pipeline performance between a single Micron 6600 ION NVMe SSD (245TB) and a 16-drive HDD RAID0 array of 16TB drives (256TB)[4] across two workload profiles: Preprocessing (read-heavy) and ingest (write-heavy), highlighting the results and architectural differences of each.

Micron 6600 ION SSD (U.2, E3.L, E3.S 1T)

1. "Real-time Data Infrastructure at Uber" (arxiv.org).
2. See "Scaling Cloud ETL: Optimizing Performance and Resolving Azure Data Factory Copy Bottlenecks" (techcommunity.microsoft.com) for additional background on ETL workloads.
3. Statements based on Micron Data Center Workload Engineering IO trace analysis.
4. Unformatted. 1GB = 1 billion bytes. Formatted capacity is less.
5. Throughout this document, the 1x 245TB Micron 6600 ION SSD configuration is referred to as "1x SSD" and the 16x 16TB HDD configuration as "16x HDD" for brevity. Different HDD vendor products may give different results.

Key findings

One Micron 6600 ION NVMe SSD consistently delivered higher ETL throughput, lower latency, superior power efficiency, and scalable concurrency than a 16-drive HDD[5] array as measured by FIO under defined ETL workload profiles, enabling faster and more predictable pipelines.

8.6x

Preprocessing average throughput

ETL preprocessing throughput averaged 5,415 MB/s with 1x SSD versus 632 MB/s with 16x HDD, an 8.6x difference

3.4x

Ingest average throughput

1x SSD sustained an average of 3.7 GB/s vs ≈1.1 GB/s for the 16x HDD array, a 3.4x difference.

23x

Lower preprocessing read latency

1x SSD completed 256KB random reads in ≈0.36ms versus ≈8.4ms for the 16x HDD array, a 23x difference.

29x

Lower ingest read latency

The 1x SSD completed 256KB verification reads in 0.32ms versus 9.22ms for the 16x HDD array, a ≈29x difference.

84x

Storage-level power efficiency

In preprocessing, the 1x SSD delivered ≈292.7 MB/s per watt vs ≈3.5 MB/s per watt for the 16x HDD array, which improved suitability for power- and cooling-constrained deployments.

micron.com/6600-ION

Configurations tested

This comparison uses a simple, approximately capacity-matched pairing for the evaluated workload profiles: 1x 245TB Micron 6600 ION SSD (available in E3.L and U.2 form factors) and 16x capacity-focused, data center, 16TB HDDs (representative of widely deployed, capacity-class data center HDDs), creating practical frameworks for choosing the right storage architectural building blocks.

Each configuration's capacity was held roughly constant to isolate architectural effects. The drive interfaces reflect those in common use: NVMe for the single 245TB Micron 6600 ION and 6Gb/s SATA for a multi-drive HDD configuration (housed in a JBOD).

Parameter	SSD configuration	HDD configuration
Drive type	Micron 6600 ION SSD	Capacity-focused, data center HDD
Capacity per drive	245TB	16TB
Drive count	1	16
Configuration capacity	245TB	256TB
Interface	NVMe (PCIe Gen5)	SATA 6 Gb/s

Table 1: Configurations overview

Throughput analysis[6]

Storage throughput can determine whether storage constrains ETL progress. Because ETL combines scattered reads with sustained writes, we measured throughput under defined transfer sizes: 256KB random reads, 1MB sequential writes, and a single 4KB write stream capped at 5MB/s to represent metadata updates.[7]

These fixed patterns isolated how each architecture sustained ETL throughput under identical read/write assumptions.

Preprocessing throughput

Preprocessing throughput reflects how quickly raw data is converted into pipeline-ready input. In this test, the pipeline pulled data from multiple source locations (modeled as 256KB random reads), applied transformation logic, and streamed results to storage in a structured format (modeled as 1MB sequential writes).

As seen in Figure 1, the 1x SSD configuration showed (2,782 MB/s + 2,633 MB/s) for a combined 5,415 MB/s, while the 16x HDD configuration showed (119 MB/s + 513 MB/s) for a combined 632 MB/s, an ≈ 8.6x difference.

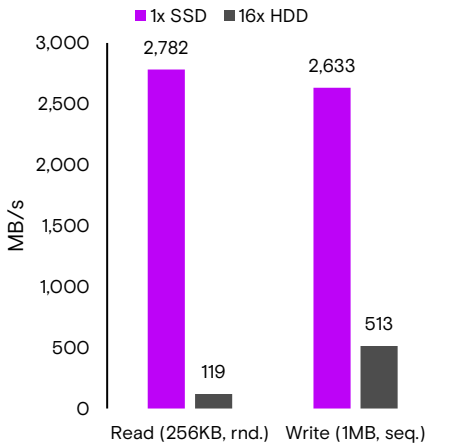

Figure 1: Preprocessing throughput

Ingest throughput

Ingest throughput measures the rate at which prepared ETL output is committed to durable storage.

We modeled ingest as write-dominant (1MB sequential writes) with limited verification reads (256KB random reads). Holding these behaviors constant enabled a repeatable 1x SSD versus 16x HDD comparison, indicating whether the landing rate would slow due to I/O queuing. Figure 2 shows that the 1x SSD configuration delivered 3.4x the ingest throughput of the 16x HDD configuration (786 + 2,863) / (27 + 1,036) ≈ 3.4x.

Next, we quantified completion latency under the same workload.

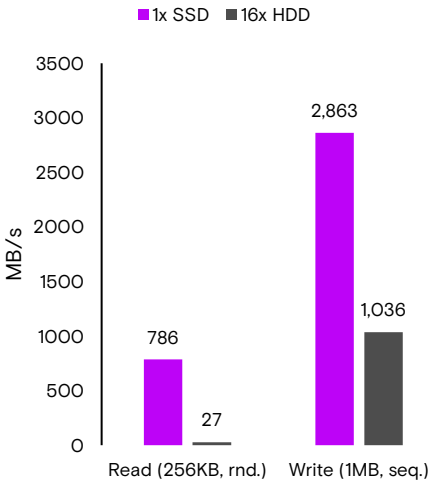

Figure 2: Ingest throughput

6. Throughput was measured with FIO "numjobs" = 1 (numjobs is shorthand for the number of FIO tester jobs used). See "fio - Flexible I/O tester" (fio.readthedocs.io) to learn more about FIO.
7. The 4KB write stream was capped at 5MB/s and did not materially affect the dominant rates shown.

Latency analysis

Latency analysis quantifies I/O responsiveness under ETL load, a key determinant of pipeline concurrency and predictability. We tested read/write latency in milliseconds (ms) for 256KB random reads and 1MB sequential writes to indicate pipeline throughput predictability.

Preprocessing latency

Preprocessing latency analysis measures how quickly the ETL pipeline completes I/O operations. For the preprocessing profile, we measured the mean completion latency for the streams noted earlier: 256KB random read (0.36ms on the 1x SSD versus 8.39ms on the 16x HDD array) and 1MB sequential write (0.76ms on the 1x SSD vs 3.90ms on the 16x HDD array), as seen in Figure 3.[8]

Higher per-I/O completion time can increase queueing and reduce responsiveness, potentially making preprocessing a pipeline pacing constraint. The 1x SSD's 0.36ms average read latency demonstrates the headroom to preserve concurrency as ETL workloads scale.

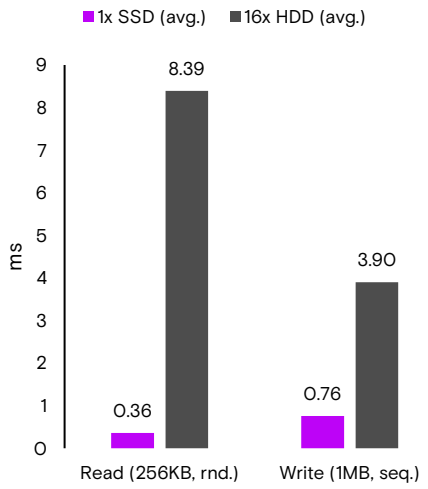

Figure 3: Preprocessing average latency

Ingest latency

Ingest was modeled as a write-heavy "data landing" phase. Its latency tells about the landing rate predictability under concurrent load. The modeled ingest workload emphasized 1MB sequential writes with deeper write depth (iodepth = 4), while maintaining only shallow 256KB random reads (iodepth = 1) for sampling/verification, plus ongoing metadata/index updates.[9]

Figure 4 shows that read latency averaged 0.32ms vs 9.22ms, and write latency averaged 1.40ms versus 3.86ms (for the 1x SSD and the 16x HDD, respectively). In this workload, metadata operations accounted for a small fraction of total I/O and did not materially affect the dominant read and write completion times.[10]

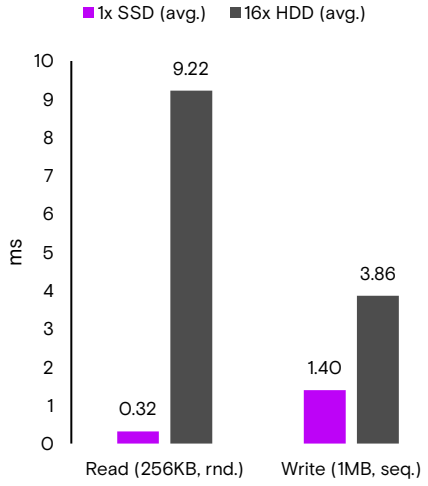

Figure 4: Ingest average latency

Power efficiency analysis

As power and cooling become primary constraints on data center expansion, power efficiency determines which storage architectures are deployable at scale, not just which perform well in isolation. While throughput and latency show ETL pace, power efficiency shows whether that pace fits within power budgets. We analyzed power efficiency from two views: 1. Total system power under load, and 2. Storage subsystem power in isolation. Together, they expressed the metric that matters—useful ETL work per watt—and showed where the watts were spent.

System-level power efficiency

System-level power efficiency converts ETL performance into deployability by dividing throughput by total infrastructure watts (HDD JBOD + IPMI), showing how much usable pipeline work you can sustain in the same power envelope.

Figure 5 shows system-level ETL power efficiency (the amount of usable preprocessing delivered per watt, including total infrastructure power). For preprocessing, the 1x SSD delivered 17.8 MB/s per watt, compared with 1.3 MB/s per watt for the 16x HDD, while for ingest, that difference is 12.2 MB/s per watt compared to 2.1 MB/s per watt.[11]

8. Preprocessing read latency difference calculated as (8.39 / 0.36) ms ≈ 23x difference.
9. The term iodepth refers to the number of concurrent, outstanding I/O requests.
10. Ingest read latency difference calculated as (9.22 / 0.32) ms = ≈29x difference. Ingest read latency matters because it can gate how fast data can be pulled into the pipeline—regardless of downstream compute or write bandwidth.
11. Ratios calculated as (17.8 / 1.3) ≈ 13.7, and (12.2 / 2.1) ≈ 5.8.

Storage-level power efficiency

Storage-level power efficiency isolates the storage subsystem and answers a question operators can use to evaluate storage architecture power use: how much ETL work is delivered per watt of storage power?

Here, storage power was measured at the storage layer (16x HDD + 1x JBOD subsystem power is shown for the 16x HDD path), so the result reflected the storage architecture's overhead rather than the platform's. Figure 6 shows storage-level ETL storage power efficiency for preprocessing and ingest.

In preprocessing, the 1x SSD delivered 292.7 MB/s per watt, while the 16x HDD + JBOD configuration delivered 3.5 MB/s per watt, an ≈ 84x difference.[12] In the ingest workload, the 1x SSD delivered 194.1 MB/s per watt, while the 16x HDD array delivered just 5.6 MB/s per watt, an increase of ≈ 34.7x. This shows that, at the storage layer, 1x SSD enabled higher ETL throughput within fixed power and cooling budgets.

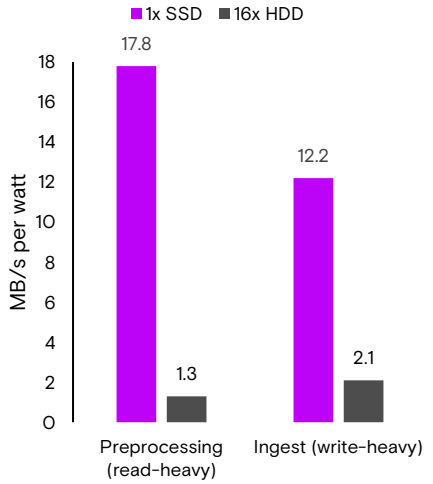

Figure 5: System-level power efficiency

Scaling analysis

Power efficiency determines whether an ETL design is deployable within real power and cooling constraints, but scaling determines whether that advantage holds as concurrency increases. The next section evaluates scaling by increasing worker parallelism (numjobs) to show how each architecture converted additional threads into additional throughput.

Preprocessing scaling | throughput and latency

Figure 7 shows that the 1x SSD combined throughput scaled from 5,415 MB/s to 9,612 MB/s as numjobs increased, while the 16x 16TB HDD stayed constant at about 632 MB/s. The 1x SSD combined throughput advantage widened from ≈ 8.6x to ≈ 15x, indicating that additional concurrency (increasing numjobs) demonstrably benefited the SSD-based architecture.

Figure 8 shows preprocessing latency results and explains why the SSD-backed ETL pipeline achieved lower per-operation completion times under the same workload.

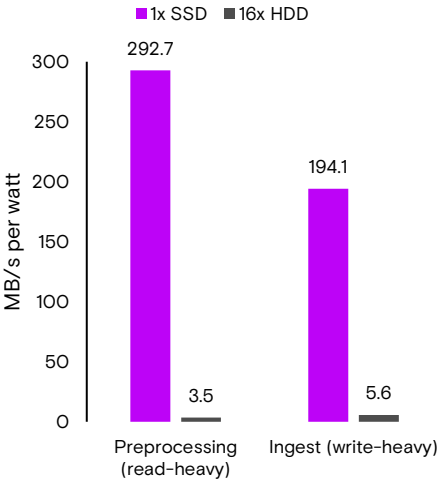

Figure 6: Storage-level power efficiency

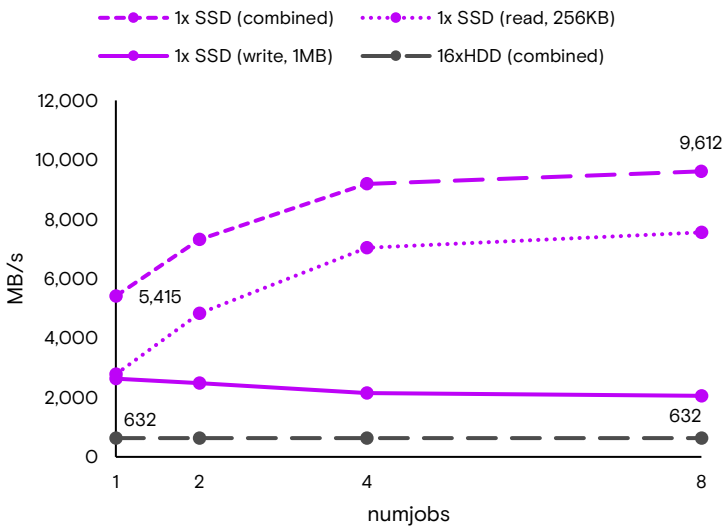

Figure 7: Preprocessing throughput scaling

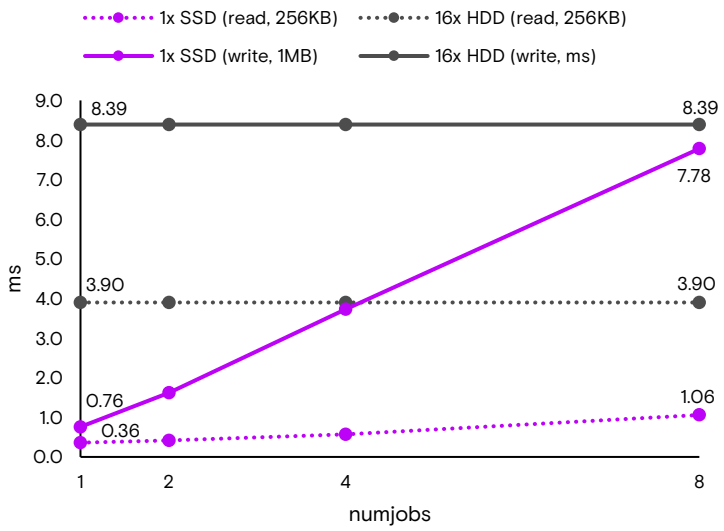

Figure 8: Preprocessing scaling latency

12. Figure 8 shows preprocessing latency results and explains why the SSD-backed ETL pipeline achieves lower per-operation completion times under MB/s per watt calculated as (292.7 / 3.5) ≈ 84.

In preprocessing, the 1x SSD completed 256KB reads at a minimum of 0.36ms, while the 16x HDD array took a minimum of 3.90ms. Writes also showed a meaningful difference, where the 1x SSD took as little as 0.76ms versus 8.39ms for the HDD array. Even with maximal concurrency, the 1x SSD 1MB write latency was lower than the 16x HDD 1MB write latency.

Ingest scaling | throughput and latency

Ingest throughput scaling matters because ingest is where ETL lands the data; if throughput does not scale with concurrency, backpressure builds, and time-to-usable-data increases.[13]

As ingest scales, shared resources (such as storage) contend, increasing latency variability and reducing predictability.

Under the ingest profile shown in Figure 9, the 1x SSD combined throughput increased from 3,649 to 7,258 MB/s as numjobs rose from 1 to 8, while the 16-drive HDD baseline remained at 1,063 MB/s. The SSD advantage expanded from ≈ 3.4x to ≈ 6.8x ((3,649 / 1,063) ≈ 3.4), (7,258 / 1,063) ≈ 6.8.

Write latency is expected to increase as concurrency increases. At numjobs = 8, Figure 10 shows the 1x SSD 1MB write latency of 13.27ms, a value to consider alongside Figure 9, which shows that the 1x SSD also achieved 6.8x, the throughput of the HDD array at that same concurrency. HDD latencies are unchanged (9.22ms read; 3.86ms write). Latency values are best evaluated alongside useful work completed.

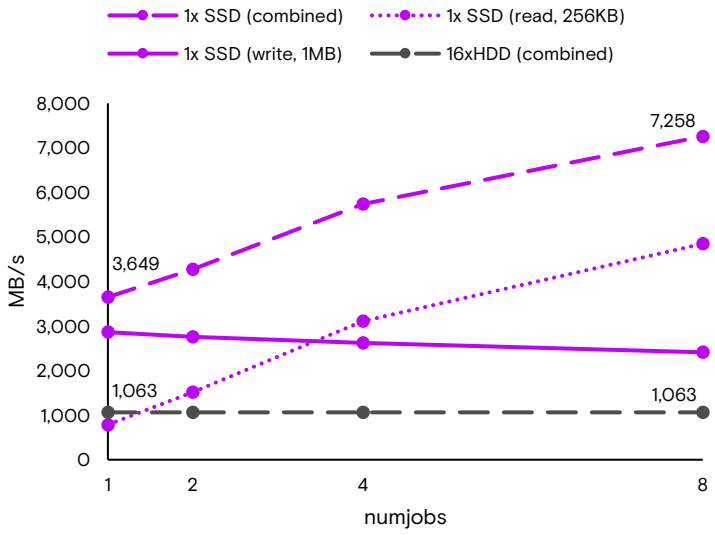

Figure 9: Ingest throughput scaling

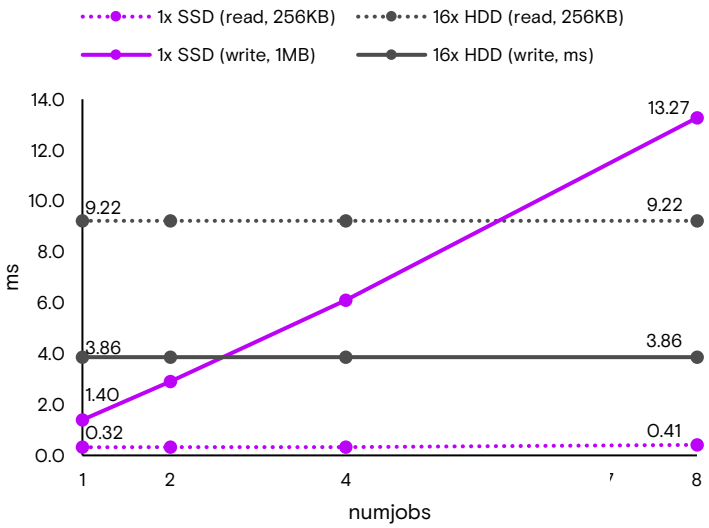

Figure 10: Ingest scaling latency

13. In this instance, "backpressure" means a resistance to processing incoming requests, exhibited as latency.

Conclusion

These results demonstrate a consistent and structural performance advantage for NVMe-based storage across every measured ETL dimension, including throughput, latency, power efficiency, and scalability under concurrency. Under realistic ETL assumptions storage architecture is a major factor in whether pipelines scale predictably. In these results, the NVMe-based design continued to convert additional parallelism into useful work, while the HDD-based design plateaued early due to queueing and contention.

Lower completion latency preserves responsiveness as throughput rises, limiting backpressure between ETL stages and improving schedule predictability. Power efficiency then sets deployability: delivering more ETL work per watt directly constrains rack density, cluster sizing, and operating cost.

At scale, ETL performance is governed by moving data efficiently under concurrency, not peak specs or raw capacity.

Visit www.micron.com/6600-ION to learn more.

micron.com/6600-ION