Technical Snippet

New Computing Resources

Many businesses got their heads cracked when semiconductors approach the wall of physics. As the recent advent of Generative AI showed, powerful AI models need an immense amount of computing and memory resources. Computing is indeed a factor for business success or failure. How do we get extra computing performance whilst hitting the wall of physics?

Vision. The recent rise of Nvidia to the world's most valuable company is not without technological reasons. It is not because its CEO is a very good stage performer at all. The reality is that the cost of computing by GPU has improved faster than Moore's Law applied to CPU developments over the last 50 years. Although GPU is not suitable for all applications, its meteoric success does give us the vision that it is not impossible to obtain extra computing power around the wall.

> Solution 1. A GPU may have thousands of simple computing cores doing parallel computing on multiple data. Can we have thousands of cores in a CPU? A CPU can compute a bigger range of applications than GPU due to its complex circuits. In theory, we can have more cores in a CPU but not as many as in a GPU. Sounds fair? A desktop with 6 cores and 12 threads in a CPU is commonplace these days (such as Compucon Diamond Plus based on i5-12400). More cores can be obtained by linking them on the same wafer during manufacturing to give multiples of 2, 4, or 6 cores for instance. A well-known example of such technology is Chip-on-Wafer-on-Substrate (CoWoS) deployed by TSMC in Taiwan. It has been applied successfully since 2023 on 3nm semiconductors close to the wall of physics. A startup company called Ampere Computing in California has announced to produce a CPU with 512 non-complex cores in 2025/26 timeframe.

COMPUCON
SYSTEMS DIVISION

Article first published 2024-09
Compucon New Zealand
234 Bush Road, Albany, Auckland
www.compucon.co.nz; www.cnz.co.nz
editor@compucon.co.nz

Solution 2. The impressive performance of GenAI models such as Copilot and Llama has made vendors think of democratizing the models so that an end-user can get instant help from the models in real time. This requires a change from cloud computing to edge (or local) computing. In order to further reduce the capital and energy cost of computing on the desktop, model inferencing can be done with 8-bit or less integer data in lieu of 16 or 32-bit floating point data. Computing performance will speed up by an order of magnitude easily. Development efforts have started. Note the adverb easily above.

Solution 3. Computing time depends on the processors (CPU and GPU), memory, and data storage. Solid state disk (SSD) without mechanical rotations that enable HDD has come out for over a decade and has helped speed up overall computer performance. A further step of speed up comes from Non-Volatile Memory Express (NVMe) SSD. The performance gain comes from sitting closer and talking directly with the CPU. Users can tell the speedup of overall performance over SSD easily (again). It has been shipping for over 2 years since.

One may say one tide floats all boats. It is true in general but not applicable in applications of technology because some boats are smarter than others. People who understand the solutions always do better than those who don’t. Smart businesses should leverage the availability of technology smartly (meaning correctly and timely).

COMPUCON
SYSTEMS DIVISION

Article first published 2024-09
Compucon New Zealand
234 Bush Road, Albany, Auckland
www.compucon.co.nz; www.cnz.co.nz
editor@compucon.co.nz