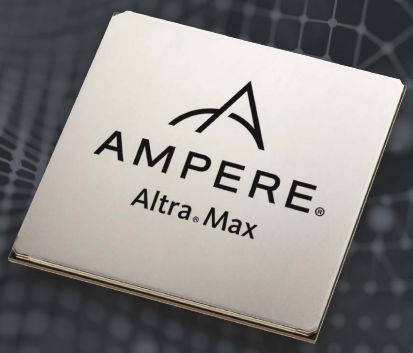
Workload Brief

Memcached In-Memory Key-Value Store on Ampere® Altra® Max



Ampere®—Empowering What's Next

Ampere Altra Max processors is complete system-on-chip (SOC) solutions built for cloud native applications. Ampere Altra Max supports up to 128 cores. In addition to incorporating a large number of high-performance cores, the innovative architecture delivers predictable high performance, linear scaling and high energy efficiency.

Memcached is an open source, in-memory, key-value data store that is typically used for small chunks of arbitrary data (strings, objects) from results of database calls, API calls, or page rendering. Due to its in-memory nature, Memcached is intended for use in speeding up dynamic web applications by caching data and objects in RAM to alleviating database loading. It continues to be ranked as one of most popular key-value stores in the cloud, according to DB-engines.

In this workload brief, we compare Ampere Altra Max M128- 30 to Intel® Xeon® 8380 and AMD EPYC™ 7763 processors running Memcached while measuring the throughput and latencies on each of these processors.

Memcached on Ampere® Altra® Max

Ampere Altra Max is designed to deliver exceptional performance for cloud native applications like Memcached. This is accomplished though an innovative architectural design, operating at consistent frequencies, and using single-threaded cores that make applications more resistant to noisy neighbor issues. This allows workloads to run in a predictable manner with minimal variance under increasing loads.

The processor is also designed to deliver exceptional energy efficiency. This translates to industry leading performance/watt capabilities and a lower carbon footprint.

Benefits of running Memcached on Ampere Altra Max

- **Cloud Native**: Designed from the ground up for 'born in the cloud' workloads like Memcached, Ampere Altra Max can deliver up to 74% higher performance than the best x86 servers.
- **Power Efficient**: With up to 128 energy-efficient Arm cores, Ampere Altra Max can consume up to 34% lower power while maintaining competitive levels of performance.
- **Lower Carbon Footprint**: Industry-leading performance and high energy efficiency result in Ampere Altra Max demonstrating up to 2.6x higher Performance/watt, leading to lower TCO and a smaller carbon footprint.
- **Consistency & Predictability:** Single-threaded cores running at fixed maximum frequencies ensure linear scaling under stringent SLAs and at high loads while running Memcached.

Benchmarking Configuration

We have used memtier_benchmark (developed by Redis Labs) as a load generator for benchmarking Memcached. Each test was configured to run with multiple threads, multiple clients per thread, and with pipelining enabled.

We recommend compiling Memcached server with GCC 10.2 or newer as newer compilers have made significant progress towards generating optimized code that can improve performance.

We used CentOS 8.4 (kernel 4.18) with Memcached 1.6.9 compiled with GCC 10.2 for our tests. We compared Ampere Altra Max M128-30, AMD EPYC 7763 and Intel Ice Lake Refresh (refer to the chart below for results). For each of the tests, we used similar clients to generate requests to Memcached-server.

Since it is realistic to measure throughput under a specified Service Level Agreement (SLA), we have used a 99th percentile latency (p.99) of 1 millisecond. This ensures that 99 percent of the requests have a response time of 1 ms in the worst case.

The test ran for 2 minutes with a 1:10 set:get ratio (1 key/value write and 10 key/value read) and 128 bytes payload, which is common for in-memory caches. We initially used an appropriate number of clients and threads/client to load one instance of Memcached server, while ensuring the p.99 latency was at most 1 ms. Pipelining feature in Memcached allows client to pack multiple requests into one single request packet which can reduce packet processing overhead. This feature can dramatically reduce response times and we used 126 concurrent pipelined requests for Ampere Altra Max M128- 30.

Next, we successively increased the number of Memcached instances till one or more instances violated the p.99 latency SLA. The aggregate throughput of all instances was used as the primary performance metric. We ran the test three times and saw minimal run-to-run variations.

Ampere Altra Max

- 128 Armv8.2+ 64-bit cores at 3.0GHz
- 64KB i-Cache, 64KB d-Cache per core
- 1MB L2 Cache
- 16MB-32MB System Level Cache
- Coherent mesh-based interconnect

Memory

- 8x72 bit DDR4-3200 channels
- ECC and DDR4 RAS
- Up to 16 DIMMs (2 DPC) and 4TB/socket

Connectivity

- 128 lanes of PCIe Gen4
- Coherent multi-socket support
- 4x16 CCIX lanes

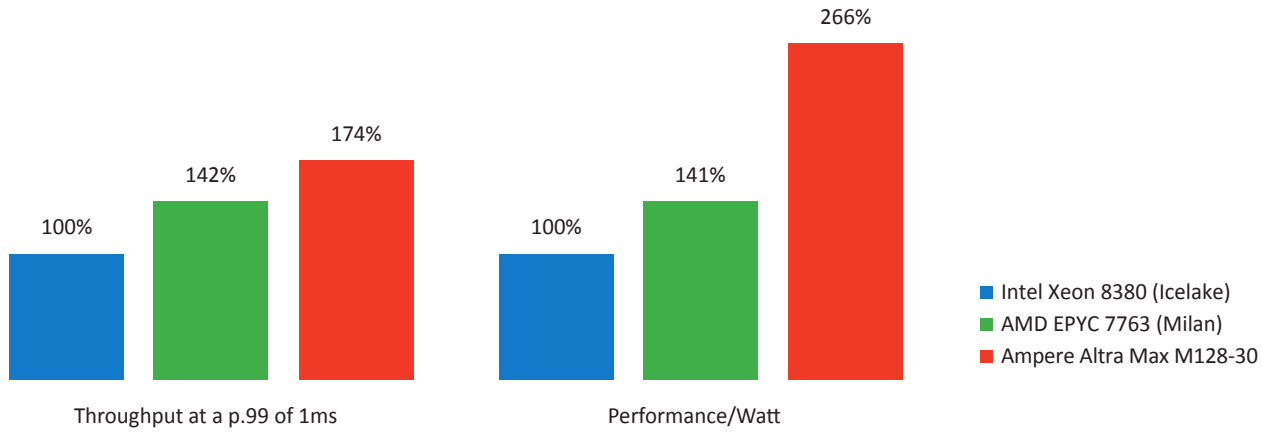
System

- Armv8.2+, SBSA Level 4
- Advanced Power Management

Performance

- SPECrate®2017_int_base:350

Ampere Altra Max M128-30 Industry-leading Performance and Energy Efficiency on Memcached



Benchmarking Results and Conclusions

As can be seen in the chart above, we observed up to a 1.74x on Ampere Altra Max compared to Intel Xeon 8380.

We observed up to a 1.23x improvement in performance on Ampere Altra Max compared to AMD EPYC 7763.

For large-scale cloud deployments energy efficiency as measured by performance/watt is an important metric in addition to raw performance. Ampere Altra Max processors have 2.66x better performance/Watt under a specified SLA than that on Intel servers and 1.87x higher performance/Watt compared to that on AMD servers.

Fast in-memory caches are used in most cloud usages today. Memcached is a popular high throughput in-memory key-value store that is applicable to low latency applications in a scale out configuration. Ampere Altra Max is designed to deliver exceptional performance and energy efficiency for cloud native applications like Memcached. In Ampere's testing, the processor demonstrated up to 1.74x performance improvements and they achieved up to 2.66x energy efficiency improvements. For more information on this workload or other workloads our engineers have been working on, please visit https://developer.amperecomputing.com/.

Ampere Computing® / 4655 Great America Parkway, Suite 601 / Santa Clara, CA 95054 / amperecomputing.com