X-Gene 3 Challenges Xeon E5

By Linley Gwennap
Principal Analyst

April 2016

www.linleygroup.com

X-Gene 3 Challenges Xeon E5

By Linley Gwennap, Principal Analyst, The Linley Group

AppliedMicro's new X-Gene 3 processor design combines 32 ARMv8-compatible CPU cores to deliver an estimated 550 SPECint_rate, competitive with the performance of today's fastest Xeon E5 processors. With eight DDR4 memory channels, X-Gene 3 even outclasses Xeon E5 in memory bandwidth, making it well suited to memory-intensive applications. AppliedMicro expects to sample the new design in 2H16, leading to production shipments in 2H17, about the same schedule as for Intel's Skylake server products.

The server market's need for processor competition remains unfulfilled, as various ARM challenges have yet to prove they have the right stuff to wrestle market share from Intel. The newest challenge comes from AppliedMicro (APM), which has announced the most powerful ARM server processor yet, a 32-core behemoth that will generate more performance than many Xeon E5 products. Customers must be patient, however, as the 16nm chip is just nearing tapeout and should be in production in 2H17, about the same time as Intel's Skylake server products.

At a peak speed of 3.0GHz, X-Gene 3 will score about 550 SPECint_rate2006, according to the company's estimates. Now that the microarchitecture is locked down, APM is running full-chip simulations on an FPGA emulator. Although the simulator isn't fast enough to execute an entire SPECint run, the company is testing the full suite using reduced data sets, which are cross-correlated against X-Gene 2 tests to ensure accuracy. As Figure 1 shows, X-Gene 3's total performance is well within the range of Xeon E5 products. On this metric, it is similar to the fastest parts typically used in cloud servers and trails only very expensive models such as the $4,115 Xeon E5-2699v4.

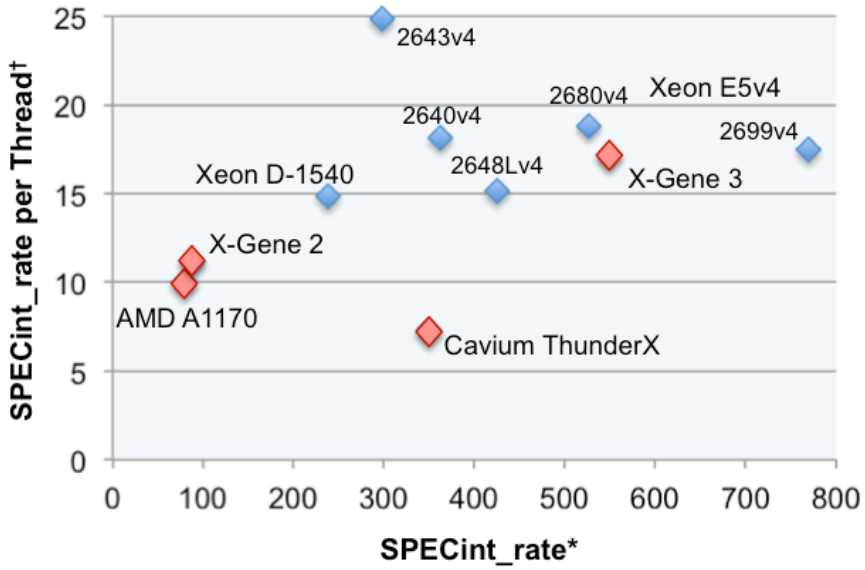

Figure 1. Comparison of server-processor performance. X-Gene 3 delivers better per-thread performance than any other ARM server processor and matches the newest Xeon E5 products in per-thread and total performance. *SPECint_rate2006 (base) for GCC; all ICC scores reduced by 15%. †at maximum thread count. (Source: vendors)

In addition to rip-roaring per-socket performance, X-Gene 3 delivers respectable per-thread performance, as Figure 1 shows. While most other ARM server-processor vendors are using smartphone-class CPUs, APM designed its own high-performance quad-issue microarchitecture backed with robust cache and memory subsystems. Although not quite as powerful as Intel's finely honed CPUs, the X-Gene design comes close, providing enough per-thread and per-socket performance for mainstream Web scale workloads. And all this performance comes within a power envelope similar to Xeon E5's.

	Compute	Memory	I/O	Networking
Storage	Moderate	Moderate	High	High
In-memory Database	Moderate	High	Moderate	Moderate
Web Search	High	Large	Moderate	High
Web services	High	Moderate	Moderate	Moderate
Machine Learning	High	High	Low	Low
HPC	High: Scalar and vector	High BW	Moderate	High/Fabric

Table 1. Cloud application requirements. The processing requirements of modern cloud workloads vary widely. Not all need high compute performance. (Source: The Linley Group)

Table 1 shows the key performance characteristics of mainstream web-scale workloads. X-Gene 3's biggest advantage will be on memory-intensive applications such as web search, big data, machine learning, and high-performance computing (HPC). These workloads require a combination of high per-thread and per-socket performance coupled with a large memory subsystem.

AppliedMicro is also working on its next generation, called X-Gene 3XL. This generation will further improve compute performance with 64 cores per socket, targeting 1000 SPECint_rate. It also extends coherency across sockets to deliver a 2P SMP system.

Third Time's the Charm

Previous X-Genes were small chips designed for the mature 40nm and 28nm HKMG nodes at TSMC. Although using older process technologies reduced execution risk, it incurred a huge performance-per-watt disadvantage given that Intel was already shipping server processors in 2012 and 2013 using the 22nm FinFET node. X-Gene 3 is designed for TSMC's 16nm FinFET+ node, which will go a long way toward narrowing the gap.

Taking advantage of the greater process density and a larger die area, X-Gene 3 features 32 CPUs. The new chip uses an improved version of the existing X-Gene core. To speed time to market, the improvements are relatively minor, including enlarging the branch prediction tables and the TLBs, adding 10% to IPC. The advantages of a FinFET technology enable the CPU cores to operate at 3.0GHz—a sizable gain over the 2.4GHz of X-Gene 2. At this speed, the chip is estimated to have a TDP between 110W and 125W,

similar to that of a high-end Xeon E5. The chip also supports Turbo speeds up to 3.3GHz to deliver higher throughput while staying within the TDP envelope.

As in the previous generation, each pair of cores shares 256KB of L2 cache and acts as a single requesting agent. To accommodate the greater number of cores, the L3 cache size is quadrupled to 32MB, distributed among the cores. More important, the coherent fabric is scaled up in bandwidth to provide efficient scaling. AppliedMicro estimates that, as the thread count increases from 8 to 32, X-Gene 3 delivers three times better SPECint_rate performance, a very good scaling factor of 75%.

As Figure 2 shows, X-Gene 3 has a totally redesigned memory subsystem with eight DDR4 channels operating at 2.67GT/s, theoretically yielding 170GB/s, with a maximum capacity of 1TB of memory. This change will place the design well beyond Xeon D (two DDR4 channels) and ahead of current Xeon E5 models (four DDR4 channels).

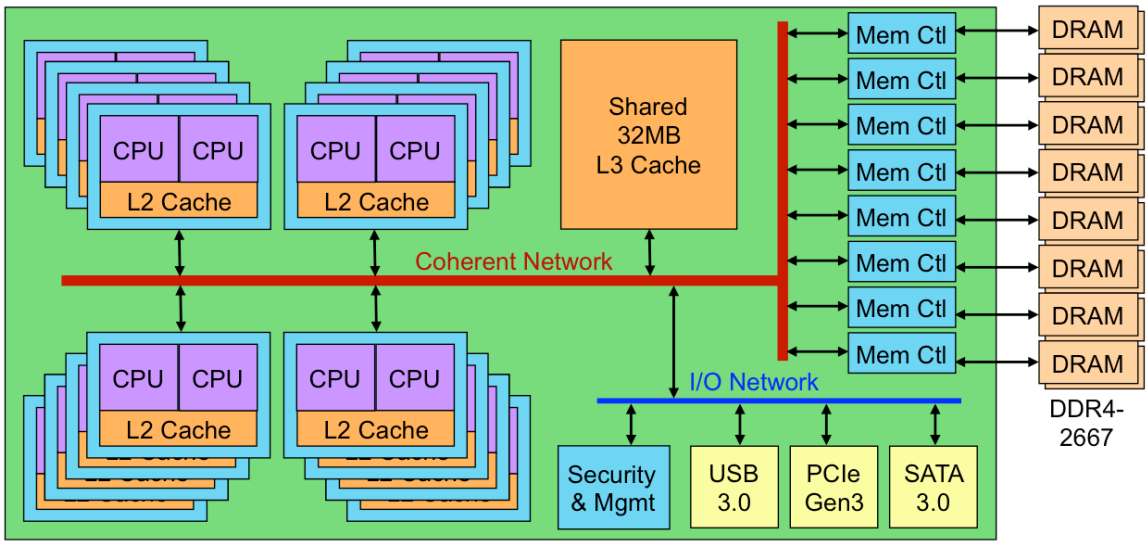

Figure 2. Block diagram of X-Gene 3. The ARMv8 processors includes 32 CPU cores, 32MB of L3 cache, eight channels of DDR4-2667 DRAM, and 42 lanes of PCI Express Gen3.

AppliedMicro has increased the reliability features of the chip to support enterprise-class servers. For example, the memory controllers support advanced ECC options to detect and correct errors even if a complete DRAM chip does not respond. The processor is designed for end-to-end data poisoning; instead of terminating a process upon an uncorrectable memory or I/O error, it tags the bad data and allows it to pass through the system, only causing a terminal error if the CPU actually attempts to use the bad data. The large L3 cache also includes ECC, and the processor can conduct background scrubbing of the L3 cache and DRAM to locate and correct single-bit errors before they accumulate into uncorrectable errors.

Flexible I/O System

AppliedMicro has also comprehensively overhauled X-Gene 3's I/O capabilities. Instead of dedicated Ethernet MACs with RDMA support, the new chip simply provides 42 lanes of PCI Express 3.0 that can connect to external Ethernet controllers. This approach

is the same that Intel uses in its Xeon E5 processors, making it easier for server designers to substitute X-Gene 3. This approach also allows designers the flexibility of choosing any combination of Ethernet speeds and links; they can even use in-house NICs.

Unlike earlier X-Gene chips, the new processor does not include any specific offload engines. Although studies indicate that accelerators can provide a sizable performance boost, many end customers have been reluctant to use them for fear of losing software portability across processors from multiple vendors. Of the few end customers that are using accelerators, most have designed their own instead of taking advantage of the processor's integrated engines. Thus, X-Gene 3 allows customers to connect external accelerators such as FPGAs, ASICs, and GPUs using PCIe.

Similarly, the PCIe links can connect to external storage controllers of the customer's preference. Given the data center's transition away from hard disks, we expect most customers to rely on NVM Express, which uses PCIe to connect directly to SSDs. X-Gene 3 includes four SATA3 links to directly support traditional hard drives.

X-ing Out Xeon D

X-Gene 3 should be a big leap forward for AppliedMicro. Compared with current ARM-based server processors, it boasts 2–6x better aggregate compute performance, memory bandwidth, and I/O bandwidth. Compared with Cavium's ThunderX, X-Gene 3 will deliver more total performance using fewer cores, as Table 1 shows. As a result, the Cavium chip lags well behind in per-thread performance.

Even at slightly greater power, X-Gene 3 is still more power efficient than ThunderX, which is built in 28nm technology. Its memory bandwidth and memory capacity double that of ThunderX. Cavium may move ThunderX to 16nm in the same timeframe that X-Gene 3 enters production, but even then, that chip is unlikely to reach the same level of per-thread or per-socket performance.

Although a good showing relative to other ARM-based products is nice, the real test is how AppliedMicro lines up against the incumbent products from Intel. Unlike previous generations, X-Gene 3 clearly outstrips Atom and Xeon D processors in per-socket performance, per-thread performance, and memory capacity. Table 2 matches X-Gene 3 against the latest 14-core Xeon E5-2680v4, which is near the top of the E5 product line. Both chips use roughly the same power, but X-Gene 3 is rated at greater performance, indicating an advantage in power efficiency. (As we did above, this chart uses SPECint scores but derates Intel's scores by 15% to compensate for its use of ICC for benchmarking.)

X-Gene 3 offers about the same amount of cache as the E5-2680v4, but its eight DDR4 DRAM channels provide more than twice the memory bandwidth. This advantage will help X-Gene 3 in memory-intensive applications. The 14-core Xeon has a lead in per-core performance, but when running with two threads per core, its per-thread performance is about the same as X-Gene 3's. In 2017, Intel is likely to have its Skylake processor in servers; this 14nm product is likely to deliver slightly better performance than Xeon E5v4 and similar power efficiency.

	AppliedMicro X-Gene 3	Intel Xeon E5-2680v4	Cavium ThunderX CP	Intel Xeon D-1540
CPU Core	Potenza++	Broadwell	Thunder	Broadwell
Max Sockets	1S	2S	2S	1S
Cores, Threads	32C / 32T	14C / 28T	48C / 48T	8C / 16T
Max Integer IPC	4 IPC	5 IPC	2 IPC	5 IPC
Base Clock Speed	3.0GHz	2.4GHz	2.5GHz	2.0GHz
Total Cache	32MB	35MB	16MB	12MB
Memory Bandwidth	170.7GB/s	76.8GB/s	76.8GB/s	34.1GB/s
Memory Capacity	1,024GB	1,536GB	512GB	128GB
Ethernet Interfaces	None	None	2x 100GbE + 10x 10GbE	2x 10GbE
PCIe Bandwidth	82.8GB/s	78.9GB/s	31.5GB/s	55.3GB/s
SPECint_rate*	550	527*	350	238*
SPEC/Thread*	17.2	18.8	7.3	14.9
Power (TDP)	110–125W	120W	95W	45W
IC Process	16nm FF+	14nm HP	28nm HKMG	14nm HP
List Price	Not disclosed	$1,745	$600–$800	$581
Production	2H17†	1Q16	4Q15	1Q15

Table 2. X-Gene 3 versus selected processors. The AppliedMicro offering is much more capable than other single-socket server chips—most importantly, Intel's Xeon D. In many respects, it compares favorably with a high-level Xeon E5v3. *Using GCC; ICC scores derated by 15%; SPECint_rate per thread at maximum thread count. (Source: vendors, except †The Linley Group estimate)

One area where AppliedMicro should have no trouble besting Intel is price. Although the company has not announced a list price for X-Gene 3, the chip has similar throughput as high-end Xeon E5 products that have a list price in excess of $1,500. We expect AppliedMicro to significantly undercut these prices.

X-Gene 3 Leads the Third Wave

X-Gene 3 is an aggressive departure from AppliedMicro's previous product strategy. Earlier X-Gene designs were 1–2 process nodes behind Intel and offered few relatively cores and high I/O integration; these parts target the Xeon E3 and Xeon D segments. In contrast, the latest version is designed for the leading-edge 16nm FinFET node and expands the die size to achieve a high core count, larger L3 cache, and generous number of DDR4 memory interfaces, enabling X-Gene 3 to compete head-to-head with mainstream Xeon E5 server processors.

One advantage of this crawl-walk-run approach is that AppliedMicro's investment has scaled with the overall maturity of the ARM ecosystem. X-Gene 1 and X-Gene 2 have gained several customers in scale-out storage, in-memory database, SDN/NFV, and HPC applications, including HP Enterprise, Kontron, ODMs such as Gigabyte and Mitac, and several cloud service providers. As the first vendor to commercialize ARMv8

server processors, AppliedMicro has cleared roadblocks to enable the ARM ecosystem for compute, storage, and networking applications. In this process, the company has gone through server qualification cycles with several OEMs and end users, and it has gained valuable insights into the workload requirements for cloud services, SDN/NFV applications, and embedded applications. X-Gene 3 builds upon this architectural maturity.

The new design is the first of a new wave of ARM-based server processors that could effectively challenge Xeon E5. At 32 cores, eight DDR4 channels, and 42 PCI Express 3.0 lanes, X-Gene 3 sets a high bar for its ARM competitors. We don't expect Cavium or Qualcomm to match X-Gene 3's performance in 2017. Furthermore, as a third-generation product, X-Gene 3 includes more tuning of the CPU, cache, and DRAM controllers for server applications, and it offers a stronger suite of reliability features.

AppliedMicro's real competition, however, is Intel. X-Gene 3 is on track to surpass Xeon D and even many Xeon E5 processors in per-socket throughput while delivering respectable per-thread performance. X-Gene 3 crushes the current Xeon E5 in memory bandwidth, and we don't expect Intel to address this problem in Skylake to avoid undercutting its high-margin Xeon E7 business. Thus, X-Gene 3 should win in memory-intensive applications, and it can undercut Intel's pricing to win general-purpose designs. But AppliedMicro must first demonstrate that its new chip can meet performance and power expectations, as these are the metrics that matter most to customers.

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.*