REPORT REPRINT

Neoverse is Arm's big opportunity to expand influence into datacenter infrastructure

MARCH 1 2019

By John Abbott

The first extensive CPU core roadmap from Arm, Neoverse is aimed specifically at infrastructure, as distinct from consumer IP lines. It comes in the wake of significant signs of traction for the firm in the systems and cloud infrastructure space.

451 TAKE

In 2015, Arm started forecasting that its chips would power between 15% and 20% of overall server shipments by 2020. In January 2016, it upped that figure to an astonishing 25%. To say the least, everyone was skeptical. Today, Arm doesn't talk about servers in isolation, but does talk about the 'global internet infrastructure,' of which it now claims a 27% and growing market share, up from 5% in 2011. That's a lot more convincing, given the growth of the public cloud as the dominant server compute model, and – coupled with its extension toward the intelligent edge and the huge amount of data likely to be generated from the massively expanding mesh of automated and connected IoT devices supporting user – quite distinct from devices that users carry themselves and maintain personally. Making a case for the enterprise use of Arm servers, particularly at an early stage of development without a strong supporting ecosystem in place, was never going to be easy. It's a different story for cloud hyperscalers, where efficiency and integration flexibility are essential, and where Linux and open source software layers are already in widespread use, reducing barriers to entry. The upgrading of users of its existing embedded systems to the intelligent edge then also becomes a more compelling proposition.

Context

We detailed the history of Arm's 11-year effort to move into datacenter infrastructure in our previous report. But since then, there have been three significant developments, aside from the initial launch of Neoverse itself last October.

The first came just a few weeks after that launch: Amazon's announcement of the custom Graviton chip, based on the 16nm Neoverse Cosmos platform (equivalent to the current Cortex-A72 64-bit core). Amazon claims that Graviton will cut the costs of running scale-out workloads by up to 45% for workloads using small instances, such as containerized micro-services or web tier applications.

Of more immediate significance for Arm is the availability of EC2 A1 instances through AWS, providing customers with instant access to Arm-based servers for testing purposes for the first time. The second development was VMware's ESXi demonstration, which represents its first foray for the hypervisor beyond x86. It's important because the cloud is extending to the edge, and hyperscale clouds don't typically use VMware, meaning that VMware will miss out on this emerging market if it doesn't evolve.

Finally, Huawei's new Kunpeng 920 chip adds a significant new architectural licensee for Arm. The high-end chip joins Fujitsu's A64FX, Ampere's eMag and Cavium's ThunderX2 as contender engines for HPC and hyperscale systems.

Products

The Neoverse family represents the first extensive CPU core roadmap from Arm aimed specifically at infrastructure – as distinct from its consumer IP lines, focused particularly on smartphones. Four generations of the new family were first described in general terms last October.

Cosmos is the infrastructure rebranding for the existing generation, based on 16nm Cortex A72 and A75 cores. Ares, using a 7nm process, is the follow-on from that, and in a sense it is the first 'real' Neoverse offering. Beyond that, the roadmap shows Zeus (7nm+) in 2020, and Poseidon (5nm) in 2021. Remarkably, Arm is promising annual systems performance gains of 30% per generation.

Neoverse N1 is the first specific Ares product, and Arm says it overdelivered – providing 60% and higher performance on key cloud workloads compared to Cosmos (A72) at the same frequency, and a 30% greater efficiency gain in the same process. Its design point was to deliver datacenter-class performance, but it could also be utilized to achieve density in the cloud (i.e., maximizing compute per socket) and for performance at the edge.

Performance gains come from multiple design features. Central are the sophisticated pipelining techniques – specifically an 11-stage 'accordion' pipeline, which cuts down stages for branch misses and lengthens for normal operation, and separated pipes for integer and vector – and the two-level caching structure designed for large, branch heavy infrastructure workloads. There's also the coherent mesh architecture and coherent I-cache (a first for Arm) that enables high performance to be sustained over large-scale, many-core systems, and a memory hierarchy supporting low latency, high bandwidth and scalability.

For networking, there's PCIe Gen4 full-bandwidth support and data preloading in core caches. Ares designs can scale up to 128 cores at 150 watts power footprint, or down to 8 cores (25 watts) for network, storage and security applications. Edge devices might require core counts somewhere between 16 and 64. Chip designers can break 64-core dies down into chiplets for more cost-effective manufacturing, connecting them via CCIX links. Arm has built in RAS (reliability, availability, serviceability) features such as error handling and ServerReady certification (ensuring that servers boot up and run in a standard way).

Neoverse E1 is a variant with smaller cores designed for next-generation data planes, throughput applications, such as 5G and edge cloud. Arm believes 5G and software-defined networking will require a rethinking of how data is transported from edge to core. The E1 architecture aims to maximize throughput while balancing compute and efficiency requirements. Arm claims it offers 2.7x the throughput performance of the current Coretex-A53, with a 2.4x throughput to power efficiency and 2.1x compute performance. A likely use case for an 8-core E1 would be a Power over Ethernet-driven wireless access device, or low-power 5G edge transport node. But higher core versions might be used as an engine for multi-port 100Gbps devices such as a firewall appliance.

Arm says that the design of both the N1 and E1 have been influenced by software-driven hardware design, where performance analysis and feedback from (for instance) cloud-native workloads has been used by the design teams to optimize performance. Specific applications optimized in this way include the NGINX web server, OpenJDK, MemcachedD, MySQL, Deepbench, the DPDK data plane development kit and Open vSwitch. There are also now about 100 open source projects and 25 standards organizations working around Arm architectures.

Strategy

The launch of Neoverse finally rounds out ARM's portfolio for datacenter deployments with a more complete base platform for its partners to build out systems and services. It helps that some key ecosystem elements have been developed out over the years so that they are now mature. Red Hat Linux, for instance, has been evolving since 2011, so it was available on day one when Amazon Web Services put up its first Arm instances at the end of last year.

It may be that AWS availability will be the tipping point, providing widespread, mainstream and instant availability

for the first time for testing and development purposes. The Neoverse launch also acts as a focus for the software ecosystem to fill in any remaining gaps that the datacenter and edge market will require.

For widespread deployments, more than just CPU IP is needed. There must also be multiple architectural and reference designs for specific use cases, verified IP blocks, electronic design automation platforms for the integration and taping out process, foundry partnerships and optimizations, compiler optimizations, standards and certifications (such as ServerReady), integrated environments in the shape of boards, servers and cloud deployments, and a developing open source and commercial software ecosystem.

Competition

Intel has previously looked unassailable in the datacenter chip market, but the slowing down of Moore's Law combined with difficulties in rolling out its 10nm process refresh and new interest in supplementing general-purpose CPUs with specialist accelerators have all combined to make it look more vulnerable.

Meanwhile, the very different requirements of the hyperscale and HPC companies, along with the extension of the cloud to the edge, have opened up brand new market opportunities that Intel won't necessarily dominate, as it has with enterprise servers and scale-out clusters for the past several decades. The reasons for choosing Arm cores over Intel might include their support of on-chip and off-chip heterogeneity, and the potential design freedom they can offer.

Arm will potentially displace other CPUs, as it is already been doing with SPARC in the HPC market (in the case of Fujitsu). Unlike Sparc, IBM's Power architecture is still seeing active investment, and has an evolving ecosystem based around industry standards. Its strength is at the very high end, potentially running demanding enterprise workloads in the cloud. MIPS, still strong as an embedded systems architecture, was recently merged with AI and automotive chip company Wave Computing. RISC-V, based on open systems cores, is gaining a lot of traction and interest recently, and has been adopted by both new silicon startups (such as Esperanto, InCore, Intensivate and SiFive) and large companies (Western Digital).

HPC is a particularly bright spot for Arm adoption, with some high-profile wins recently, including Astra, a collaboration with Hewlett Packard Enterprise, Sandia National Labs and the US Department of Energy that looks like it will be the world's largest Arm supercomputer, made up of 2,592 dual processor (Marvell/Cavium) servers. Integration flexibility with new interconnects and accelerators, along with its huge installed base at the edge, gives Arm a new opportunity to make inroads into datacenters as they are redesigned and reconfigured to address distributed intelligence.

The commonality of architectures between cloud and edge could be an advantage here. The advent of coherent scale-out architectures supporting heterogeneous computing – such as GenZ and CCIX – could further ease ARM's adoption in the datacenter. However, while Amazon has adopted Arm cores as the basis of its cloud chips, others – including Alibaba, Google and Tencent – have built their own custom designs. Google's TPU, for example, now also includes an edge version.

SWOT Analysis

STRENGTHS

Three recent developments have increased ARM's chances of building its infrastructure business to challenge Intel – endorsement by Amazon Web Services, the first availability of VMware, and the introduction of Huawei's new high-end Kunpeng chip.

WEAKNESSES

While extensive, particularly in phones and embedded systems, the Arm installed base has been fragmented due to the company's business model of licensing cores and letting its customers customize. Even today, standards in the software stack continue to evolve.

OPPORTUNITIES

The future of datacenters must be heterogeneous, and that gives Arm its biggest opportunity yet to expand its influence beyond smart phones and embedded systems.

THREATS

Intel has been fighting back with a series of acquisitions. Meanwhile, many hyperscalers have acquired their own silicon expertise and designing their own processors from scratch.