Study Notes: The New Golden Age of Computer Architecture

I. Power Consumption and the Dark Silicon Problem

- **Power Density:** $P_{dynamic} \sim \alpha \times C \times V^2 \times f \times N$; $P_{leakage} \sim N \times V \times e^{-Vt/2}$
  - α: Average switches per cycle
  - C: Capacitance
  - V: Voltage
  - f: Frequency (linear with V)
  - N: Number of transistors
  - V_t: Threshold voltage
- **Moore's Law:** Transistors get smaller, allowing higher frequencies.
- **Dennard Scaling Discontinued:** Voltage cannot be lowered further.
- **Dark Silicon Problem:** Given a fixed area and power budget, only a portion of the chip can operate at full speed. The rest remains "dark" (powered off) to stay within the power constraints.
  - **Goal:** Maximize efficiency per chip within power constraints.
  - **Example:** A chip with some cores running at top speed while others are off vs. all cores running at a lower frequency.

II. Solutions to the Dark Silicon Problem

- **Frequency Adjustment:** Aggressively adjust the frequency of processor cores based on demand.
  - Run compute-intensive programs at full speed, others at lower frequencies or turn them off.
- **Many-Core Processors (GPUs):** Improve throughput per chip through massive parallelism, with each processing element operating at a lower speed.
  - Not ideal for latency-sensitive workloads.
- **Heterogeneous CMPs:** Balance trade-offs for general-purpose workloads (e.g., big.Little cores).

III. GPU Architecture

- **Core Components:** L1 instruction cache, L1 data cache/shared memory, TEX units, warp scheduler, dispatch unit, register file, INT32/FP32/FP64 units, MXU/Tensor Cores.
- **Vector Processing:** Each core acts as a vector processing unit.

IV. The Rise of Hardware Accelerators (ASICs)

- **Motivation:** Address the limitations of general-purpose processors and GPUs for specific tasks.
- **Advantages:**
  - Better performance-per-watt.
  - Better performance-per-area.
- **Disadvantage:**
  - Worst programmability compared to CPUs/GPUs.
  - More likely to be memory-bounded.
- **AI/ML Accelerators:**

    - Reduced functionality, focusing on matrix multiplications.
    - Large, high-throughput on-chip memory.
    - RISC-style instruction set architecture.

V. Tensor Processing Units (TPUs)

- **Google TPUs:** TPU v1, v2, v3, v4, v5e, Edge TPU, Tensor G1/G2/G3.
- **Matrix Processing:** Core of AI/ML accelerators.
- **Datapath Specialization:** Local & optimized memory, parallelism, reduced overhead.
- **Programming Interface:** TensorFlow (high-level mathematical functions).

VI. Programming in Turing Architecture (NVIDIA)

- **Tensor Cores:** Use tensor cores for 16-bit operations.
- **Example:** `cublasGemmEx` for matrix multiplication.
- **Performance:** Tensor cores can be significantly faster due to performing multiple operations in one cycle.

VII. Data Movement Overhead

- **Challenge:** Data movement between CPU, GPU, and memory can be a bottleneck.
- **Example:** Copying data to and from GPU memory using `cudaMemcpy`.
- **PCIe Speed:** The speed of PCIe can limit performance compared to the computational speed of GPUs.

VIII. SIMD2: A Generalized Matrix Instruction Set

- **Concept:** A unified instruction set for various matrix operations beyond GEMM.
- **Operations:** Matrix multiplication, all-pair shortest path, critical path, minimum/maximum reliability paths, minimum spanning tree, transitive closure, L2 distance.
- **Integration:** Can be integrated into GPU cores alongside Tensor Cores.

IX. Ray Tracing for Sparse Matrix Computation

- **Technique:** Mapping sparse GEMM as ray tracing.
- **Steps:**
    1. Matrix B as scene objects.
    2. Matrix A as rays.
    3. Ray-object intersection test.
    4. MAC on vector cores.

X. Conclusion: The New Golden Age

- Computer architecture is crucial in the dark silicon era.
- Hardware specialization and domain-specific architectures are becoming increasingly important.
- Need for system-level programming to utilize new ASICs effectively.

Glossary

- **ASIC:** Application-Specific Integrated Circuit - a chip designed for a particular use.

- **CMP:** Chip Multiprocessor - a single chip containing multiple processors or cores.
- **Dark Silicon:** The phenomenon where, due to power constraints, not all transistors on a chip can be powered on simultaneously.
- **Dennard Scaling:** The historical trend where as transistors got smaller, power density remained constant. This scaling has ended.
- **FPGA:** Field-Programmable Gate Array - an integrated circuit that can be configured by the user after manufacturing.
- **GEMM:** General Matrix Multiplication.
- **GPU:** Graphics Processing Unit - a specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device.
- **ISA:** Instruction Set Architecture - an abstract model of a computer, defining how software controls the hardware.
- **RISC-V:** An open standard instruction set architecture (ISA) based on established reduced instruction set computer (RISC) principles.
- **SMT:** Simultaneous Multithreading - a processor design that allows multiple independent threads of execution to better utilize the resources provided by modern processors.
- **TPU:** Tensor Processing Unit - a custom-developed accelerator chip by Google designed specifically for neural network workloads.