World's First High-Performance x86 SoC with Integrated AI Coprocessor

Technology Announcement

CenTaur Technology

Who is Centaur Technology?

- www.centtech.com
- Based in Austin, Texas
- Founded 24 years ago by Glenn Henry, Chief Architect
  - Ex-IBM and Dell
- Developed 26 different x86-based designs with millions of units sold
- Now the first x86 vendor to introduce an integrated AI Coprocessor!

IT WORKS!

Official MLPerf[1] Scores
(Fastest MobileNet classifier latency – 330µs)

Demo at ISC East
Booth #751

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

Industry Challenge: Fast Inference Hardware Requires External x86 Host Processor

Corollary: Current x86 Processors need External Inference Acceleration for Efficiency

- Huge inference markets beyond hyperscale Cloud or low-power mobile/IOT
  - On-premises or private data centers for applications using security video, medical images, or other sensitive data (also reducing cost for bandwidth to Cloud)
  - "Edge Analytics Servers" run inference on multiple streams of data (cameras, IOT sensors, etc.)
  - Currently addressed by x86 PC/server hardware with GPU add-in card
  - Power consumption and peak performance less important than cost and form-factor
- New chips coming to market to meet exploding demand for fast/accurate inference
  - SoCs based on ARM (and soon RISC-V): designed for power-constrained applications that don't need x86
  - High-end accelerator chips for x86-based platforms: connect through PCIe/M.2 to compete with GPUs
  - x86 processors' new instructions for DL (ex: VNNI): requires many expensive CPU cores for specialized DL task
- External accelerators add latency, cost, power, board space, another point of failure, etc.
- But x86 CPUs aren't efficient: VNNI yields approximately 53fps/core on ResNet-50 in MLPerf[1]

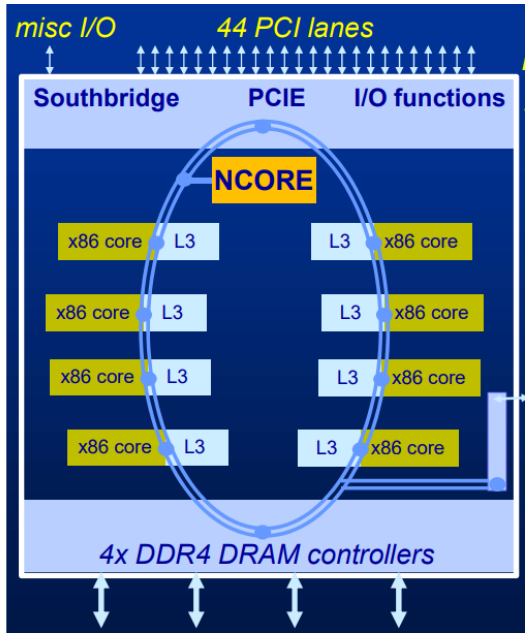
[1] MLPerf Inf-0.5-23. Dual Intel® Xeon® Platinum 9282 (112 total cores). Offline/Available ResNet-50 v1.5 (5965.62 fps) = ***53.26 fps/core****.*

CenTaur Technology

The Solution: Integrate a Dedicated AI Coprocessor with Fast x86 cores

"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."
Linley Gwennap, Editor-in-Chief, Microprocessor Report

Microprocessor Report will publish a deep-dive into the technical details on Dec 2nd, 2019.

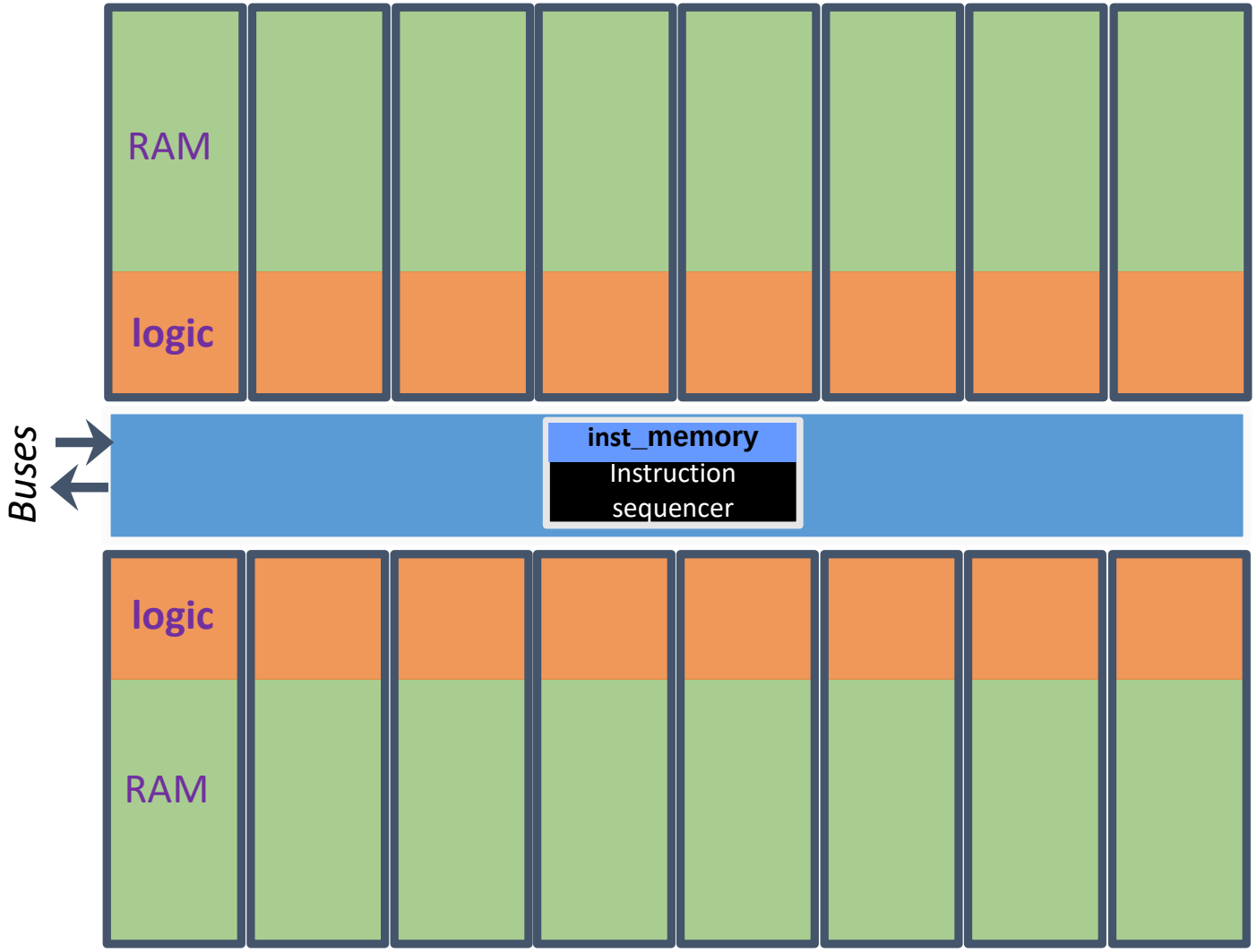


- Centaur developed a new x86 microprocessor with high instructions/clock (IPC)
  - Microarchitecture designed for server-class applications with extensions such as AVX-512
  - New x86 technology now proven in silicon with 8 CPU cores and 16MB L3 caches
  - SoC architecture provides an extensible platform with 44 PCIe lanes and 4 channels of PC3200
- Including AI coprocessor, requires less than 195mm^2 in 16nm TSMC
- Reference platform running at 2.5GHz today
  - Simultaneous execution of x86 cores and 20 TOPS AI Coprocessor
  - Delivers 20 peak terabytes/sec to AI Coprocessor from dedicated 16MB SRAM

Introducing the Centaur AI Coprocessor

- Centaur's internal code name: "NCORE"
  - SoC design is "CHA", and x86 core is "CNS"
- Designed from a Clean Sheet with goals for efficiency, scalability and flexibility
  - PLUS low latency!
- 32,768-bit very-wide SIMD architecture organized in vertical "slices" for easy size reconfiguration
  - SIMD made it easier to accelerate non-MAC operations (activation functions, etc.)
  - We at Centaur know how to make SIMD fast and efficient
  - Results for 4096 computations available in one clock! Hence: very low latency!
- Much better than huge array of MACs (used by many new chips)
  - Hard to do fast non-MAC things, not best latency, hard to scale
- Challenge: Memory
  - Many millions of bytes required to process an image in a video stream
  - Thus lots of fast (20 TB/s) internal RAM required, plus fast-as-possible access to DRAM and L3 cache
  - We have more dedicated memory per MAC than most (4 KB/MAC)

Current Version of Centaur AI Coprocessor



16 slices

X

256 bytes wide

= 4,096 bytes wide

All running at

2.5 GHz

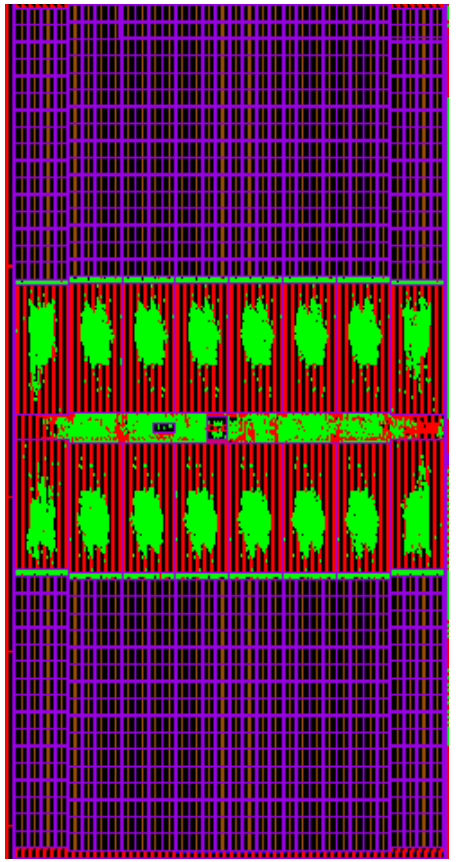
2,048 RAM "rows"

X

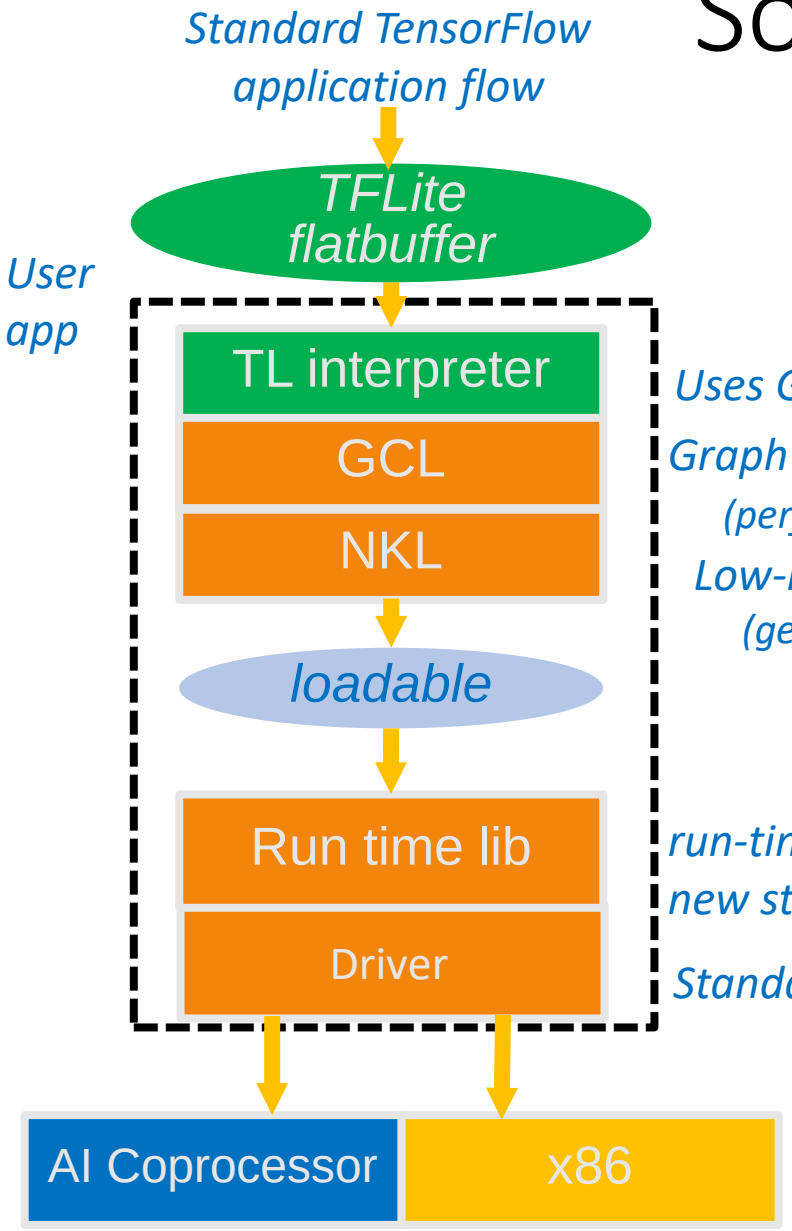
4KB/row x 2 RAMs

= 16MB of RAM

34.4mm² in16FFC



Software Tools for AI Coprocessor



Uses Google Delegate API

Graph Compiler Library

(performs data optimizations and DMA scheduling)

Low-level kernel library

(generates AI Coprocessor code)

Standard Linux driver (we run on Ubuntu)

In progress

Support for more frameworks: TF, PyTorch, etc.

Roadmap focused on utilizing MLIR for both graph-level and code-generation optimizations.

Significant optimization of x86 code done… & more to come

run-time includes some x86 optimized functions (AVX512), new streaming x86 insts, etc.

Delegate API for TFLite will require only 3 lines of code change in applications to support Centaur's AI Coprocessor. This was the flow used for MLPerf submission.

MLPerf Benchmarking

"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."
Vijay Janapa Reddi, Harvard University and MLPerf Inferencing Co-Chair

- Great benchmark effort with over 100 companies and universities
  - For over a year, Centaur helped inference working group define the benchmark
  - Rigorous, transparent, and audited methodology – Check out the v0.5 results and MLPerf white paper
- Centaur submitted to 4 of the 5 applications (Closed/Preview)
  - Software team had one month with working silicon, so MLPerf results not optimized
  - Centaur will publish unofficial results showing higher throughput; SSD MobileNet will go up by 3X
- As expected, big and expensive systems get higher throughput than a single 195mm^2 x86 chip
  - Google's 128 TPU v3's achieved one million frames/sec on ResNet-50!
  - Nvidia GPUs and start-up Habana score well, but require an external host Xeon® processor
  - Intel's new NNP-I has h/w equivalent of 12 NCORE blocks (but only 4.3X higher MLPerf throughput[1])
- These add-in cards will also work in a Centaur system (44 lanes of PCIe)
  - Best of both worlds with low latency of Centaur AI Coprocessor and high throughput of add-in cards

[1] MLPerf Inf-0.5-33. Dual Intel® Nervana™ NNP-I + 4116 Processor. Offline/Preview ResNet-50 v1.5 (10567 fps = 5284 fps/chip).

Highlights for Centaur's MLPerf Results

- MobileNet-V1 image classification
  - Centaur achieved the best MobileNet latency of any submission (330µs)
  - Throughput equivalent to 23.2 cores[1] of Intel VNNI; unofficial result will be over 6500 fps
- SSD MobileNet-V1 object detection
  - 1.54ms latency was within 140µs of fastest submission
  - Thoughput was throttled by lack of time to optimize; unofficial result will be over 2000 fps (3X)
- ResNet-50 V1.5 image classification
  - Latency was roughly 1ms per frame for a model requiring 25.5MB weights and over 4GMACs/image
  - Throughput equivalent to 22.9 cores[2] of Intel VNNI; unofficial result will be over 1400 fps
- GNMT text translation
  - Only chip vendor to submit results: proves architecture handles 263.3MB weights, RNNs and bfloat16
  - Unoptimized result is 12 sentences/sec (4 GMACs/sentence); unofficial result will be at least 30

*[1] MLPerf Inf-0.5-23. MobileNet-V1 on 112 cores (29203 fps) = 260.7 fps/core. It would **require 23.17 cores to match Centaur** (6042 fps)*
*[2] MLPerf Inf-0.5-23. ResNet-50 v1.5 on 112 cores (5965.6 fps) = 53.3 fps/core. It would **require 22.9 cores to match Centaur** (1218.5 fps)*

CenTaur Technology

Summary – World's First x86 Integrated AI Coprocessor!

- Paired with new high-performance x86 microprocessor design
- Official MLPerf results comparable to leading AI hardware companies
- Latency and cost advantage from integrating AI into x86 host
- Specialized AI Coprocessor more efficient than new x86 instructions
  - Requires 23 of Intel's world-class x86 VNNI cores to match Centaur Coprocessor
  - Saves power and cost while freeing up x86 cores for general-purpose tasks

- And IT WORKS! Demonstrating video analytics at ISC East

CenTaur Technology