Performance Arms X-Gene 3 for Cloud

By Linley Gwennap
Principal Analyst

March 2017

The Linley Group

www.linleygroup.com

Performance Arms X-Gene 3 for Cloud

By Linley Gwennap, Principal Analyst, The Linley Group

Cloud data centers run many different workloads with different performance requirements. Today's ARM-based processors address only a few of these workloads, but the new X-Gene 3 offers a new level of performance that satisfies most cloud applications. The 16nm chip is the industry's first ARM-compatible processor that matches Xeon E5 in CPU throughput, per-thread performance, and power efficiency. It offers significant advantages in memory bandwidth and cost of ownership. AppliedMicro has already validated performance on first silicon and is now sampling X-Gene 3.

Modern cloud data centers run a variety of workloads, many of which have different performance characteristics. Traditional web hosting, for example, comprises many simple requests, whereas a web search is highly CPU intensive. Storage servers require high I/O bandwidth but little processing power. Big-data applications often depend more on DRAM bandwidth and capacity than on CPU performance. HPC applications, by definition, require high compute performance, but emerging variants such as neural-network inferencing don't have the floating-point requirements of traditional scientific code.

Therefore, assessing the performance of a cloud processor is a complex task. Simply looking at the CPU speed (GHz) and number of cores isn't adequate. Compute benchmarks such as SPEC_int provide a good measure of basic CPU performance but don't reflect the capabilities of the memory and I/O subsystem to meet the needs of any given workload. A variety of metrics are needed to better understand how a specific processor maps to the needs of various applications. A single processor cannot optimally serve all workloads.

The X-Gene 3 processor delivers several major improvements to the ARM server lineup, allowing it to serve a broader range of applications than previous models. The third-generation design improves the microarchitecture, making it the most powerful ARM CPU available today. The CPU runs at a base frequency of 3.0GHz in a 16nm FinFET process, faster than current 28nm ARM chips. Using 32 cores, up from 8 in the previous generation, X-Gene 3 is 6x more powerful than its predecessor.[*]

In addition to this powerful set of CPUs, X-Gene 3 includes 32MB of on-chip L3 cache that is shared among the cores. When software requires additional data that is not in this large cache, the processor employs eight channels of high-speed DDR4 DRAM to fetch the data, as Figure 1 shows. For high-speed I/O, 42 lanes of PCI Express Gen3 can connect to external Ethernet NICs or storage adapters. X-Gene 3 is now sampling to lead customers.

[*] More information on the X-Gene 3 processor design is available in our white paper *X-Gene 3 Challenges Xeon E5.*

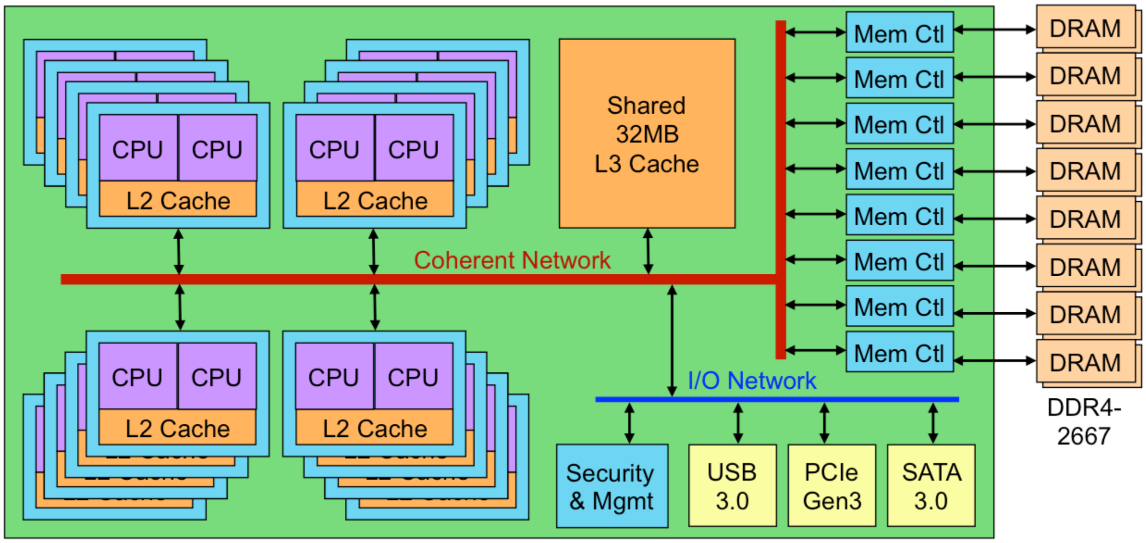

Figure 1. Block diagram of X-Gene 3. The ARMv8 processors includes 32 CPU cores, 32MB of L3 cache, eight channels of DDR4-2667 DRAM, and 42 lanes of PCI Express Gen3.

X-Gene 3 Up and Running

AppliedMicro received first silicon of X-Gene 3 in late 2016, and validation proceeded smoothly. The processor quickly booted CentOS 7.3 Linux using AMI MegaRAC BMC management software and Aptiov UEFI BIOS; it has since booted other 32-bit and 64-bit operating systems. The processor runs with all 32 cores enabled, and the 32MB of L3 cache is fully functional. The CPUs operate at the specified base speed of 3.0GHz and have been validated at the peak turbo speed of 3.3GHz. All PCIe ports can operate in Gen3 mode, and other integrated I/Os such as SATA and USB work at full speed. All eight memory channels operate at up to DDR4-2400; the company is still working to boost the memory speed to the rated DDR4-2667. This type of speed tuning is normal during initial silicon bringup.

The company has run several performance benchmarks on the initial silicon. Based on testing of the current configuration of 3.0GHz CPU frequency and DDR4-2400, the company expects the chip to deliver a SPECint_rate2006 (peak) score of at least 500 when running at its peak speed of 3.3GHz and DDR4-2667 and with some additional hardware and compiler tuning. This score is well ahead of that of any other ARM processor and similar to that of mainstream Xeon E5 processors.

In addition, the processor should achieve a single-thread SPEC_int2006 (peak) score of 24. Other CPU tests such as CoreMark and Dhrystone deliver similarly impressive results, outscoring leading ARM cores such as Cortex-A72 and custom designs from Cavium, Qualcomm, and Samsung. With eight DDR4 channels, X-Gene 3 also posts excellent scores on memory tests. For example, the processor scores 67.1GB/s on Stream Copy using DDR4-2133 and could exceed 80GB/s when it uses DDR4-2667.

The company has not disclosed the power of the initial silicon but still expects the final product to satisfy a TDP rating of 125W. Most server designs can easily handle a processor with this thermal capacity. To assist customers, the company has developed two evaluation boards for X-Gene 3 that supports 8 DIMMs (one per channel) and 16 DIMMs (two per channel) respectively as well as several PCIe ports, other standard I/O, and an AST2500 board-management controller (BMC).

The First Xeon E5–Class ARM Processor

X-Gene 3 compares well to a 14nm Xeon E5-2680v4, which has 14 Broadwell CPU cores running at 2.4GHz and a TDP of 120W. Although Intel has not released specifications for the next-generation Xeon E5-2680v5, we expect that product to include 16 Skylake CPU cores at a similar speed and TDP, offering a 10–20% boost in performance.

Comparing the performance of X-Gene 3 to Xeon E5 is difficult. Intel's high benchmark scores depend on its proprietary ICC compiler, which includes benchmark-specific optimizations that are not useful for real code. Most software developers prefer GCC, which produces code that is easier to debug. GCC provides good performance on typical applications, but benchmark scores generated using GCC are considerably lower than those using ICC. Taking into account this compiler difference, we expect X-Gene 3's performance to wind up in the same ballpark as that of the E5-2680v5 (Skylake).

Intel offers some Xeon processors with more cores and greater performance than the E5-2680. These models will deliver much more CPU performance than X-Gene 3, but their list prices are above $2,000, in some cases several thousand dollars. Most customers buy server processors with a list price of less than $1,500, so X-Gene 3 covers the full range of mainstream server configurations.

X-Gene 3 does better on certain workloads. Using eight DRAM channels (versus four for current Xeon models and six for some Skylake products), it will deliver top marks on memory tests such as Stream. These benchmark scores will translate to superior performance on memory-intensive applications such as in-memory databases.

	Xeon E5-2680v4	Xeon E5-2680v5	X-Gene 3
CPU Type	64-bit x86 (Broadwell)	64-bit x86 (Skylake)	64-bit ARMv8 (custom)
Cores/Threads	14C / 28T	16C /32T†	32C / 32T
CPU Speed*	2.4GHz / 3.3GHz	Not disclosed	3.0GHz / 3.3GHz
L3 Cache	35MB	Not disclosed	32MB
DRAM Channels	4x DDR4-2400	6x DDR4-2667†	8x DDR4-2667
DRAM B/W	76.8GB/s	128.0GB/s†	170.7GB/s
PCI Express	40 lanes Gen3	Not disclosed	42 lanes Gen3
South Bridge	External	External†	Integrated
TDP	120W + SB	120W + SB†	125W
List Price	$1745 + SB	$1745 + SB†	Not disclosed

Table 1. High-end server processors. X-Gene 3 has sizable advantages in memory bandwidth and price and is similar in other parameters. SB=south bridge. *base/turbo. (Source: vendors except †The Linley Group estimate)

X-Gene 3 lags Xeon E5 in floating-point performance, but that doesn't rule out all HPC applications. X-Gene 3 will do well on integer-only workloads such as deep learning, where deploying a neural network (inferencing) often uses integer calculations. Data centers use neural networks for voice services (e.g., Alexa and Siri), image classification, and other emerging tasks.

For customers concerned about power efficiency, X-Gene 3 is rated at 125W TDP, so it should have performance per watt similar to the Xeon chip's. Note that Intel's TDP rating does not include the south-bridge chipset, whereas X-Gene 3 integrates the south bridge, tipping this comparison in X-Gene's favor, as Table 1 shows. The big win for X-Gene 3 is in performance per watt per dollar. The E5-2680v4 carries a list price of $1745, whereas X-Gene 3 will cost about a third less. Thus, the ARM processor should finish far ahead of Xeon on any metrics involving cost.

ARM Competitors Lag in Performance

X-Gene 3 far outclasses current ARM server processors, including the initial X-Gene products, AMD's A1100, and Cavium's ThunderX. Except for ThunderX, these products are limited to eight cores and thus have modest CPU performance. Although ThunderX includes up to 48 cores, each of its cores is quite small, resulting in very low per-core performance. Thus, ThunderX is limited to scale-out applications that lightly load the CPUs. Furthermore, X-Gene nearly doubles that chip's total SPEC_int throughput, whereas ThunderX is limited to only the low end of the Xeon range. Current ARM products also lag in power efficiency due to their 28nm manufacturing nodes.

X-Gene 3 will compete against other next-generation ARM processors that are built in 16nm/14nm technology. ThunderX2 aims for a large increase in per-core performance, but we expect it will still be well below X-Gene 3 and Skylake-EP. Even with this improvement, we don't expect ThunderX2 to compete for CPU-intensive (scale-up) workloads. Furthermore, ThunderX2 is behind X-Gene 3 in reaching the market and may not meet its target specifications, which have not yet been validated in silicon.

Cavium recently acquired a second ARM processor, known as Vulcan. Broadcom originally developed the Vulcan design but decided to exit the server-processor market after the design failed to meet its original schedule and performance targets. We believe Cavium will bring both the ThunderX2 and Vulcan processors to market, creating some overlap in its product line. To resolve this overlap, we expect ThunderX3 will use a CPU based on the Vulcan design.

Qualcomm's forthcoming Centriq-2400 is another competitor, but the company has released little information about its initial product other than that it will have 48 cores. We expect its performance will be similar to that of ThunderX2, since Qualcomm's CPU is a modified smartphone design. Furthermore, the first Centriq design will lack the performance tuning that X-Gene has developed over three generations.

A Mature ARM Platform for Cloud Workloads

With two products in production and a third on the way, X-Gene has already built a strong ecosystem that competitors have yet to match. The platform includes a full set of software, firmware, and development boards. The CPU design has been proven over time and evolved to improve performance and power efficiency. X-Gene 2 is already deployed in a large Asian data center, and leading OEMs such as HP Enterprise are shipping X-Gene systems in volume. Even as it faces new competitors, X-Gene is the most mature ARM-compatible cloud processor.

X-Gene 3 is the industry's first Xeon E5–class processor based on the ARM instruction set. It matches up well against current E5 products in overall CPU throughput, cache size, and I/O configuration while offering big advantages in memory bandwidth, integration, and price. Using 16nm FinFET technology similar to Intel's, X-Gene 3 fits into standard server thermal envelopes and offers power efficiency similar to Xeon's. It uses the most powerful ARM CPU yet announced and the first that delivers per-thread performance similar to Xeon E5's.

As a result, X-Gene 3 can handle a broad range of cloud workloads, including scale-up and scale-out applications. The processor excels on big data, particularly in-memory databases, because of its high memory bandwidth. It can even handle some HPC applications, including deep learning. Customers who move to X-Gene 3 can expect a mature platform with TCO (total cost of ownership) superior to Intel's. With X-Gene 3, ARM is ready for the cloud.

Linley Gwennap is principal analyst at The Linley Group and editor-in-chief of Microprocessor Report. *The Linley Group offers the most comprehensive analysis of microprocessor and SoC design. We analyze not only the business strategy but also the internal technology. Our in-depth reports also cover topics including embedded processors, mobile processors, IoT processors, and processor IP cores. For more information, see our web site at www.linleygroup.com.*

The Linley Group prepared this paper, which AppliedMicro/Macom sponsored, but the opinions and analysis are those of the author.