

## Centaur Demonstrates Industry's First High-Performance x86 SoC with Server-Class CPUs and Integrated AI Coprocessor Technology

*Certified MLPerf benchmark wins on latency performance (image classification in less than 330 microseconds)*

**Austin, Texas – November 18, 2019** – Centaur Technology revealed the technology behind its outstanding results on the MLPerf<sup>1</sup> inference benchmarks, which were officially certified on a development system for key customers and software developers. Centaur's first design with its new artificial intelligence (AI) technology combines eight new server-class x86 CPU cores with a 20 tera-operations-per-second (TOPS) coprocessor optimized for inference applications in server, cloud and edge products. Centaur's system-on-a-chip (SoC) technology allows users to save substantial cost and power over systems that require both an x86 host processor and external AI accelerators.

Linley Gwennap, Editor-in-Chief, Microprocessor Report, confirms Centaur's achievement, "Centaur Technology is the first to announce an x86 processor design that integrates a specialized coprocessor to accelerate deep learning. This coprocessor delivers greater AI performance than any CPU and frees the x86 cores to focus on general-purpose tasks that continue to require x86 compatibility."

Centaur Technology submitted audited results for four MLPerf inference applications in the Closed/Preview category and had the fastest latency score of all submitters for image classifier benchmark Mobilenet-V1. To demonstrate the flexibility of its architecture, Centaur was the only chip vendor to submit scores for GNMT, which uses a recurrent neural network for text translation of English to German. The results of all submissions are at the [MLPerf website](#). Centaur expects its MLPerf throughput scores to improve by up to 3X using its new software tool flow.

According to Vijay Janapa Reddi, Harvard professor and an MLPerf founder while on sabbatical at Google, "Centaur Technology has been a major contributor to the MLPerf initiative for over a year, and we were pleased to see a small company submit official results that can be directly compared to the industry leaders. Centaur's fast inference latency stands out as a key point of comparison."

Glenn Henry, an industry luminary from IBM and Dell before founding Centaur Technology almost 25 years ago, is the chief architect of Centaur's AI coprocessor. Glenn is well-known in the microprocessor industry for his early battles against bad benchmarks that allow marketing to distort performance. For MLPerf, Glenn offered this high praise, "This is one of the best benchmark efforts I've ever seen, and I'm a real skeptic of benchmarking."

Glenn Henry explains, “We set out to design an AI coprocessor with 50x the inference performance of a general-purpose CPU. We achieved that goal. Now we are working to enhance the hardware for both high-performance and low-power systems, and we are disclosing some of our technology details to encourage feedback from potential customers and technology partners.”

Eliminating the need to move data to an off-chip accelerator yields extremely low latency on inference tasks and enables new low-cost form-factors. The first x86 SoCs using Centaur’s AI technology will allow the best of both worlds by integrating 44 lanes of PCIe connectivity for those wanting to add external GPUs and AI accelerators for even higher throughput.

Centaur worked with Qvis (a video-security company) to port its x86-based network video recording (NVR) and video-analytics software to a prototype system that will be showcased in booth #751 at ISC East in New York on November 20th and 21st. According to the CEO of Qvis Labs, Wayne Heideman, “Bringing up our NVR software on a Centaur motherboard was a straightforward process, since our high-end systems already rely on x86 processors. At ISC East we will show the high-end video-analytics features that are possible with the high level of performance and integration that Centaur achieves with its new processor technology.”

Centaur’s x86-based AI coprocessor accelerates more than just image and video analytics. Al Loper, President of Centaur, explains, “Our scalable AI-technology platform fits into the software ecosystems for both x86 and the rapidly evolving AI infrastructure based on popular AI frameworks such as TensorFlow and PyTorch. Our long-term goal is to accelerate diverse AI applications for imaging, speech, text and other emerging use cases without customers needing to use specialized software tools. Centaur has been involved from the beginning with new industry initiatives such as MLIR to create a common software infrastructure.”

#### About Centaur Technology

Austin, Texas-based Centaur Technology is a small group of very talented engineers that have been designing AI accelerator technology for high-performance, low-cost x86-compatible microprocessors. Over the past 24 years, Centaur Technology has shipped 26 different x86-based designs with millions of units sold. More information is at [www.centtech.com](http://www.centtech.com).

*[1] MLPerf v0.5 Inference Closed/Preview audited submission, Sept. 2019. MLPerf name and logo are trademarks. See [www.mlperf.org](http://www.mlperf.org) for more information.*

**Press contact:** Paula Jones – [CentaurPress@centtech.com](mailto:CentaurPress@centtech.com) – 650-279-8997