

Limitations of performance of Exascale Applications and supercomputers they are running on

JÁNOS VÉGH

Abstract

The paper highlights that the cooperation of the components of the computing systems receives even more focus in the coming age of exascale computing. It discovers that inherent performance limitations exist and identifies the major critical contributions of the performance on many-many processor systems. The extended and reinterpreted simple Amdahl model describes the behavior of the existing supercomputers surprisingly well, and explains some mystical happenings around high-performance computing. It is pointed out that using the present technology and paradigm only marginal development of performance is possible, and that the major obstacle towards higher performance applications is the 70-years old computing paradigm itself. A way to step forward is also suggested.

Keywords: processor single-threaded performance, Amdahl's law, supercomputer application efficiency, supercomputer performance limitation, cooperative computing, explicitly many-processor approach, extending the computing paradigm

1. Introduction

In the age of exascale applications new challenges for the application writers and users appear. The supercomputers –and especially the exascale ones– are strongly custom-made, rather than general purpose computers. The cooperation of the components will be as crucial as it was never before, so the users and the programmers of the exascale applications –as well as the architecture designers– must exactly understand how the strongly parallelized and distributed computing system works. Otherwise minor imprecisions, a poorly organized cooperation or a small amount of poorly organized code can result in catastrophics performance losses.

With the expected and wanted advent of exascale computing a lot of application developments has been started practically in all fields of science, military, industry, services, etc. The increased interest is partly motivated by real economical needs (as proved by the growing number of industry-hosted supercomputers), by the need of elaborating "big data", by the increased importance of computer modeling or providing platform for applications utilizing artificial

intelligence, etc. In contrast with those "commodity supercomputers", the "racing supercomputers" are mainly motivated by the prestige value of the better position on the list of high-performance supercomputers, "green" supercomputers or total supercomputer capacity. Of course the acquired experiences and developed technologies can be effectively utilized later in the "commodity supercomputers" as well. The ongoing race between institutions, nations, companies, processors, interconnection and memory access methods, etc. resulted in a kind of "gold rush". The computing performance quietly approached its technological bounds, but the "computing stack" has not been revised, in the past 70 years.

It is well known, that computing has its numerous limitations and even those limitations are limited [1], as well as that the computing growth (like Moore's observation) is only initially exponential [2]. It is also a common experience on all fields that when approaching some extremities (big masses, small sizes, big sizes, high speed), the behavior of the studied subject drastically changes. Followers of the high-performance computing field faced strange experiences, like aborting projects in the very last phase of their development, utilizing only a fragment of the available nominal performance when nominating to the TOP500 [3], withdrawn from the competition in a half year, different ranking on different benchmarks, tragically low efficiency of some application on some architecture, strategic alliance of manufacturers with conflicting interests for producing new processor/accelerator/supercomputer, or calling for project ideas for a supercomputer with unknown architecture and features. All these happenings seem to support the presumption that parallel processing has approached some mystical performance bound.

Amdahl [4] called first the attention to that *the parallel systems built from single processors have serious performance limitations*. His warning was used successfully on quite different fields [5], sometimes misunderstood, abused and even refuted (for a review see [6]). Despite of that "*Amdahl's Law is one of the few, fundamental laws of computing*" [7], for today it is quite forgotten because it is commonly assumed to be valid for software (SW) only, although Amdahl was speaking about complex computing systems.

The paper attempts to review some issues important for the coming exascale computing age and is structured as follows. Amdahl's law is reinterpreted for modern computing systems in section 2 in its original spirit, rather than as formulated by the successors of Amdahl and commonly accepted today. Section 3 explains why in consequence of Amdahl's Law the present computing technologies have an inherent (although depending on many factors) upper bound for the achievable performance gain. In the light of this, section 4 considers the documented history of supercomputers, and demonstrates that the efficiency of parallelization (in other words the achievable parallelization performance gain) governed the supercomputer development. Based on the idea of Amdahl that the different non-parallelizable contributions, *independently of their origin*, act as a single non-parallelizable fraction, section 5 introduces a by intention strongly simplified model, that assigns some separated fractions to some major contributors, enabling to prepare a semi-technical model. The behavior of the intro-

duced contributions is analyzed in section 6. Section 7 explains what attribute of supercomputers is measured by the different benchmarks and why selecting a proper benchmark is important for ranking the supercomputers. As section 8 presents, the model enables to understand what factors affect the efficiency of an application on supercomputers having different configurations. The behavior of applications depend heavily on the supercomputer hardware (HW). Based on the behavior of those different contributions, section 9 provides some short-term predictions on the development of supercomputer performance. As it can be concluded from the previous sections, and was guessed by many researchers, one of the major obstacles on the road towards even higher performance is the computing paradigm itself. This issue is discussed in section 10, where also a possible solution is suggested. The discussion is concluded in section 11 with repeating the prophecy of Amdahl: *the age of single-processor approach development is over, other than marginal advances will only be possible if using cooperating processors.*

2. Reinterpreting Amdahl's Law

2.1. Amdahl's idea

A general misconception (introduced by the successors of Amdahl) is to assume that Amdahl's law is valid for software only and that the non-parallelizable fraction contains something like the ratio of the numbers of the corresponding instructions. Actually, *Amdahl's law is valid for any partly parallelizable activity (including computer unrelated ones) and the non-parallelizable fragment shall be given as the ratio of the time spent with non-parallelizable activity to the total time.* Amdahl in his famous paper [4] speaks about "*the fraction of the computational load*" and explicitly mentions, in the same sentence and same rank, algorithmic reasons like "*computations required may be dependent on the states of the variables at each point*"; architectural aspects like "*may be strongly dependent on sweeping through the array along different axes on succeeding passes*" as well as "*physical problems*" like "*propagation rates of different physical effects may be quite different*". His point of view is valid also today: one has to consider the *workload of the complex HW/SW system, rather than some segregated component*, and his idea describes parallelization imperfectness of any kind. *When applied to a particular case, especially in the case of exascale systems, it shall be scrutinized which contributions can be neglected. Notice that the eligibility of neglecting some component changes with time, technology and conditions.*

2.2. Deriving the effective parallelization

Successors of Amdahl expressed Amdahl's law with the formula

$$S^{-1} = (1 - \alpha) + \alpha/k \quad (1)$$

where k is the number of parallelized code fragments, α is the ratio of the parallelizable part within the total code, S is the measurable speedup. The assumption can be visualized that (assuming many processors) in α fraction

of the running time the processors are executing parallelized code, in $(1-\alpha)$ fraction they are waiting (all but one), or making non-payload activity. That is α describes how much, in average, processors are utilized, or how effective (at the level of the computing system) the parallelization is.

For a system under test, where α is not *a priori* known, one can derive from the measurable speedup S an *effective parallelization factor* ([8]) as

$$\alpha_{eff} = \frac{k}{k-1} \frac{S-1}{S} \quad (2)$$

Obviously, this is not more than α expressed in terms of S and k from Equ. (1). For the classical case, $\alpha = \alpha_{eff}$; which simply means that in the *ideal* case the actually measurable effective parallelization achieves the theoretically possible one. In other words, α describes a system the *architecture* of which is completely known, while α_{eff} characterizes the *performance*, which describes both the complex architecture and the actual conditions. It was also demonstrated [8] that α_{eff} can be successfully utilized to describe parallelized behavior from SW load balancing through measuring efficacy of the on-chip HW communication to characterize performance of clouds.

The value α_{eff} can also be used to refer back to Amdahl's classical assumption even in the realistic case when the parallelized chunks have different lengths and the overhead to organize parallelization is not negligible. The speedup S can be measured and α_{eff} can be utilized to characterize the measurement setup and conditions, how much from the theoretically possible maximum parallelization is realized. Numerically $(1 - \alpha_{eff})$ equals with the f value, established theoretically [9].

The distinguished constituent in Amdahl's classic analysis is the parallelizable fraction α , all the rest (including wait time, non-payload activity, etc.) goes into the "sequential-only" fraction. When using several processors, one of them makes the sequential calculation, the others are waiting (use the same amount of time). So, when calculating the speedup, one calculates

$$S = \frac{(1 - \alpha) + \alpha}{(1 - \alpha) + \alpha/k} = \frac{k}{k(1 - \alpha) + \alpha} \quad (3)$$

hence the efficiency is

$$E = \frac{S}{k} = \frac{1}{k(1 - \alpha) + \alpha} \quad (4)$$

This explains the behavior of diagram $\frac{S}{k}$ in function of k experienced in practice: the more processors, the lower efficiency. At this point one can notice that $\frac{1}{E}$ is a linear function of number of processors, and its slope equals to $(1 - \alpha)$. Equ. (4) also underlines the importance of the single-processor performance: the lower is the number of the processors used in the parallel system having the expected performance, the higher can be the efficacy of the system.

Notice also, that through using Equ. (4), the efficiency $\frac{S}{k}$ can be equally good for describing the efficiency of parallelization of a setup, provided that the

number of processors is also known. From Equ. (4)

$$\alpha_{E,k} = \frac{Ek - 1}{E(k - 1)} \quad (5)$$

If the parallelization is well-organized (load balanced, small overhead, right number of processors), α_{eff} is close to unity, so tendencies can be better displayed through using $(1 - \alpha_{eff})$ in the diagrams below.

The importance of this practical term α_{eff} is underlined by that the achievable speedup (performance gain) can easily be derived from Equ. (1) as

$$G = \frac{1}{(1 - \alpha_{eff})} \quad (6)$$

2.3. The original assumptions

The classic interpretation implies three¹ essential restrictions, but those restrictions are rarely mentioned in the textbooks on parallelization:

- the parallelized parts are of equal length in terms of execution time
- the housekeeping (controlling parallelization, passing parameters, waiting for termination, exchanging messages, etc.) has no cost in terms of execution time
- the number of parallelizable chunks coincides with the number of available computing resources

Essentially, this is why *Amdahl's law represents a theoretical upper limit for parallelization gain*. It is important to notice, however, that *a 'universal' speedup exists only if the parallelization efficiency α is independent from the number of the processors*. As will be discussed in section 6, this assumption is only valid if the number of processors is low, so the usual linear extrapolation of the actual performance on the nominal performance will not be valid any more in the case of the exascale computing systems.

2.4. The additional factors considered here

In the spirit of the *Single Processor Approach* (SPA) the programmer (the person or the compiler) has to organize the job: at some point the initiating processor splits the execution, transmits the necessary parameters to some other processing units, starts their processing, then waits for the termination of started processings; see section 3. Real-life programs show sequential-parallel behavior, with variable degree of parallelization [10] and even apparently massively parallel algorithms change their behavior during processing [11]. All these make Amdahl's original model non-applicable, and call for extension.

As discussed in [1]

¹An additional essential point which was missed by both [9] and [2], that *the same computing model was used in all computers considered*.

- many parallel computations today are limited by several forms of communication and synchronization
- the parallel and sequential runtime components are only slightly affected by cache operations
- wires get increasingly slower relative to gates

In the followings

- the main focus will be on synchronization and communication; they are kept at their strict absolute minimum; and their effect is scrutinized
- the effect of cache will be neglected, and runtime components not discussed separately
- the role of the wires is considered in an extended sense: both the importance of physical distance and using special connection methods will be discussed

3. The inherent limit of parallelization

As it was mentioned in the previous section, initially and finally only one thread exists, i.e. the minimal absolutely necessary non-parallelizable activity is to fork the other threads and join them again. With the present technology, no such actions can be shorter than one clock period. That is, the non-parallelizable fraction will be given as the ratio of the time of the two clock periods to the total execution time. The latter time is a free parameter in describing the efficiency, i.e. the value of the effective parallelization α_{eff} also depends on the total benchmarking time (and so does the achievable parallelization gain, too).

This dependence is of course well known for supercomputer scientists: for measuring efficiency with better accuracy (and also for producing better α_{eff} values) hours of execution times are used in practice. For example in the case of benchmarking the supercomputer *Taihulight* [12] 13,298 seconds benchmark runtime was used; on the 1.45 GHz processors it means $2 * 10^{13}$ clock periods. This means that (at such benchmarking time) the inherent limit of $(1 - \alpha_{eff})$ is 10^{-13} (or equivalently the achievable performance gain is 10^{13}). In the followings for simplicity 1.00 GHz processors (i.e. 1 ns clock cycle time) will be assumed.

The supercomputers, however, are distributed systems. In a stadium-sized supercomputer the distance between processors (cable length) about 100 m can be assumed. The net signal round trip time is cca. 10^{-6} seconds, or 10^3 clock periods. The presently available network interfaces have 100...200 ns latency times, and sending a message between processors takes time in the same order of magnitude. This also means that *making better interconnection is not a bottleneck in enhancing performance*. This statement is underpinned also by statistical considerations [13].

Taking the (maybe optimistic) value $2 * 10^3$ clock periods for the signal propagation time, the value of the effective parallelization $(1 - \alpha_{eff})$ will be at least

in the range of 10^{-10} , only because of the physical size of the supercomputer. This also means that the expectations against the absolute performance of supercomputers are excessive: assuming a 10 Gflop/s processor, the achievable absolute *nominal* performance is $10^{10} \times 10^{10}$, i.e. 100 EFlops. Because of this, in the name of the company PEZY² the last two letters are surely obsolete. It looks like that in the feasibility studies an analysis for whether this inherent performance bound exists is done neither in USA [14] nor EU[15].

Another major issue arises from the computing principle Single Processor Approach (SPA): only one computer at a time can be addressed by the first one. As a consequence, minimum as many clock cycles are to be used for organizing the parallel work as many addressing steps required. Basically, this number equals to the number of cores in the supercomputer, i.e. the addressing in the TOP10 positions typically needs clock cycles in the order of $5 \times 10^5 \dots 10^7$; degrading the value of $(1 - \alpha_{eff})$ into the range $10^{-6} \dots 2 \times 10^{-5}$. Two tricks may be used to reduce the number of the addressing steps: either the cores are organized into *clusters* as many supercomputer builders do, or the processor itself can take over the responsibility of addressing its cores [16]. Depending on the actual construction, the reducing factor can be in the range $10^2 \dots 5 \times 10^4$, i.e. the resulting value of α_{eff} is expected to be in the range of $10^{-8} \dots 2 \times 10^{-6}$. Notice that utilizing "cooperative computing" [16] enhances further the value of $(1 - \alpha_{eff})$, but it means already utilizing a (slightly) different computing principle.

An operating system must also be used, for protection and convenience. If one considers the context change with its consumed 2×10^4 cycles [17], the absolute limit is cca. 5×10^{-8} , on a zero-sized supercomputer. This value is also very close to the "danger zone" derived above. This is why *Taihulight* runs the actual computations in kernel mode [16].

It is crucial to understand that the decreasing efficiency (see Equ. (4)) is coming from the computing principle itself rather than from some kind of engineering imperfection. This inherent limitation is of principal nature and cannot be mitigated without changing the computing principle. For validating these limitations, see also the measured performance data in Fig. 5.

Although not explicitly dealt with here, notice that the data exchange between the first thread and the other ones also contribute to the non-parallelizable fraction and typically uses system calls, for details see [10, 18] and section 7.

4. The history of supercomputer development

As the supercomputer performance is traditionally characterized by the data R_{Max} and R_{Peak} , their efficiency can be derived and using Equ. (5) the value of α_{eff} can easily be calculated. Performing that calculation for the data in TOP500 [3] for the first 25 supercomputers in the first 25 years, the "parallelization hillside" for the top supercomputers can be derived, see Fig. 1. From

²https://en.wikipedia.org/wiki/PEZY_Computing: The name PEZY is an acronym derived from the greek derived Metric prefixes Peta, Eta, Zetta, Yotta

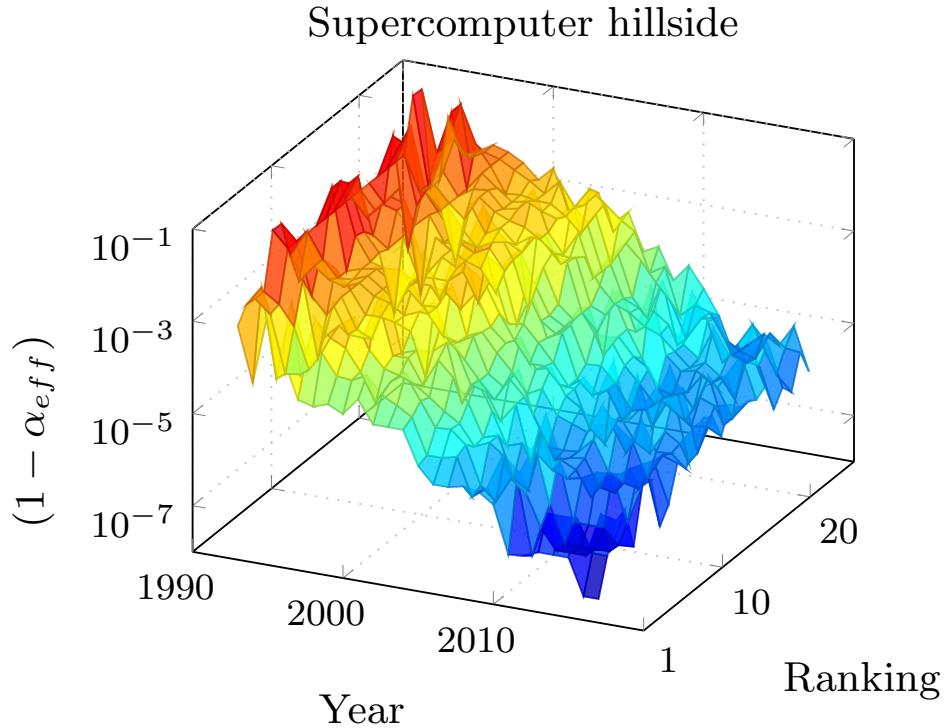


Figure 1: The Top500 supercomputer parallelization efficiency. The $(1 - \alpha)$ parameter for the past 25 years and the (by R_{max}) first 25 computers. Data derived using the HPL benchmark.

the figure one can conclude that the value of $(1 - \alpha_{eff})$ decreases with the number of years (as technology develops) and increases with the ranking (the engineering ingenuity).

Fig. 2 displays the same data from another point of view. On the vertical axis the performance gain (the single-processor performance is not included, see Equ.(6)) is shown. This diagram is for enhancing parallelization what Moore observation is for enhancing electronic density, i.e. *the development of the parallelization of supercomputers is governed by the Amdahl's Law*. Notice that the data are derived from the *efficiency*, so neither single processor performance nor clock frequency are included. It was theoretically concluded in section 3 that the performance gain data are expected to saturate around the value 10^7 , as displayed in the figure.

The data in the big circle show up a kind of stalling, not experienced in the former years. In other words, the saturation effect [2] is displayed. As was discussed in section 3, the processor of *Taihulight* deploys cooperative computing (i.e. a slightly different computing principle) that keeps its achievable performance gain at a value slightly higher than the predicted maximum. The newly shined up supercomputer *Summit*, deploying the conventional comput-

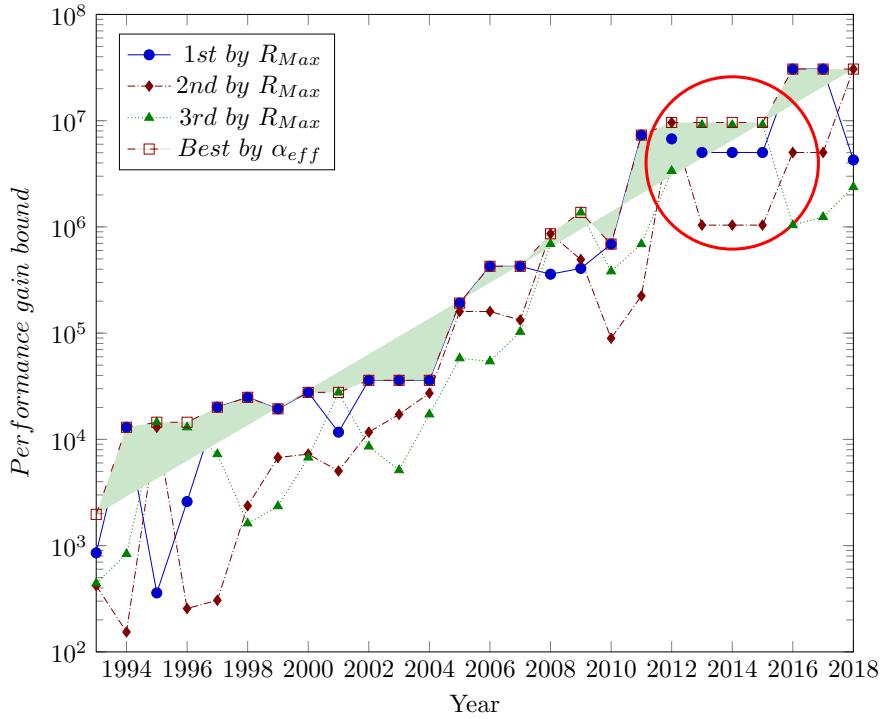


Figure 2: The trend of the development of computing performance gain in the past 25 years, based on the first three (by R_{Max}) and the first (by $(1 - \alpha)$) in the year in question. Data derived using the HPL benchmark.

ing principle, shows up the traditionally achievable parallelization gain; it could conquer slot #1 only thanks to its better single processor performance. The advantage of its processor for supercomputing was also underpinned by statistical considerations [13].

5. A simplified model for parallel operation

5.1. The performance losses

When speaking about computer performance, a modern computer system is assumed, which comprises many sophisticated components (in most cases embedding complete computers), and their complex interplay results in the final performance of the system. In the course of efforts to enhance processor performance through using some computing resources in parallel, many ideas have been suggested and implemented, both in SW and HW [19]. All these approaches have different usage scenarios, performance and limitations. Because of the complexity of the task and the limited access to the components, empirical methods and strictly controlled special measurement conditions are used to

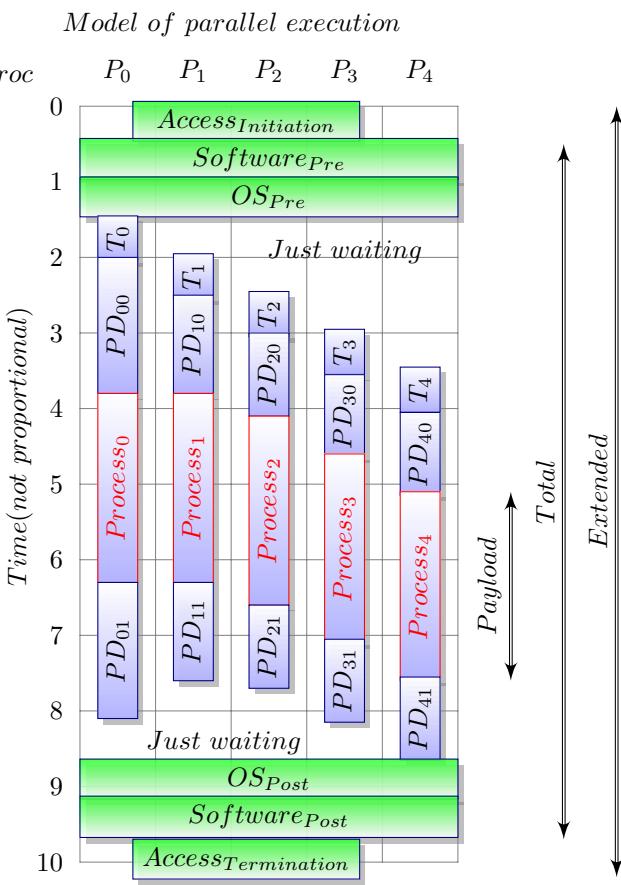


Figure 3: The extended Amdahl's model (strongly simplified)

quantitatively performance [20]. Whether a metric is appropriate for describing parallelism, depends on many factors [9, 21, 22].

As mentioned in section 2, Amdahl listed different reasons why losses in the "computational load" can occur. To understand operation of computing systems working in parallel, one needs to extend Amdahl's original (rather than that of the successors') model in such a way, that non-parallelizable (i.e. apparently sequential) part comprises contributions from HW, operating system (OS), SW and Propagation Delay (PD)³, and also some access time is needed for reaching the parallelized system. The technical implementations of different parallelization methods show up infinite variety, so here a (by intention) strongly simplified model is presented. Amdahl's idea enables to put everything that cannot be parallelized into the sequential-only fraction. The model is general enough to discuss qualitatively some examples of parallelly working systems, neglecting different contributions as possible in the different cases. The model can easily be converted to a technical (quantitative) one and the effect of inter-core communication can also easily be considered.

5.2. The principle of the measurements

When measuring performance, one faces serious difficulties, see for example [23], chapter 1, both with making measurements and interpreting them. When making a measurement (i.e. running a benchmark) either on a single processor or a system of parallelized processors, an instruction mix is executed many times. The large number of executions averages the rather different execution times [24], with an acceptable standard deviation. In the case when the executed instruction mix is the same, the conditions (like cache and/or memory size, the network bandwidth, Input/Output (I/O) operations, etc) are different and they form the subject of the comparison. In the case when comparing different algorithms (like results of different benchmarks), the instruction mix itself is also different.

Notice that the so called "algorithmic effects" – like dealing with sparse data structures (which affect cache behavior) or communication between the parallelly running threads, like returning results repeatedly to the main thread in an iteration (which greatly increases the non-parallelizable fraction in the main thread) – manifest through the HW/SW architecture, and they can hardly be separated. Also notice that there are fixed-size contributions, like utilizing time measurement facilities or calling system services. Since α_{eff} is a *relative* merit, the *absolute* measurement time shall be long. When utilizing efficiency data from measurements which were dedicated to some other goal, a proper caution must be exercised with the interpretation and accuracy of the data.

³This separation cannot be strict. Some features can be implemented in either SW or HW, or shared among them, and also some apparently sequential activities may happen partly parallel with each other.

5.3. The formal introduction of the model

The extended Amdahl's model is shown in Fig. 3. The contributions of the model component XXX to α_{eff} will be denoted by α_{eff}^{XXX} in the followings. Notice the different nature of those contributions. They have only one common feature: *they all consume time*. The vertical scale displays the actual activity for processing units shown on the horizontal scale.

Notice that our model assumes no interaction between the processes running on the parallelized systems in addition to the absolutely necessary minimum: starting and terminating the otherwise independent processes, which take parameters at the beginning and return results at the end. It can, however, be trivially extended to the more general case when processes must share some resource (like a database, which shall provide different records for the different processes), either implicitly or explicitly. Concurrent objects have inherent sequentiality [25], and synchronization and communication among those objects considerably increase [10] the non-parallelizable fraction (i.e. contribution $(1 - \alpha_{eff}^{SW})$), so *in the case of extremely large number of processors special attention must be devoted to their role on the efficiency of the application* on the parallelized system.

Let us notice that all contributions have a role during measurement: contributions due to SW, HW, OS and PD cannot be separated, though dedicated measurements can reveal their role, at least approximately. The relative weights of the different contributions are very different for the different parallelized systems, and even within those cases depend on many specific factors, so in every single parallelization case a careful analysis is required.

5.4. Access time

Initiating and terminating the parallel processing is usually made from within the same computer, except when one can only access the parallelized computer system from another computer (like in the case of clouds). This latter access time is independent from the parallelized system, and *one must properly correct for the access time when derives timing data for the parallelized system*. Amdahl's law is valid only for properly selected computing system. This is a one-time, and usually fixed size time contribution.

5.5. Execution time

The execution time *Total* covers all processings on the parallelized system. All applications, running on a parallelized system, must make some non-parallelizable activity at least before beginning and after terminating parallelizable activity. This SW activity represents what was assumed by Amdahl as the total sequential fraction⁴. As shown in Fig. 3, the *apparent* execution time includes the real payload activity, as well as waiting and OS and SW activity.

⁴Although some OS activity was surely included, Amdahl assumed some 20 % SW fraction, so the other contributions could be neglected compared to SW contribution.

Recall that the execution times may be different [24, 23, 26] in the individual cases, even if the same processor executes the same instruction, but executing an instruction mix many times results in practically identical execution times, at least at model level. Note that the standard deviation of the execution times appears as a contribution to the non-parallelizable fraction, and in this way increases the "imperfectionness" of the architecture. This feature of *processors* deserves serious consideration when utilizing a large number of processors. Over-optimizing a processor for single-thread regime hits back when using it in a parallelized many-processor environment, see also the statistical underpinning in [13].

6. The non-parallelizable contributions

6.1. The contribution of the operating system

All applications must use OS services and some HW facilities to initialize themselves as well as to access other processors. Because operating system works in a different (supervisor) mode, a considerable amount of time is required for the context switching⁵.

The OS initiates only the accessing of the processors, after that HW works partly in parallel with the next action of the OS and with the other actions initiating accessing other processors. This period is denoted in Fig. 3 by T_x . After the corresponding signals are generated, they must reach the target processor, that is they need some propagation time. PDs are denoted by PD_{x0} and PD_{x1} , corresponding to actions delivering input data and result, respectively. This propagation time (which of course occurs in parallel with actions on other processors, but which is a sequential contribution within the thread) depends strongly on how the processors are interconnected: this contribution can be considerable if the distance to travel is large or message transfer takes a long time (like lengthy messages, signal latency, handshaking, store and forward operations in networks, etc.).

6.2. The contribution of the physical size

Although the signals travel in a computing system with nearly the speed of the light, with increasing the physical size of the computer system a considerable time passes between issuing and receiving a signal, causing the other party to wait, without making any payload job. At the today's frequencies and chip sizes a signal cannot even travel in one clock period from one side of the chip to the other, in the case of a stadium-sized supercomputer this delay can be in the order of several hundreds clock cycles. Since the time of Amdahl, the ratio of the computing to the propagation time drastically changed, so –as [1] calls the attention to it– it cannot be neglected any more, although presently it is not (yet) a dominating term.

⁵This is usually not a crucial contribution, but under the extremal conditions represented by supercomputers, specialized operation systems must be used and every single core must run a lightweight OS [16]

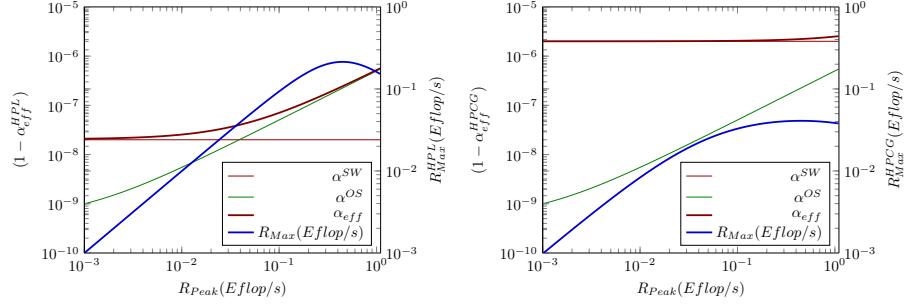


Figure 4: Contributions $(1 - \alpha_{eff}^X)$ to $(1 - \alpha_{eff}^{total})$ and max payload performance R_{Max} of a fictive supercomputer ($P = 1\text{Gflop/s} @ 1\text{GHz}$), imitating behavior of benchmarks HPL and HPCG. The $(1 - \alpha_{eff})$ values refer to the left scale, the R_{Max} values to the right scale

6.3. The contributions critical for large processor numbers

Notice that the non-payload processing activity comprises two contributions of variable length, so handling them properly is crucial for high number of processors. The first one is the OS contribution loop iteration overhead: processors must be handled one-by-one: they receive arguments and return results. This contribution simply increases linearly with the loop count. The second one is the propagation delay overhead PD that increases with the physical size of the supercomputer. The computer components can be in proximity in the range of mm , as well as in the range of 100 m . Similarly, they can be addressed in the first iteration or in the last one. These two contributions may be very different for different processing units, so combining the short ones from the first overhead class with long ones from the second overhead class reduces the overall overhead time, see Fig. 3.

From the figure one can find out the meaning of the introduced metric α_{eff} : it is simply the ratio of the *Payload* time (the processors are utilized to do actual work) to the *Total* execution time on the parallel system, exactly as in the classic Amdahl model. Notice that from the point of view of the processing units there is no difference *why* they cannot do *Payload* work: all (in)activities are considered as contributions to the non-parallelizable fraction. Notice also that using the *Extended* time in place of the *Total* time falsifies characteristics of the parameters of the parallelized system: it adds some foreign contribution (time consumed by systems *other* than the parallelized system) to the correct time.

6.4. How the dominance of contributions changes with the performance

Fig. 4 may help to understand the role of both the different kinds of contributions α_{eff}^X and the benchmarks. The figure displays diagrams for a hypothetic

supercomputer⁶, with two parameter sets imitating the behavior of the benchmarks HPL and High Performance Conjugate Gradients (HPCG). For clarity, only the dominant contributions from OS and SW are shown.

The supercomputer is the same in both cases, the only essential difference is in the value of α_{eff}^{SW} (i.e. another benchmark runs on it, and utilizes the HW facilities in a different way, see also section 7). As seen, the higher SW contribution from the benchmark causes drastic changes in the dependence of the payload performance on the nominal performance. At low values of processor numbers (low payload performance) the deviation is hardly noticeable. The higher SW contribution decreases the achievable maximum R_{Max} , and as the looping delay contribution due to the increasing number of processors exceeds the SW contribution, the R_{Max} diagrams decline. In other words, as the nominal performance approaches the expected dream limit, the resulting α_{eff} starts to raise and turns back the diagram of the actual computing performance R_{Max} . This is a new phenomenon, noticeable only at high number of processors. At the time of Amdahl, α was considered as constant (and was dominated by the SW contribution), but in the exascale systems it tends to dominate.

The performance gain of supercomputers, see Fig. 2, shows up a behavior very similar to that of the Moore’s law. It looks like that it increases year-by-year by a factor of cca. 1.5 (the performance *gain*, rather than the payload performance). Some reasons were presented, however, why also this behavior is not without limitations. Let us recall, that inside the circle for three years there was no change in the parameters. Even, with the appearance of the new world recorder in 2016 *Taihulight* –which utilizes cooperative computing, a different principle– the stall period seems to be prolonged for two more years. Using some reasonable assumptions about the different contributions to the non-parallelizable fraction mentioned above, their order of magnitude were estimated, and also some saturation values were predicted in section 3.

Noticable that the performance gain in Fig. 4 even shows up a breakdown, while Fig. 2 does not. One can guess, however, that the stalling was caused by this breakdown: adding more processors *decreases* the actual performance, so no measurements data were published about the measurements with more processor and less performance; see also section 9.

7. Benchmarking supercomputer performance

As experienced in running the benchmarks HPL and HPCG and explained in connection with Fig. 4, the different benchmarks produce different payload performance and computational efficiency on the same supercomputer. The model presented in Fig. 3 enables to explain the difference.

The benchmarks, utilized to derive numerical parameters for supercomputers, are specialized *programs*, which run in the HW/OS *environment* provided

⁶The technical model is not accurate enough because the needed parameters are not provided, so only a guess for their order of magnitude can be given.

by the supercomputer under test. One can use benchmarks for different goals. Two typical fields of utilization: to describe the environment supercomputer application runs in, and to guess how quickly an application will run on a given supercomputer.

The (apparently) sequential fraction ($1 - \alpha_{eff}$), as it is obvious from our model, cannot distinguish between the (at least apparently) sequential processing time contributions of different origin, even the SW (including OS) and HW sequential contributions cannot be separated. Similarly, it cannot be taken for sure that those contributions sum up linearly. Different benchmarks provide different SW contributions to the non-parallelizable fraction of the execution time (resulting in different efficiencies and ranking [27]), so comparing results (and especially establishing ranking!) derived using different benchmarks shall be done with maximum care. *Since the efficiency depends heavily on the number of cores, different configurations shall be compared using the same benchmark and the same number of processors (or same R_{Peak}).*

If the goal is to characterize the supercomputer's HW+OS system itself, a benchmark program should distort HW+OS contribution as little as possible, i.e. the SW contribution must be much lower than the HW+OS contribution. In the case of supercomputers, the benchmark HPL is used for this goal since the beginning of the supercomputer age. The mathematical behavior of HPL enables to minimize SW contribution, i.e. HPL *delivers the possible best estimation for α_{eff}^{HW+OS} .*

If the goal is to estimate the expectable behavior of an application, the benchmark program should imitate the structure and behavior of the application. In the case of supercomputers, a couple of years ago the benchmark HPCG has been introduced for this goal, since "*HPCG is designed to exercise computational and data access patterns that more closely match a different and broad set of important applications, and to give incentive to computer system designers to invest in capabilities that will have impact on the collective performance of these applications*" [28]. However, its utilization can be misleading: *the ranking is only valid for the HPCG application, and only utilizing that number of processors.* HPCG seems really to give better hints for designing supercomputer applications⁷, than HPL does. According to our model, in the case of using the HPCG benchmark, the SW contribution dominates⁸, i.e. HPCG delivers the best estimation for α_{eff}^{SW} for this class of supercomputer applications.

Supercomputer community has extensively tested the efficiency of TOP500 supercomputers when benchmarked with HPL and HPCG [30]. It was found that the efficiency (and R_{Max}) is typically 2 orders of magnitude lower when benchmarked with HPCG rather than HPL, even at relatively low number of processors.

⁷This is why for example [29] considers HPCG as "practical performance".

⁸ Returning calculated gradients requires much more sequential communication (unintended blocking).

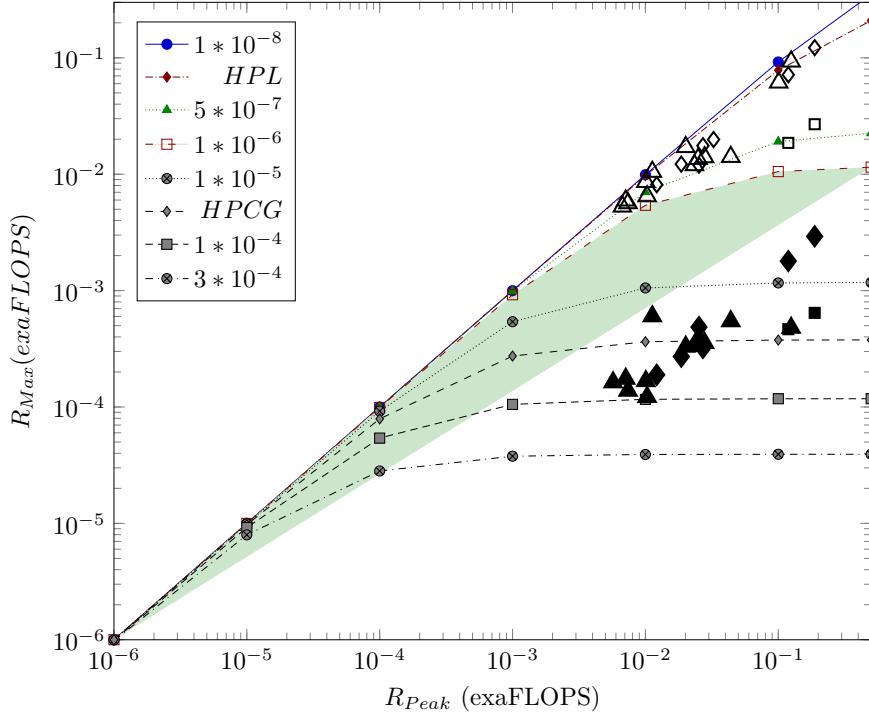


Figure 5: The R_{Max} payload performance in function of the peak performance R_{Peak} , at different $(1 - \alpha_{eff})$ values. The figures display the measured values derived using HPL and HPCG benchmarks, for the TOP15 supercomputers.

8. The efficiency of the applications

As discussed above, the value of $(1 - \alpha_{eff})$ differs for the two famous benchmarks by more than two orders of magnitude. For the users of exascale applications, it is primarily interesting how their application will run on an exascale supercomputer. Fig. 5 attempts to orient them in this question. The shading shows how nonlinearly the actual performance behaves when the nominal performance approaches the dream limit 1 $Eflop/s$.

The figure displays how the payload performance depends on the nominal performance, using $(1 - \alpha_{eff})$ as parameter. The diagram assumes that α_{eff} does not depend on the nominal performance, and the nominal performance is set by changing virtually the number of the cores. For orientation, the best benchmark results (as of 2018 July) are shown for the supercomputers in the TOP10 (either by HPL or by HPCG). The empty marks refer to the HPL case, the filled ones to HPCG. The diamonds denote GPU accelerated supercomputers, the triangles unaccelerated ones.

As predicted in section 3, the HPL payload performance data all fit in the $(1 - \alpha_{eff})$ band $10^{-7} \dots 10^{-6}$ (near to the $5 \cdot 10^{-7}$ value) and the HPCG payload

peformance data all fit in the α_{eff} band $10^{-5} \dots 10^{-4}$ (near to the $5*10^{-5}$ value). The corresponding measured values in Fig. 5 should be around the corresponding diagram lines marked by HPL and HPCG, repectively. In the light of the model above, these experiences should be easy to comprehend. The HPL and HPCG benchmarks measure the α_{eff}^{HW} and α_{eff}^{SW} , respectively.

The HPCG case is easier to explain. The HPCG measures the (dominant) contribution of the same software, so the measured performance value points are expected to scatter around the same value, provided that the single processor performance is not very different. The full marks clearly show the saturation effect, and the points scatter around the corresponding diagram line. The exceptions are being the new champion and its small brother. The reason is the exceptional single-processor performance of their processors. If one corrects for the performance factor (with relative to the single-processor performance of the *Taihulight*), the agreement is perfect, see the smaller filled quares at the corresponding nominal performance values.

In the HPL case the marks seem to follow the predicted HPL diagram line. The top 4 supercomputers deploy the same (or quite similar) trick: they reduce the looping delay by organizing the cores into groups. Supercomputers *Summit*, *Sierra* and *Tianhe-2* organize single processors into clusters, *Taihulight* organizes "clusters" inside the processor. As HPL measures value of α_{eff}^{HW+OS} , this benchmark is sensitive to decreasing the looping delay, the dominant contribution. This is why these supercomputers are on the top of the list. In addition, the higher single-processor performance also raises their value of R_{Max} .

The trick they use helps only, however, when the value of α_{eff}^{HW+OS} represents a relatively large contribution to the value of α_{eff} . In the HPL case decreasing α_{eff}^{OS} by two-three orders of magnitude reduces considerably the resulting effective parallelization, so maybe up to an order of magnitude better value of α_{eff} can be measured. In the case of HPCG, however, the contribution of $(1 - \alpha_{eff}^{SW})$ is about two orders of magnitude larger than the contribution $(1 - \alpha_{eff}^{HW+OS})$, so decreasing this latter by orders of magnitude has only marginal effect on $(1 - \alpha_{eff})$. The computers deploying this "clusterization" trick show up good results when benchmarked with HPL, but not so good when measured with HPCG. The exceptionally good values are made by that the performance is achieved through using much less processors (due to utilizing accelerator). See also Fig. 4.

So it looks like that the payload performance of supercomputers is limited as predicted in section 3. Even for the HPL-class applications, only a few tenths of *Eflop/s* can be achieved, for the more real-life (HPCG-class) applications achieving a few *Pflop/s* can be a realistic target.

The major difference between those two classes is the α_{eff}^{SW} contribution, mainly that the iterative nature of HPCG requires intensive data exchange with the ancestor thread, repeating thread forking and joining many times, as well as repeating calculation and communication of the new values for the next iteration many times. *All those actions increase the non-parallelizable fraction, i.e. decrease α_{eff} and R_{Max} .* Programmers of exascale applications

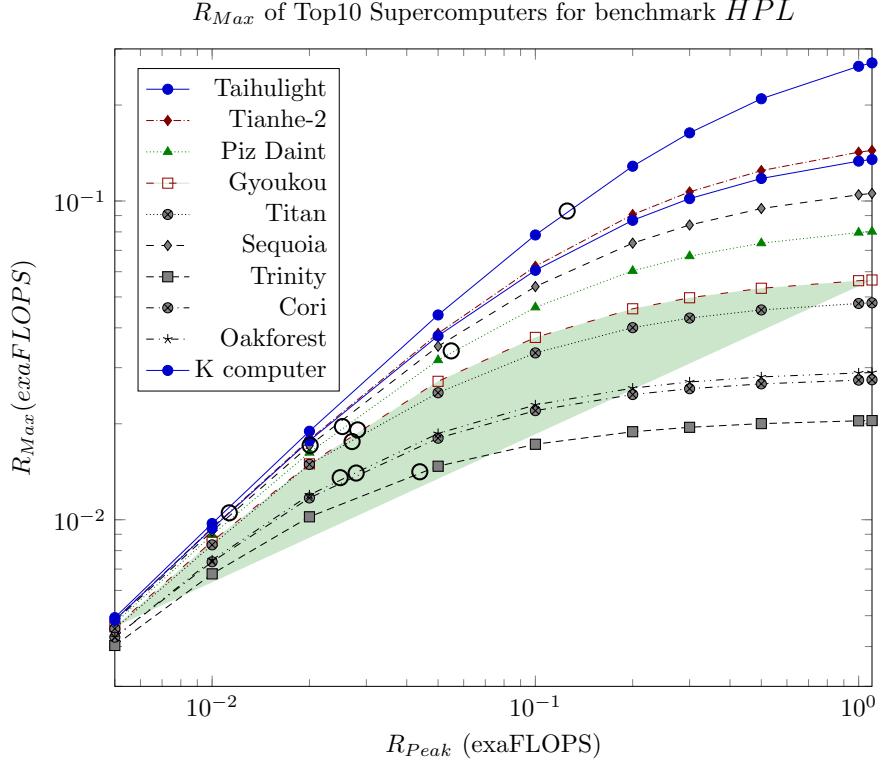


Figure 6: R_{Max} performance of selected TOP10 (as of 2017 November) supercomputers in function of their peak performance R_{Peak} , for the HPL benchmark. The actual R_{Max} values are denoted by a bubble.

shall reorganize their programs to mitigate this effect as much as possible: a kind of "SW clusterization" should be invented. Making local data handling and calculations in other than the ancestor thread, as well as avoiding not strictly necessary communication should be advantageous.

9. The perspectives of supercomputing

On the long term the "gold rush" will of course continue and (partly due to the unusual strategic alliances) new processors, connections, materials, principles will appear and maybe change the game. Some short time prediction, however, can be drawn from the tendencies and the presently available data base, again changing virtually the number of processors in the TOP10 supercomputers, hoping that α_{eff} will not increase. From this point of view, the predicted performance values are optimistic: when no drastic changes happen, the shown future performance values will be surely not exceeded. Fig. 6 shows how the performance of the TOP10 supercomputers (as of November 2017)

would change with the conditions given above. One can conclude that even the "benchmark payload performance" will not achieve the dream limit in the coming few years. See also Fig 5.

One has to consider Fig. 4 seriously: simply increasing the number of the cores or utilizing accelerators is a short dead-end street. It is worth to re-read Amdahl's classic paper [4]: "*the effort expended on achieving high parallel processing rates is wasted unless it is accompanied by achievements in sequential processing rates of very nearly the same magnitude*", see also Equ. (4). No single processor was added to the Chinese top supercomputers for years; *Gyoukou* failed (and withdrawn) because it could only use 12% of its processors; *Aurora* has been cancelled (and redesigned) because of similar reasons; *Summit* utilizes only 2.3M cores out of the available 2.8M. The newest (and probably worst) example is *AI Bridging Cloud Infrastructure* at slot #5. They have outstanding single-processor performance but –because of using the cloud infrastructure– a very poor parallelization ($\alpha_{eff} = 5.9 * 10^{-5}$), so they can only use 2.8% of their cores, although they have "only" 392K cores total.

10. The role of the computing paradigm

As it could be concluded from the discussion above, one of the major issues in increasing the payload performance of supercomputers is that the looping delay should be decreased. This obstacle originates in the SPA. Since in the paradigm only one processor exists, all components are designed in the spirit of SPA and all supercomputers are built from commodity components. As a consequence, on the bus only one processor can be addressed at a time and the cores cannot directly exchange data between each other. The OS takes over the responsibility of knowing about other processors, but it does it in a rather time-expensive way [17]. The data exchange can take place only through some kind of "far" memory, causing slowdown and sharing problems.

Reducing the looping delay can be relatively easily solved by "clustering": one can organize the cores into "nodes" and then the main thread shall iterate only through the nodes rather than the cores. In practice, it can be solved through organizing the single (many-core) processors into clusters, or delegating the cluster organization to the many-core processor [16]. Both solutions enabled a supercomputer to conquer the #1 slot for a while.

There are attempts to make direct data transfer through registers of different cores, like [31, 32], but the idea of the explicit cooperation (in form of data transfer) in processor chips has been implemented for the first time in [16] (just notice that some 50 years later after that Amdahl suggested the idea [4]). The power of the idea is clearly shown by that the supercomputer *Taihulight* stayed on the top for two years, when ranked by R_{Max}^{HPL} , and continues to stay when ranked by α_{eff}^{HPL} or G^{HPL} , thanks only to its processor.

Among others, the supercomputer experiences also underpin the need for renewing computing [33]. Working out a computing paradigm which is drastically different from the traditional one and at the same time it is upward compatible

with it, is not simple but possible [34]. Its simulator [35] proves the feasibility and viability of the concept. Utilizing that (or some similar) concept the limitations stemming out from the computing paradigm can be circumvented, similarly to some other limitations of computing [1]. In that way in the farther future the "dream limit" can be exceeded and exascale applications with reasonable efficiency can be prepared.

11. Summary

The paper discussed the exascale applications and exascale supercomputers as a complex HW/SW system. It has pointed out that such systems have inherent performance limitations, and through understanding the reasons of those limitations, they can be mitigated. The introduced naive model (based on extending Amdahl's principle) describes surprisingly well the performance values measured on the recently announced supercomputers and explains some mysteries experienced around supercomputers approaching the exaFLOPS dream limit. The model also explains the role of benchmarking and provides some practical hints for writing efficient applications for the future exascale computers. Based on the rigorously verified database of supercomputer performance data, it is shown that with the present technologies (and computing principle) the dream limit cannot be achieved. The need for "rebooting computing" is underlined by the analysis and also a possible way out of the present stalling through extending the computing paradigm is shown.

Acknowledgements

Project no. 125547 has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the K funding scheme.

References

References

- [1] I. Markov, Limits on fundamental limits to computation, *Nature* 512(7513) (2014) 147–154.
- [2] P. J. Denning, T. Lewis, Exponential Laws of Computing Growth, *Communications of the ACM* (2017) 54–65doi:DOI:10.1145/2976758.
- [3] TOP500, November 2017 list of supercomputers, <https://www.top500.org/lists/2017/11/> (2017).
- [4] G. M. Amdahl, Validity of the Single Processor Approach to Achieving Large-Scale Computing Capabilities, in: AFIPS Conference Proceedings, Vol. 30, 1967, pp. 483–485. doi:10.1145/1465482.1465560.

- [5] S. Krishnaprasad, Uses and Abuses of Amdahl's Law, *J. Comput. Sci. Coll.* 17 (2) (2001) 288–293.
URL <http://dl.acm.org/citation.cfm?id=775339.775386>
- [6] F. Dévai, The Refutation of Amdahl's Law and Its Variants, in: O. Gervasi, B. Murgante, S. Misra, G. Borruso, C. M. Torre, A. M. A. Rocha, D. Taniar, B. O. Apduhan, E. Stankova, A. Cuzzocrea (Eds.), Computational Science and Its Applications – ICCSA 2017, Springer International Publishing, Cham, 2017, pp. 480–493.
- [7] J. M. Paul, B. H. Meyer, Amdahl's Law Revisited for Single Chip Systems, *International Journal of Parallel Programming* 35 (2) (2007) 101–123.
doi:10.1007/s10766-006-0028-8.
URL <https://doi.org/10.1007/s10766-006-0028-8>
- [8] J. Végh, P. Molnár, How to measure perfectness of parallelization in hardware/software systems, in: 18th Internat. Carpathian Control Conf. ICCC, 2017, pp. 394–399.
- [9] A. H. Karp, H. P. Flatt, Measuring Parallel Processor Performance, *Commun. ACM* 33 (5) (1990) 539–543. doi:10.1145/78607.78614.
URL <http://doi.acm.org/10.1145/78607.78614>
- [10] L. Yavits, A. Morad, R. Ginosar, The effect of communication and synchronization on Amdahl's law in multicore systems, *Parallel Computing* 40 (1) (2014) 1–16.
- [11] K. Pingali, D. Nguyen, M. Kulkarni, M. Burtscher, M. A. Hassaan, R. Kaleem, T.-H. Lee, A. Lenhardt, R. Manevich, M. Méndez-Lojo, D. Prountzos, X. Sui, The Tao of Parallelism in Algorithms, *SIGPLAN Not.* 46 (6) (2011) 12–25. doi:10.1145/1993316.1993501.
URL <http://doi.acm.org/10.1145/1993316.1993501>
- [12] J. Dongarra, Report on the Sunway TaihuLight System, Tech. Rep. Tech Report UT-EECS-16-742, University of Tennessee Department of Electrical Engineering and Computer Science (June 2016).
- [13] J. Végh, Statistical considerations on limitations of supercomputers, CoRR abs/1710.08951. arXiv:1710.08951.
URL <http://arxiv.org/abs/1710.08951>
- [14] US DOE, The Opportunities and Challenges of Exascale Computing, https://science.energy.gov/~media/ascr/ascac/pdf/reports/Exascale_subcommittee_report.pdf (2010).
- [15] European Commission, Implementation of the Action Plan for the European High-Performance Computing strategy, http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=15269 (2016).

- [16] F. Zheng, H.-L. Li, H. Lv, F. Guo, X.-H. Xu, X.-H. Xie, Cooperative computing techniques for a deeply fused and heterogeneous many-core processor architecture, *Journal of Computer Science and Technology* 30 (1) (2015) 145–162.
`doi:10.1007/s11390-015-1510-9`
URL <https://doi.org/10.1007/s11390-015-1510-9>
- [17] D. Tsafrir, The context-switch overhead inflicted by hardware interrupts (and the enigma of do-nothing loops), in: *Proceedings of the 2007 Workshop on Experimental Computer Science, ExpCS '07*, ACM, New York, NY, USA, 2007, pp. 3–3.
`doi:10.1145/1281700.1281704`
URL <http://doi.acm.org/10.1145/1281700.1281704>
- [18] S. Eyerman, L. Eeckhout, Modeling Critical Sections in Amdahl's Law and Its Implications for Multicore SIGARCH Comput. Archit. News 38 (3) (2010) 362–370.
`doi:10.1145/1816038.1816011`
URL <http://doi.acm.org/10.1145/1816038.1816011>
- [19] K. Hwang, N. Jotwani, *Advanced Computer Architecture: Parallelism, Scalability, Programmability*, 3rd Edition, Mc Graw Hill, 2016.
- [20] D. J. Lilja, *Measuring Computer Performance: A practitioner's guide*, Cambridge University Press, 2004.
- [21] X.-H. Sun, J. L. Gustafson, Paper: Toward a better parallel performance metric, *Parallel Comput.* 17 (10-11) (1991) 1093–1109.
`doi:10.1016/S0167-8191(05)80028-6`
URL [http://dx.doi.org/10.1016/S0167-8191\(05\)80028-6](http://dx.doi.org/10.1016/S0167-8191(05)80028-6)
- [22] S. Orii, Metrics for evaluation of parallel efficiency toward highly parallel processing, *Parallel Computing* 36 (1) (2010) 16 – 25.
`doi:http://dx.doi.org/10.1016/j.parco.2009.11.003`
URL <http://www.sciencedirect.com/science/article/pii/S0167819109001227>
- [23] D. Patterson, J. Hennessy (Eds.), *Computer Organization and design*. RISC-V Edition, Morgan Kaufmann, 2017.
- [24] P. Molnár, J. Végh, Measuring Performance of Processor Instructions and Operating System Services in Soft Processor Based Systems, in: 18th Internat. Carpathian Control Conf. ICCC, 2017, pp. 381–387.
- [25] F. Ellen, D. Hendler, N. Shavit, On the Inherent Sequentiality of Concurrent Objects, *SIAM J. Comput.* 43 (3) (2012) 519536.
`doi:10.1137/08072646X`
- [26] R. E. Bryant, D. R. O'Hallaron, *Computer Systems: A Programmer's Perspective*, Pearson, 2014.
- [27] IEEE Spectrum, Two Different Top500 Supercomputing Benchmarks Show Two Different Top Supercomputers, <https://spectrum.ieee.org/tech-talk/computing/hardware/two-different-top500-supercomputers> (2017).

- [28] HPCG Benchmark, Hpcg benchmark, <http://www.hpcg-benchmark.org/> (2016).
- [29] T. Dettmers, The Brain vs Deep Learning Part I: Computational Complexity Or Why the Singularity Is Nowhere Near, <http://timdettmers.com/2015/07/27/brain-vs-deep-learning-singularity/> (2015).
- [30] J. Dongarra, The Global Race for Exascale High Performance Computing, http://ec.europa.eu/newsroom/document.cfm?doc_id=45647 (2017).
- [31] J. Congy, et al, Accelerating Sequential Applications on CMPs Using Core Spilling, Parallel and Distributed Systems 18 (2007) 1094–1107.
- [32] ARM, big.LITTLE technology (2011).
URL <https://developer.arm.com/technologies/big-little>
- [33] J. Végh, Renewing computing paradigms for more efficient parallelization of single-threads, Vol. 29 of Advances in Parallel Computing, IOS Press, 2018, Ch. 13, pp. 305–330.
- [34] J. Végh, Introducing the explicitly many-processor approach, Parallel Computing 75 (2018) 28 – 40.
doi:<https://doi.org/10.1016/j.parco.2018.03.001>.
- [35] J. Végh, EMPAthY86: A cycle accurate simulator for Explicitly Many-Processor Approach (EMPA) com
URL <https://github.com/jvegh/EMPAthY86>