Parallel Computing 64 (2017) 65–80

Contents lists available at ScienceDirect

Parallel Computing

journal homepage: www.elsevier.com/locate/parco

Optical interconnects for extreme scale computing systems

Sébastien Rumley[a,*], Meisam Bahadori[a], Robert Polster[a], Simon D. Hammond[b], David M. Calhoun[a], Ke Wen[a], Arun Rodrigues[b], Keren Bergman[a]

[a] *Columbia University, New York, NY, United States*
[b] *Sandia National Laboratories, Albuquerque, NM, United States*

ARTICLE INFO

Article history:
Received 31 July 2016
Revised 4 December 2016
Accepted 8 February 2017
Available online 14 February 2017

Keywords:
Exascale
High performance computing
Optical interconnects
Interconnection networks
Silicon photonics

ABSTRACT

Large-scale high performance computing is permeating nearly every corner of modern applications spanning from scientific research and business operations, to medical diagnostics, and national security. All these communities rely on computer systems to process vast volumes of data quickly and efficiently, yet progress toward increased computing power has experienced a slowdown in the last number of years. The sheer cost and scale, stemming from the need for extreme parallelism, are among the reasons behind this stall. In particular, very large-scale, ultra-high bandwidth interconnects, essential for maintaining computation performance, represent an increasing portion of the total cost budget.

Photonic systems are often cited as ways to break through the energy-bandwidth limitations of conventional electrical wires toward drastically improving interconnect performance. This paper presents an overview of the challenges associated with large-scale interconnects, and reviews how photonic technologies can contribute to addressing these challenges. We review some important aspects of photonics that should not be underestimated in order to truly reap the benefits of cost and power reduction.

1. Introduction

The performance of supercomputers has experienced a steady growth in the last decade and has mainly kept up with the intensive computation needs of advanced scientific studies. Modelers and scientific computing programmers have so far found ways to leverage the computing power made available to them to more accurately simulate complex systems. It is therefore no wonder that demands for even greater capabilities are sought to further advance research and/or open new directions. On the hardware side, the research community and industry have been successful at designing computer architectures and components that have been able to deliver a continuous growth in performance, and to fulfill the needs of majority of researchers, as shown in the Top 500 trend lines [1].

Looking at the last four years more specifically, however, one discerns a change in these trends. For the first time since the ranking was established, a top supercomputer (Tianhe-2A) stayed unchallenged for two and a half years, since its introduction in the spring of 2013 until the recent introduction of the new world leader, Sunway TaihuLight. This deceleration is apparent when one examines the sum and #500 trend lines, and even more strikingly clear if the comparison is restricted to the leading 20 Supercomputers (Fig. 1). Despite the recent introduction of Sunway TaihuLight, the steady growth observable between 2009 and 2013 has been discontinued.

[*] Corresponding author.
E-mail address: rumley@ee.columbia.edu (S. Rumley).

http://dx.doi.org/10.1016/j.parco.2017.02.001
0167-8191/© 2017 Elsevier B.V. All rights reserved.

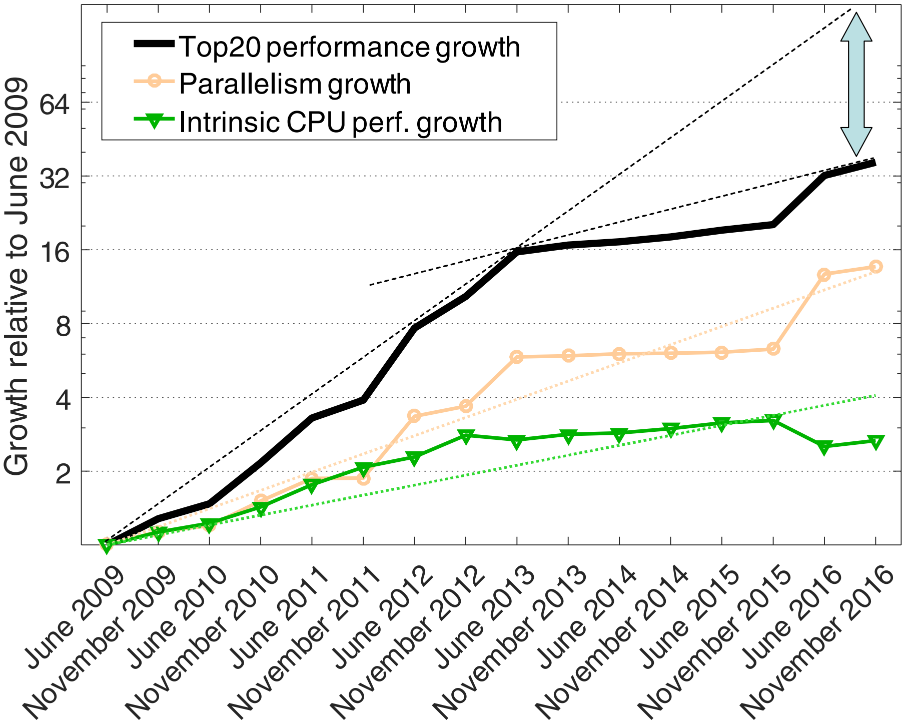

Fig. 1. Relative evolution of the average of the 20 most powerful Supercomputers in the last years, in terms of total compute power growth (thick line), CPU parallelism growth (circles) and CPU compute power growth (triangles). Normalized to June 2009.

We also observe on Fig. 1 that the compute power per CPU (measured in FLOPs—floating point operations per second, and averaged over the top 20) has significantly stalled in recent few years. One can conclude that the era during which clock speed and CPU architecture provided the key performance drivers has clearly come to an end. Beyond this point, we can only count on parallelism to reach new frontiers, in particular the one of realizing an Exascale computer.

Parallelism, if realized “simply” by aligning more discrete components, has in turn a severe impact on costs. Funding sources are typically under stringent restrictions with limited scalability [2]. As a consequence, smarter parallelism must be achieved. To limit component purchase expenses as well as system assembly costs, large efforts have been invested in maximizing the FLOPs delivered by a single Chip Multi Processor (CMP). As shown in Fig. 2, this has resulted in stability with respect to the number of nodes. Inspired by Graphical Processing Unit (GPU) architectures, and building upon the progresses made in transistor downscaling, a number of vendors have already been successful at packaging several TeraFLOPs[1] into a single chip and its associated ≈ 200–300 W power envelope and ≈ 300 mm^2 area [3]. Further steps in this direction should allow computing chips with total compute power of tens of TeraFLOPs to emerge in the coming years. Reducing the cost of compute silicon is, however, only one part of the challenge. Bulk computing power needs to be surrounded with appropriate memory resources, but also adequately interconnected to become a true supercomputer. Each FLOP available on a chip must be associated with enough bandwidth for synchronization and data exchanges with other cores and nodes.

In this paper, we focus on these interconnect resources. With the growing level of parallelism, one could expect them to increase, i.e. expect the bandwidth made available to each FLOP (expressed in byte/FLOP—B/F hereafter) to grow alongside the FLOPs, or to the least remain constant. However, looking at the top supercomputers from the last few years, the opposite is observed (Fig. 3). To obtain this curve, we collected the number of nodes present in the top 10 supercomputers in the last 7 years and calculated the FLOPs per node. We also collected performance measurements of their interconnection systems (Table 1), both at the physical transmission level and at the MPI (Message Passing Interface) level—unidirectional MPI bandwidth. We finally combined these measurements to obtain the B/F figure. The MPI bandwidth available to the nodes in the top 10 progressed by a factor of 3.3× from 2010 to 2016. In 2010, the typical interconnect was based on QDR Infiniband (4 instances in the Top 10) providing 3.4 GB/s. In 2016, the Cray Aries DragonFly is the typical interconnect, offering 9 GB/s (4 instances). The top two systems in the June 2016 ranking have custom interconnects providing ≈ 12 GB/s.

[1] If CPU architectures are considered, several TeraFLOPs-on-a-chip are obtained by aggregating 50–100 cores each providing 10–40 GigaFLOPs. In GPU architectures, several thousands of cores each providing 0.5 GigaFLOPs are aggregated. These GPU cores, however, are generally organized in clusters within which hundreds of cores share register files and caches. These clusters, providing around 100 GigaFLOPs, are sometimes regarded as cores (e.g. by the top500 ranking).