REPORT REPRINT

Intel turns heat to high with new server chips

APRIL 12 2019

By Daniel Bizo

The chipmaker has unveiled its latest generation server platform, code-named Whitley, which will be the foundation for its datacenter portfolio in the coming years. Intel designed Whitley for a dramatic increase in top processor performance via allowance for higher power consumption – much higher than ever before.

Introduction

Intel has unveiled its latest generation server platform, code-named Whitley, which will be the foundation for its datacenter portfolio in the coming years. Intel designed Whitley for a dramatic increase in top processor performance via allowance for higher power consumption – much higher than ever before. Indeed, the fastest models of the new server processor family, code-named Cascade Lake, will need all the power they can get to fend off a resurgent AMD and to match NVIDIA's claims for running deep neural networks. The cooling needs of the highest performance models will also challenge datacenter operators, most of which have never had to handle such power-hungry and thermally challenging chips at scale. With more power, further performance optimization features and support for persistent memory, the new Xeons represent a major jump even without new manufacturing technology or core design.

451 TAKE

Cascade Lake is a power demonstration from Intel in every sense of the word. The new processors should cement Intel's position in servers as a performance leader across a wide range of workloads even after AMD's upcoming launch of new products. The addition of persistent memory support shows the breadth of Intel's effort at its platform strategy and a novelty that may over time change how applications treat data. But what will largely define the extent of Cascade Lake's success is the speed at which customers will be willing to accept power-hungry models to realize the full potential of the new generation. A shift to liquid cooling is what Cascade Lake really needs to flex its muscles; something the datacenter industry has long resisted. However, if major cloud providers make the move, the rest will follow.

Context

Headquartered in Santa Clara, California, Intel is the world's biggest semiconductor vendor by revenue, largely due to its position as the predominant supplier of processors for PCs and servers. For 2018, it reported revenue of $70.8bn, a growth of 13% over 2017, and employed about 100,000 staff globally. The company's datacenter unit expanded by an impressive 21% due to demand from major IT and web service providers that outpaced expectations – Intel sold about 13% more server processors (and at 7% higher average prices) since service providers tend to buy more powerful chips that help them generate more revenue.

The view is not all rosy, however. Even though major cloud providers do not appear to be slowing down in their overall expansion anytime soon, each planning for tens of megawatts of IT capacity for 2019 and beyond, Intel conservatively guides for an overall flat 2019. The biggest setback of all was the prolonged delay of its next-generation semiconductor process. Dubbed 10-nanometer, the technology won't enter high-volume manufacturing before the second half of 2019 – at least two years behind schedule. (Even though Intel has some 10-nanometer production, it is limited to a handful of entry-level PC chips at small volume.)

The combination of semiconductor process complexity and aggressive areal density targets (to shrink the size of transistors' metal interconnects) probably contributed to the delays. In the meantime, contract manufacturers TSMC and Samsung Electronics are powering ahead with their technology roadmap in more incremental steps that position chip designers such as AMD, NVIDIA and Qualcomm to capture some share from Intel. AMD in particular is mounting a comeback after many years of near irrelevance in datacenters, which helped Intel dominate server processor sales.

Without new manufacturing technology or a major overhaul of its processor architecture, Cascade Lake was probably not planned to be a major milestone in Intel's original plans for this timeframe. Arguably, it meant to be a stopgap product that bridges the time between the previous generation (codenamed Skylake) and the upcoming 10-nanometer incarnation of Xeons still in the labs (Ice Lake) and due out in 2020. However, the combination of

silicon technology delays, growing competitive pressure and the longer than expected rollout of storage-class memory all converged around Cascade Lake and made it a processor generation of great importance to the health of Intel's datacenter business.

Technology

Intel still produces Cascade Lake on the company's 14-nanometer generation silicon technology but which the company has gradually optimized for higher performance in exchange for lower area density (bigger circuits) to compensate for the delay in the rollout of the 10-nanometer process. Cascade Lake appears to benefit from silicon technology and electrical optimizations to deliver frequency bumps. The new chips also use by and large the same processor architecture as the previous generation of Skylake chips – although there are some marginal improvements with significant effects that we discuss in more detail a bit later.

Marginal technology optimizations and architectural tweaks mean that the sole viable way for Intel to introduce major performance increases across the board is to increase the core count and allow for more power. This has happened in the past, however, not to the degree seen with Cascade Lake. In response to growing competitive threats from AMD, NVIDIA and a longer tail of chip design efforts that also include hyperscalers' in-house development, Intel decided to use multi-chip packaging and support thermal design power up to 400 watts per processor, up from the current envelope of 205 watts.

The extra power allowance means Intel can effectively double raw performance per processor over existing Xeons by fitting two full-speed Cascade Lake chips in a single processor package. This brings Intel's top-end model to 56 cores without giving up virtually any clock frequency compared with a single-chip 28-core variant. Cascade Lake also adds a new performance optimization feature called Speed Select Technology, which allows customers to assign different core performance levels to different applications that instructs the processor to allocate more power budget to cores that run latency-critical workloads.

Another key performance feature of the new chips is the acceleration of certain operations common to deep neural network inference, an emerging area of workloads that are widely expected to become a dominant force in cloud service provider infrastructure. Intel says that code that takes advantage of the new instructions (collectively called dlboost) can speed up by as much as a factor of 2 to 4 compared with Skylake Xeons.

Cascade Lake also adds official support for persistent memory modules (called Optane DC persistent memory) to its memory controllers, a feature also carried in Skylake but not commercially enabled due to some bugs in its implementation, according to Intel. Persistent memory allows applications to write and store data in memory without accessing the storage subsystem. This can bring substantial speedups to data-intensive applications such as database engines backing up online transaction processing systems or in-memory computing applications for inline analytics. Not only can applications that use memory persistence avoid the performance penalties of writing data to the storage subsystem, Optane memory modules also offer more capacity than their DRAM-based counterparts: much more (2-4x) data can reside in and be accessed from memory at near-DRAM speeds. Intel has been shipping Optane memory modules, albeit at low volume, since August 2018 for select customers (mostly cloud infrastructure providers) that participate in its early access program and have access to Cascade Lake too.

Strategy

Cascade Lake marks the start of a new phase for Intel's datacenter business. What it lacks in manufacturing technology change or architectural overhaul it makes up for in boldness and new additions. Top-end Cascade Lake models, which 451 Research expects to be in high demand from cloud providers and supercomputer users alike, represent an unprecedented jump in performance and power, ending an era of conservative steps and also highlights that the market has now accepted that power efficiency is not about low-power chips but rather the opposite.

These are drastic measures by historical standards and drastic measures can be seen as desperation. Indeed, Intel would unlikely have gone all the way to doubling its processor power envelope at one time if it weren't for the delay to its 10-nanometer technology and uncomfortable competitive pressure. But there is more to it than

a purely defensive play. With a much-expanded power envelope and the option for dual-chip packages, Intel is able to offer a wider range of customization options (number of cores, cache sizes, frequencies, thermal power) to cloud and telecom providers than before. Customization has grown to form the central part of Intel's datacenter business and growth. The company said that in 2018 half of its processors shipped to service providers were custom models.

With more performance and a much richer choice of custom configurations, Intel hopes to carry the momentum over from the strong uptake of Skylake chips in cloud infrastructure that helped its datacenter business expand by a fifth in size in 2018. It's a big ask but the performance density of Cascade Lake may be a big enough leap to trigger a strong refresh cycle against a now-sizeable install base of cloud and web infrastructure servers.

With Optane DC persistent memory, Intel strives to grab a bigger share of spending on servers (at the expense of DRAM content from third parties) but its objective is also to make its Xeon platform stickier than ever before. This puts Cascade Lake in a comfortable position where customers that want to adopt persistent memory have to also buy the new chips. The production of Optane memory modules is still just ramping up, the capacity for which Intel is still building as its joint venture with Micron is being dissolved in an unfriendly split, but if proves popular with customers it will, over time, help Intel fortify its midrange to high-end workload positions against AMD, which is set to launch its next-generation, 64-core server processor later in 2019.

A large memory footprint, many more cores with deep neural network acceleration, positions Intel better for AI-enhanced applications such as image classification, data analysis or speech generation, an area where NVIDIA has already gathered pace but other chip designers, including several startups, are developing offerings.

But probably the single biggest question is whether operators will be willing to accept 300- to 400-watt processors in their infrastructure to realize the full potential of Cascade Lake. Only a few years ago, 100 watts was still standard for processor power and 200-watt parts are still a somewhat recent development. Such a step increases power density and throws existing IT and facility infrastructure out of balance. Importantly, such thermal power calls for a move away from air cooling, the predominant form of cooling in today's datacenters, and toward more effective methods such as liquid cooling of chips with cold plates or total immersion.

Competition

In the datacenter/server market, Intel faces competition from AMD (the alternative licensed x86 supplier), which is now on the charge with its recently introduced new server architecture called EPYC. There is also some small but potentially growing competition from the Arm ecosystem (a nonvertically integrated ecosystem of suppliers and chipmakers that use Arm designs and leverage independent semiconductor foundries), and some competition from the IBM Power architecture in the high-end and HPC segments. NVIDIA has found a sweet spot for its GPU technology in high-performance computing and emerging deep-learning workloads.

Intel's competitive position is also affected by the execution of contract chip manufacturers; chiefly TSMC, the world's leading foundry, and to a lesser extent the semiconductor division of Samsung Electronics and GLOBALFOUNDRIES. TSMC and Samsung claim that their 7nm processes have entered production, with the latter detailing EUV progress that has dramatically reduced the number of pattern mask steps required. The short-term battle will be one of getting high-performance computing devices to market cost effectively. Intel is betting that the excellence in packaging that should lead to better manufacturability will give it an edge over those wrestling with yields at smaller geometries.

SWOT Analysis

STRENGTHS

Despite its setbacks, Intel still enjoys the benefits of a global network of chip factories that its rivals cannot match in cutting-edge manufacturing capacity. Its processor and system designs are highly advanced and are well understood by developers and customers.

WEAKNESSES

Intel runs a technically highly complex, rapidly moving and capital-intensive business that leaves relatively little room for error – missteps can cost billions, and the delay of its 10nm process may have cost it in unrealized sales. Its business model relies heavily on mass production of largely generic parts.

OPPORTUNITIES

The continued buildup of capacity at cloud and web service providers, as well as its new memory modules and silicon photonics connectivity products, should keep Intel's factories full for the coming years. There is no end in sight for explosive growth in data processing.

THREATS

Intel's repeatedly delayed transition to its next-generation manufacturing technology provides an opening for AMD and NVIDIA. Emerging data-intensive computing fields such as deep learning may also pave the way for rivals to challenge Intel's position in workstations and datacenters.