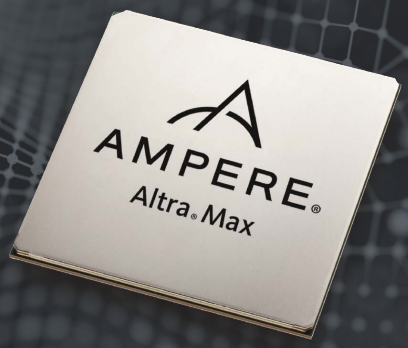
Workload Brief

Redis™ In-Memory Cache on Ampere® Altra® Max



Ampere®—Empowering What's Next

The Ampere Altra and Ampere Altra Max processors are 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.

Redis is an open source, in-memory, key-value data store that is typically used as a database or a cache. It uses an in-memory dataset but data can be persisted through periodic writes or appends to disk. Due to its in-memory nature, Redis is very fast, and it can deliver high throughput at sub-millisecond latencies. It continues to rank highly in popularity among key value stores in the cloud, according to DB-engines.

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

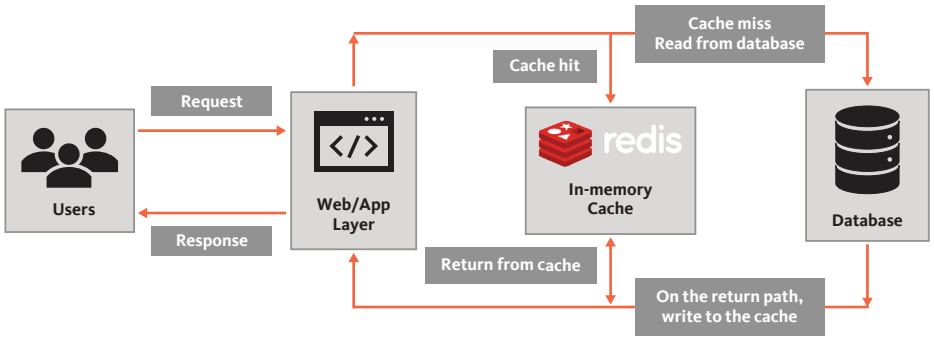


Figure 1: **Using Redis as an in-memory cache for traditional databases**

Benefits of running Redis on Ampere Altra Max

- **Cloud Native**: Designed from the ground up for 'born in the cloud' workloads like Redis, Ampere Altra Max can deliver up to 1.9x higher performance than the best x86 servers
- **Energy Efficiency:** With up to 128 energy-efficient Arm cores, Ampere Altra Max can consume up to 32% 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 2.8x 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 multiple Redis instances.

Redis on Ampere Altra Max

Ampere Altra Max processors are designed to deliver exceptional performance for cloud native applications like Redis. They do so by using 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 processors are also designed to deliver exceptional energy efficiency. This translates to industry leading performance/watt capabilities and a lower carbon footprint.

Benchmarking Configuration

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

We recommend compiling Redis server with GCC (GNU Compiler Collection) 10.2 or newer as newer compilers have made significant progress towards generating optimized code that can improve performance for Aarch64 applications.

We used CentOS 8.3 (kernel 4.18) with Redis-server 5.0.12 compiled with GCC 10.2 for our tests. For each of the tests, we used similar clients to generate requests to Redis-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 3 minutes with a 1:10 get:set ratio, which is common for in-memory caches. We initially used an appropriate number of clients and threads/client to load one instance of Redis, while ensuring the p.99 latency was at most 1 ms. Pipelining is a feature whereby Redis can process new requests even if the client has not already read older responses. This feature can dramatically reduce response times and we used 45 concurrent pipelined requests.

Next, we successively increased the number of Redis instances until 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 Processor Family

- Up to 128 Armv8.2+ 64-bit cores at 3.0GHz, 64KB i-Cache, 64KB d-Cache per core 1MB L2 Cache
- 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 Estimated: 350

Figure 2(A-B): Ampere Altra Max M128-30
Industry-leading Performance and Energy Efficiency with Redis

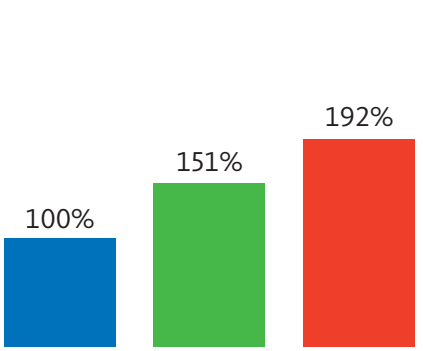


Figure 2A: Throughput (p.99 latency of 1 ms)

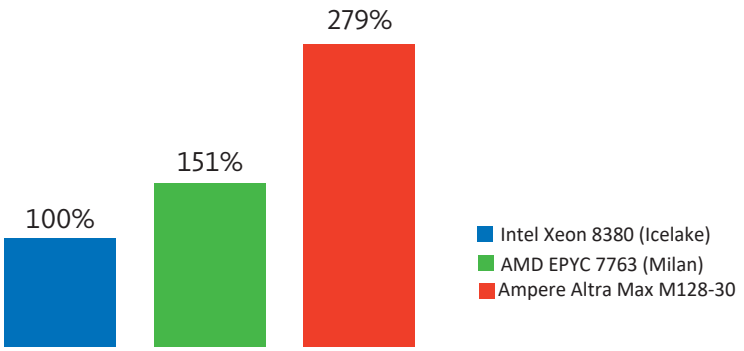


Figure 2B: Performance/Watt

Benchmarking Results and Conclusions

As can be seen in Figure 2A, we observed up to 92% higher throughput on Ampere Altra Max compared to Intel Xeon 8380 and 27% higher throughput compared to AMD EPYC 7763.

For large-scale cloud deployments, performance/watt (i.e. energy efficiency) is an important metric in addition to raw performance. Ampere Altra Max processors demonstrated 2.8x better performance/watt under a specified SLA compared to the Intel Xeon 8380 and 1.85x higher than the AMD EPYC 7763 (refer to Figure 2B).

Fast in-memory caches are used in many cloud workflows today. Redis is a popular high throughput in-memory key-value store that is used in low latency applications in scale out configurations. Ampere Altra Max processors are designed to deliver exceptional performance and energy efficiency for cloud native applications like Redis. In Ampere's testing, these processors demonstrated compelling performance and outstanding energy efficiency compared to x86 processors.

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