Moore's Law Redefined for AI/HPC

Katharine Schmidtke, Hans-Juergen Schmidtke
Eribel Systems LLC, 1250 Grant Rd., #311, Mountain View, CA 94040, USA
kschmidtke@eribelsystems.com, hjschmidtke@eribelsystems.com

Abstract: Examining the impact of AI workloads on system performance, we reapply Moore's law at the system level to uncover the implications for photonic components and the drivers that will propel the photonic industry forward.

1. Moore's Observation

Moore's prediction in 1965 for the relationship between transistor density over time has been used to guide the semiconductor industry ever since. The success of Moore's guidance was that it was ambitious and realistic at the same time and provided the controlling "tact frequency" for growth rate in a highly complex integrated circuit supply chain which became a self-fulfilling "law". Today, a doubling of performance every two years has become the default standard by which performance improvements are measured and future roadmaps are determined.

Moore's law has been re-expressed over time. Originally focused on transistor count and then density, it evolved to encompass logical operations (instructions per second) per chip. The doubling of semiconductor performance every two years has continued for over four decades of technology development. It is expected that innovations in system performance will continue to propel performance growth forward at the same rate. More recently Moore's law has driven innovation beyond just the transistor-level, with the incorporation of new multi-chip-module technologies known as 2D, 2.5D, and 3D.

In this work we discuss the impact of AI/HPC workloads and the growth in model size on the required system performance. By redefining a version of Moore's law at the system level we explore the implications for photonic components that will be needed to support system growth for AI/HPC clusters. Re-applying Moore's guidance in this context, we analyze the drivers that will propel the photonic industry and explain the technology requirements that will be needed for the next decade.

2. System Performance and Moore's Law

System performance is an important performance metric because it is what the end user experiences. Before 2004, system performance improvements were mostly dominated by the regular improvements in transistor density, in line with the original expression of Moore's law, and an increase in chip clock-frequency according to Dennard scaling [1,2]. These advances in silicon technology directly benefited performance at the system level. Systems designs therefore focused on incorporating these exponential improvements rapidly through flexible designs, and increased programmability.

This led to rapid increase in the performance of large computer systems as can be seen in the performance charts shown in Figure 1. [3]. Since the ending of Dennard scaling around 2004 [1,2] the performance increase of the individual components due to clock-speed slowed, but was compensated for by the introduction of multi-core chip architectures allowing system performance to keep improving. As can be seen in Figure 1., the performance of HPC systems consistently grew at the rate of 3.3 times every 2 years until around 2012 when the growth rate reduced to 2.2 times every 2 years. This can also be explained by a limitation of available memory bandwidth known as the "memory wall" [4,5] and interconnection transfer rates which were only scaling up at 1.4 times every 2 years. System designs were forced to scale out to wider parallel interconnections to add additional memory and interconnect compute cores. This increasing disparity between the growth in computer performance of HPC systems and GPU's in comparison to the growth in memory bandwidth and interconnection bandwidth can be seen in Figure 1.

From a system perspective, Moore's law does not fully capture the performance growth because it does not capture the other design factors which are equally important in a system such as; system size, power consumption, cost efficiency, I/O bandwidth, and the control aspects. Full system design includes the on-chip memory buses, transport performance, as well as the I/O off-the-chip, and the connection between chips. These key factors are perhaps even more critical because since the 1980s, the movement of logical bits has consumed more energy than the processing of the bits themselves.

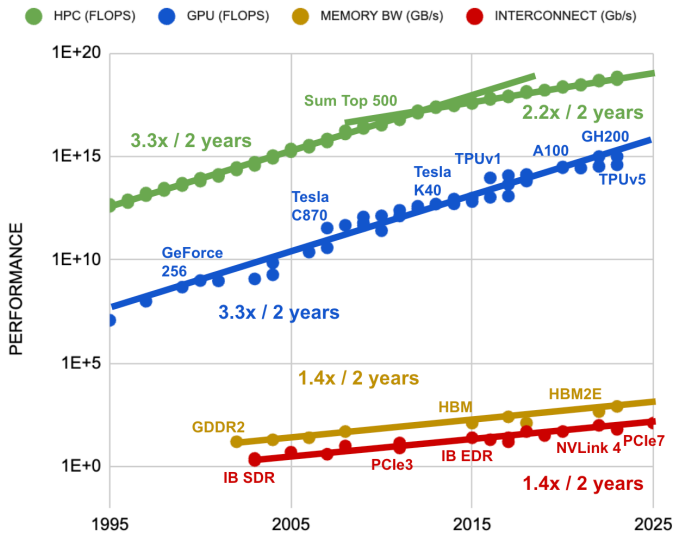

Fig. 1. Performance versus time for: the sum of TOP500 supercomputers in floating point operations per second (FLOPS), GPU and TPU peak performance in FLOPS, memory bandwidth in GB/s for generations of GDDR & HBM, and interconnection bandwidth in Gb/s for PCIe, IB, & NVlink.

3. Impact of AI/ML Workloads

A discussion of AI/ML workloads and its impact on system performance is highly complex and dependent on many factors including the algorithms, software-models, software-compilers, and specific hardware architecture used.

Large Language Models (LLM) and Deep Learning and Recommendation Models (DLRM) are among the most used by hyper-scale datacenter companies. The size of these AI models has been increasing at approximately ten times per year, far outpacing the increase in compute performance [5,6]. To accommodate the large size of models, the computer chips are optimized for the particular needs of the model type, focusing on specific requirements for the accelerators (xPU/GPU), dedicated network interface (NIC), and memory interconnect (CXL/PCIe, HBM's, DDR's). In addition, a unique architecture design is used for the workload which is referred to as Domain Specific Architecture (DSA) [5]. In this case the orchestration, control, and interconnection of components becomes a design tool to improve the system performance. This is referred to as the co-design of AI hardware and software.

It's clear that interconnects play an important role in AI/HPC system design because of effects like the 'memory wall' [4,5,7]. For low operational-intensity workloads (i.e. those with low compute to memory ratio) cluster designs are often memory bandwidth limited rather than compute-limited [5,8].

4. Consequences for Optical Interconnects

The trends in system performance of both AI and HPC systems drive the requirements for all the components including optical interconnects and therefore provide a useful predictor for interconnect requirements. Assuming AI clusters grow at similar rates to those of HPC and GPUs, this would indicate AI cluster performance improvement of 2 to 3 times every 2 years. An analysis of growth rates in switched network architectures shows a similar behavior with switching capacity increasing by a factor 2 every 2 years. Since interconnect bandwidth has historically scaled up at a rate of only 1.4 times every 2 years (equivalent to doubling every 4 years), then systems must compensate for the factor (2/1.4) shortfall by also scaling out at a rate of 1.4 times every 2 years (doubling the number of interconnects every 4 years) to achieve a total interconnect capacity growth of 2 times every 2 years.

AI/HPC systems consist of several types of interconnections of which the most important are listed here:

1. Scale-out at the system level is achieved by interconnecting the compute-nodes through a switched network (today up to 1 Tbit/s/GPU). Optical fiber technologies are applied today. Depending on the network architecture optical reaches of up to 2km might be needed [9].
2. The back-end network manages tightly interconnected communication between compute-cores that are collectively working a compute-problem. Sometimes that is also referred to "scale-up". This is a highly

interconnected network (currently around ~4-8 Tbit/s/GPU) within a rack. As the domain of the GPU's that tied to a scale-up network is expected to grow, optical interconnects will become necessary for this type of network in the next few years. Optical technologies are viewed as an enabler for future large scale-up domains, but are still relatively power inefficient in comparison to copper for short reach (<6m) [9]

3. CXL/PCIe switches that interconnect memory, CPU and GPU's. Today these are implemented with PCIe copper-based technologies.
4. Memory bandwidth between HBM and chip (currently up to 38.4Tbit/s/GPU) with short-reach and wide-bus interconnections [10].

The requirements on front-end and back-end networks have different optimization points in terms of bandwidth, reach, and latency. Scale-up networks today are typically located within a shelf or rack with very high bandwidth (several Tbit/s) and currently use copper connections. In order to address the opportunities for much larger scale-up designs for GPU's, optics will need to be delivered with ~10Tbit/s across several racks (~20m) at increased manufacturing volumes.

Over the past decade the scale-out network switch capacity and the I/O capacity has grown by increasing the number of layers of switches required to connect all the GPU's. Since switch capacity will be limited, as the network grows, additional layers will be required to support the end-to-end flows. These additional switch layers will require interconnect and therefore increase the total number of interconnects further. The exact number of additional interconnects that are required depends on the type of switch architecture used and the number of switch layers.

5. Conclusion

Anticipating a doubling of HPC systems and AI cluster performance every two years, in line with Moore's law, can provide guidance to the photonics industry. If historical growth of interconnect bandwidth can continue to double every four years, then the number of interconnects needed to scale out the systems will also need to double every four years to achieve a combined doubling of capacity every two years. This will dictate new requirements for interconnects in the following areas:

1. Increased capacity of data connections (particularly GPU-networks for scale-out networks)
2. Number of connections will increase the need for volume manufacturing techniques
3. Increased length or reach of each link as the system size grows
4. Miniaturization along edge or beach-front to increase I/O density
5. Improved energy efficiency as power is a limiting factor
6. Increased cost pressure as volume increases

Many of the above requirements are familiar to the photonics industry. Photonics will play a critical role in the growth of HPC systems and AI clusters to provide the data movement in larger and more complex systems. Moore's law can provide guidance for future roadmaps including; data-rate, miniaturization, energy efficiency, cost reduction, and manufacturing volume.

6. References

[1] J. Shalf and R. Leland, "Computing beyond Moore's Law." Computer. 48. 14-23. 10.1109/MC.2015.374, (2015)

[2] Dennard Scaling, https://en.wikipedia.org/wiki/Dennard_scaling

[3] https://www.top500.org/

[4] Dr. Mark Liu, TSMC Keynote ISSCC https://research.tsmc.com/assets/download/Chairman_2021_ISSCC.pdf (2021)

[5] John L. Hennessy, David A. Patterson, "Computer Architecture - A Quantitative Approach", (2019)

[6] https://ourworldindata.org/grapher/artificial-intelligence-number-training-datapoints economist.com, (April 2023)

[7] Wikipedia "Random access memory" https://en.wikipedia.org/wiki/Random-access_memory#Memory_wall accessed (2023)

[8] Vivienne Sze et al, "Efficient Processing of Deep Neural Networks", Springer (2020)

[9] Madeleine Glick, Ling Liao, Katharine Schmidtke, "Integrated Photonics for Data Communication Applications" Elsevier (2023)

[10] Nvidia https://www.nvidia.com/en-us/data-center/h200/