Agenda

● GPU News
  ○ Nvidia Hopper GH100
  ○ Intel Arc
About Us

- Interested in building fast computing systems.
- Spanning across system software, compilers and architecture.
- Currently working with Prof. Hadi on compiling DL models for hardware accelerators and runtime software for heterogeneous systems.

Edwin Mascarenhas
About Us

● Like rock/blues music, running and travelling.
● In 2021: 48 runs, 203.3 kms and 18hrs,31 mins,10 secs.
● La Jolla Half marathon: 2h27m
● 8 countries visited so far.
● Top two artists/bands: Eric Clapton, The Eagles

Edwin Mascarenhas
About Us

- Interested in Computer and SoC Architecture.
- Recently developed an interest in hardware security.
- Currently working with Prof. Kastner on hardware security and verification.
- Like travelling, reading, new types of music, and most recently Formula 1.

Gandhar Deshpande
GH100

- Hopper Architecture
- Successor to A100 Ampere architecture chips
- Paired with NVIDIA Grace CPU to form the Grace Hopper Subsystem
- Architecture Whitepaper
GH100

- SM arch changes compared to A100
  - 6x faster: per-SM arch changes + higher clocks + additional SM count
GH100

- SM arch changes compared to A100
  - 6x faster: per-SM arch changes + higher clocks + additional SM count
  - 2x faster MMA on equivalent data types, 4x on FP8 as compared to 16-bit
GH100

- SM arch changes compared to A100
  - 6x faster: per-SM arch changes + higher clocks + additional SM count
  - 2x faster MMA on equivalent data types, 4x on FP8 as compared to 16-bit
  - 7x faster dynamic programming algorithms with DPX instructions (umm, how?)
GH100

- SM arch changes compared to A100
  - 6x faster: per-SM arch changes + higher clocks + additional SM count
  - 2x faster MMA on equivalent data types, 4x on FP8 as compared to 16-bit
  - 7x faster dynamic programming algorithms with DPX instructions
  - Thread block clusters to control locality at a larger granularity. Threads < Thread blocks < Thread block clusters < Grids
GH100

- SM arch changes compared to A100
  - 6x faster: per-SM arch changes + higher clocks + additional SM count
  - 2x faster MMA on equivalent data types, 4x on FP8 as compared to 16-bit
  - 7x faster dynamic programming algorithms with DPX instructions
  - Thread block clusters to control locality at a larger granularity. Threads < Thread blocks < Thread block clusters < Grids
  - SM-SM communication for loads, stores and atomics across multiple SM shared memory blocks
GH100

- SM arch changes compared to A100
  - 6x faster: per-SM arch changes + higher clocks + additional SM count
  - 2x faster MMA on equivalent data types, 4x on FP8 as compared to 16-bit
  - 7x faster dynamic programming algorithms with DPX instructions
  - Thread block clusters to control locality at a larger granularity. Threads < Thread blocks < Thread block clusters < Grids
  - SM-SM communication for loads, stores and atomics across multiple SM shared memory blocks
  - Tensor Memory Accelerator (TMA) unit to transfer large blocks of data efficiently
GH100

- SM arch changes compared to A100, 6x perf improvement
- Transformer engine
  - Software + Tensor core to accelerate transformer computations
  - At each layer analyze the statistics of the output values from Tensor Core
  - With next layer type + precision for next layer transformer engine decides target format to convert tensor to before storing to memory
  - Computes scaling factors computed from tensor statistics.
GH100

- SM arch changes compared to A100
- Transformer engine
  - Accelerate transformer model training and inference
- HBM3 memory subsystem
- 50MB L2 cache
- 4th gen NVLink, PCIe gen 5
- TSMC 4N process

<table>
<thead>
<tr>
<th></th>
<th>A100</th>
<th>A100 Sparse</th>
<th>H100 SXM5</th>
<th>H100 SXM5 Sparse</th>
<th>H100 SXM5 Speedup vs A100</th>
</tr>
</thead>
<tbody>
<tr>
<td>FP8 Tensor Core</td>
<td>2000</td>
<td>4000</td>
<td>6.4x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP16</td>
<td>78</td>
<td>120</td>
<td>1.5x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP16 Tensor Core</td>
<td>312</td>
<td>624</td>
<td>3.2x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BF16 Tensor Core</td>
<td>312</td>
<td>624</td>
<td>3.2x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP32</td>
<td>19.5</td>
<td>60</td>
<td>3.1x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF32 Tensor Core</td>
<td>156</td>
<td>312</td>
<td>3.2x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP64</td>
<td>9.7</td>
<td>30</td>
<td>3.1x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP64 Tensor Core</td>
<td>19.5</td>
<td>60</td>
<td>3.1x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INT8 Tensor Core</td>
<td>624 TOPS</td>
<td>1248 TOPS</td>
<td>3.2x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Intel Arc

- Alchemist GPUs
- Xe-HPG Microarchitecture
- HPG microarchitecture
Intel Arc

- Intel Xe Matrix Extensions (XMX) Engines for matrix operations for AI workloads.
- Ray tracing accelerator for Ray traversal, ray box intersection computation and ray-primitive intersections.
- XeSS to support AI enhanced supersampling.
Intel Arc

- Intel Xe Matrix Extensions (XMX) Engines for matrix operations for AI workloads.
- Ray tracing accelerator for Ray traversal, ray box intersection computation and ray-primitive intersections.
- XeSS to support AI enhanced supersampling.
- Accelerators for AV1 encoding & decoding
Intel Arc

- Intel Xe Matrix Extensions (XMX) Engines for matrix operations for AI workloads.
- Ray tracing accelerator for Ray traversal, ray box intersection computation and ray-primitive intersections.
- XeSS to support AI enhanced supersampling.
- Accelerators for AV1 encoding & decoding.
- Dynamic Power Share to Intelligently distribute power between Intel Core processors and Intel Arc graphics based on usage.
- Stream Assist: Offload streaming workloads from dedicated graphics to integrated graphics.
- TSMC N6 node
- Variable Rate Shading (At what granularity?)
RISC-V
A brief overview

Intel in RISC-V
RISC-V in AI
Intel in RISC-V

Intel has joined the RISC-V International with a Premier Membership.

Intel Foundry System to provide broad range of IPs based on RISC-V.

IPs will be optimized to make sure that IPs run best on IFS silicon across different segments.

Partner products manufactured on IFS technologies.

RISC-V cores licensed as differentiated IP.

Chiplet building blocks based on RISC-V, leveraging advanced packaging and high-speed chip-to-chip interfaces.
SiFive and BrainChip partnering to deploy ML applications on the edge.

SiFive provides multicore, multicluster solutions with leading performance and efficiency for ML and AI workloads.

They will now integrate Akida, BrainChip’s advanced neural network processor into their chips for accelerated performance on the edge.

Akida mimics the human brain to analyze essential sensor inputs and processes the data with unparalleled efficiency and energy economy.

RISC-V and AI

Esperanto Technology’s AI/ML solution in initial evaluation stages.

Esperanto makes chips for accelerated AI inferencing. First evaluation shows steady performance and near-linear scaling across different configurations of AI clusters.

It harnesses over 1000 RISC-V cores to achieve great performance.

RISC-V and AI

Alibaba Cloud’s Xuantie C906 processor attained firsts in MLPerf Tiny v0.7, an AI benchmark focusing on IOT devices.

Excelled in 4 categories: visual wake words, image classifications, keyword spotting, and anomaly detection.

The Xuantie C906 is Alibaba’s custom-built processor based on the RISC-V instruction-set architecture.

Alibaba used an optimized software stack, the hardware toolset and library and has optimised AI operators

Article Link
Quantum computing
ARM CHERI chip
Microsoft Azure ARM VMs
AMD to acquire Pensando
Intel BlockScale
Intel and Quantum Computing

Intel and QuTech (TU Delft and TNO) have collaborated to create the first silicon qubits at scale in Intel’s Oregon facility.

10000 arrays with several silicon-spin qubits on a single wafer with yield > 95%.

Process uses same equipment as the one used for Intel’s latest CMOS technology.

Proves that Quantum computing may not require an entirely different manufacturing process and strategy.

A new hardware focused on secure computing.

Stands for Capability Hardware Enhanced RISC Instructions

Makes computing safe for memory-unsafe languages like C and C++

Changes the way memory is accessed in hardware and changes are fundamentally designed in silicon.

Integrates with MMUs to ensure that memory accesses are done using capabilities which store permissions, not just integer addresses.

Deterministically mitigates vulnerability classes rather than just detecting high probability scenarios.

ARM has shipped chips for testing and evaluation for various use-cases.

https://msrc-blog.microsoft.com/2022/01/20/an_armful_of_cheris/
Azure VMs ARM’ed Ampere Altra chips

Azure VMs preview with Ampere Altra Arm-based processor

VMs outperform equivalently sized Intel and AMD instances from the same generation by 39% and 47%, respectively on Spec CPU 2017
**AMD to acquire Pensando**

Pensando, a distributed services provider is to be acquired by AMD for their data center solutions.

Pensando brings a high-performance packet processor and a system software stack to the table.

Enhances the AMD product line in datacenters with software that provides 8x to 13x performance boost.

Solutions from Pensando are already deployed at scale by all major data center providers.

Expected to finish the acquisition by end of Q2 2022.

Intel BlockScale

Intel BlockScale ASICs to start shipping from Q3 2022.

Given the requirement of decentralization and distribution of system, compute requirement for blockchain based systems increasing rapidly.

BlockScale to provide low power consumption and efficient hashing solutions for companies like Argo Blockchain, Block Inc., Hive Blockchain Technologies and GRIID Infrastructure

Manufacturing of these chips should be able to continue without compromising CPU/GPU supply.

Thank you!