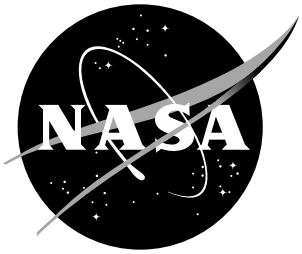
NASA/CR-1998-207652



A Petaflops Era Computing Analysis

Frank S. Preston
Computer Sciences Corporation, Hampton, Virginia

March 1998

The NASA STI Program Office ... in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA Scientific and Technical Information (STI) Program Office plays a key part in helping NASA maintain this important role.

The NASA STI Program Office is operated by Langley Research Center, the lead center for NASA's scientific and technical information. The NASA STI Program Office provides access to the NASA STI Database, the largest collection of aeronautical and space science STI in the world. The Program Office is also NASA's institutional mechanism for disseminating the results of its research and development activities. These results are published by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counter-part of peer reviewed formal professional papers, but having less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services that help round out the STI Program Office's diverse offerings include creating custom thesauri, building customized databases, organizing and publishing research results ... even providing videos.

For more information about the NASA STI Program Office, you can:

- Access the NASA STI Program Home Page at http://www.sti.nasa.gov/STI-homepage.html

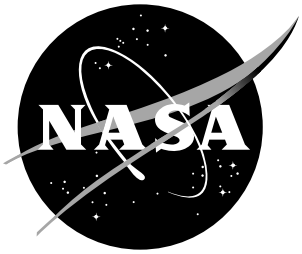
- E-mail your question via the Internet to help@sti.nasa.gov

- Fax your question to the NASA Access Help Desk at (301) 621-0134

- Phone the NASA Access Help Desk at (301) 621-0390

- Write to:
  NASA Access Help Desk
  NASA Center for AeroSpace Information
  800 Elkridge Landing Road
  Linthicum Heights, MD 21090-2934

NASA/CR-1998-207652



A Petaflops Era Computing Analysis

Frank S. Preston
Computer Sciences Corporation, Hampton, Virginia

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23681-2199

Prepared for Langley Research Center
under Contract NAS1-20048

March 1998

Available from the following:

URL address: ***http://techreports.larc.nasa.gov/ltrs/ltrs.html***

NASA Center for AeroSpace Information (CASI)
7121 Standard Drive
Hanover, MD 21076-1320
(301) 621-0390

National Technical Information Service (NTIS)
5285 Port Royal Road
Springfield, VA 22161-2171
(703) 487-4650

CONTENTS

(Contents continued on next page)

CONTENTS (Continued)

ABSTRACT

This report covers a study of the potential for petaflops (10^{15} floating point operations per second) computing. This study was performed within the past year and should be considered as the first step in an on-going effort. The analysis concludes that a petaflop system is technically feasible but not feasible with today's state-of-the-art. Since the computer arena is now a commodity business, most experts expect that a petaflops system will evolve from current technology in an evolutionary fashion. To meet the price expectations of users waiting for petaflop performance, great improvements in lowering component costs will be required. Lower power consumption is also a must. The present rate of progress in improved performance places the date of introduction of petaflop systems at about 2010. Several years before that date, it is projected that the resolution limit of chips will reach the now known resolution limit. Aside from the economic problems and constraints, software is identified as the major problem. The tone of this initial study is more pessimistic than most of the other material available on petaflop systems. Workers in the field are expected to generate more data which could serve to provide a basis for a more informed projection. This report includes an annotated bibliography.

SUMMARY

A number of important areas and tasks require performance well beyond the teraflops range (10^{12} floating point operations per second) — just now becoming available. This demand has focused some attention on the next major milestone of petaflops (10^{15} FLOPS) computing. Even this high performance is not sufficient for some problems but such capability, into the ExaFLOPS range (10^{18} FLOPS) is now beyond the limit of informed forecasts.

The present rate of progress, if continued, will reach a petaflop in about a dozen years. The present state-of-the-art in computers will be adequate to meet this goal from a performance standpoint, but not within reasonable power consumption. It is unlikely to yield sufficient cost improvement to place a petaflops system within the economical range of users and programs. The present computer environment has swung to a commodity economy. It is not clear that future trends, driven by the mass market, will produce computer and memory elements that are suited to petaflops computers. Despite the future uncertainties, most users and experts in the field expect that petaflop systems are certain.

To a manager of a program dependent upon supercomputing resources, and the managers of these facilities, the greatest problem and challenge is economic rather than technical. Existing systems are not well balanced and the trend has worsened this balance in a number of areas. Memory access speeds have not kept up with the growth in processing performance. Support subsystems are barely adequate for gigaflops systems and must be increased in major proportions to match petaflops — a million times faster. With today's shrinking budgets, supporting this with facilities and personnel is a real challenge. The biggest imbalance lies in the software area. New algorithms, new application programs, new operating systems, and new programming languages are all required.

This report concludes with a number of recommendations (page 26) applicable to NASA and Langley. Further study of the coming petaflop era is recommended.

INTRODUCTION

Purpose of Study

This study's objective was to examine the potential for petaflops (10^{15} floating point operations per second) computing. To place this in a current framework, the present operation of supercomputers falls in the range of from one to a few tens of gigaflops (10^9 to 10^{10}) for the upper limit of most supercomputing applications running in the year 1996. There are exceptions to this for demonstrations, speed contests, and special purpose systems. Some record levels have been demonstrated at greater than 100 GFLOPS (100 x 10^9) and computers are being ordered for performance of up to about 3 TFLOPS (3 x 10^{12}). Thus, a Petaflop computer will be 1,000 times as fast as the current fastest and generally 100,000 times as fast as most users are currently able to achieve. *Of course, petaflops systems are not for everyone.*

Before going on, it is critically important to qualify these numbers. Users are only able to enjoy the effective rate of computers. Manufacturer's, of course, can only quote peak rates or demonstrated rates on specific benchmarks, since the effective performance rate depends upon the application. The manufacturer is giving a figure of the speed "guaranteed not to be exceeded". Some years ago, these two numbers differed generally by about a factor of from two to perhaps eight. Now with old applications running on massively parallel systems, these numbers can differ by as much as a factor of from about 5 to 25 or more. Both users and suppliers have fallen into the bad practice of speaking of speed without specifying if it is PEAK or EFFECTIVE. Generally PEAK is implied — or the listener is assumed to be knowledgeable enough to know which. For this report, PEAK will be used unless otherwise stated. [For definitions of these and other terms, refer to the DEFINITIONS section of this report.]

This study started in March 1996 and the literature search portion was largely ended about September 1996. Although this report bears a 1997 date, it covers only work through the end of 1996. Computer technology, being a fast paced field, will change in the future whereas this report is "frozen" into the technology knowledge of 1996. It is even worse than that because most of the references are for earlier than 1996 and are reporting on work generally a year or more before than that.

One reason for doing this study is that Dan Goldin, NASA Administrator, has stated that looking just at teraflops computing is short-sighted, so NASA participates in an Interagency Petaflops Initiative Computing Group. (See Bibliography for Taub96 as a source) *Refer to References on Page 5 for reference coding.*

Scope of Study

As is generally now forecast, Petaflop systems will be technically feasible about the year 2007 — experts are using the range of 2007 to 2010. My projection in this report falls in the year 2014. {In 1994, my projection for a Teraflop system was 2003. It now appears that this 9 year span may be compressed into about 6. Some of the difference may be accounted for by differences between delivery of a machine and its working to its specification.} Thus this study, must include technologies that are suited to Petaflop systems and the time period to at least 2015.

This study is specifically limited to scientific computing and, with only minor exceptions, to general purpose (i.e. not special purpose) computers. Business computers have in the past lagged scientific by about five years, but this may not be the case in the 21st Century.

The original scope when this study was started, was to include ExaFLOPS (10^{18}) computing. Since references found for petaflops were rather limited — and nothing found on exaflops — this report will not treat exaflops. It is idle to speculate about 1,000 petaflops. Technology progress at the current rate would place exaflops about 11 years further in the future beyond petaflops. My view is that exaflops will require a far different technology and likely more than 11 additional years. It is entirely possible that these projections will turn out wrong as the result of some breakthrough, not now envisioned or at least reported.

Pre-conceived Problems, Questions, and Driving Forces

At the outset of this study, there were a number of questions and problems that immediately came to mind about petaflops computing. The study scope boundaries were set to address these points. Now, looking back, it must be admitted that not all of the problems and questions were satisfactorily resolved.

One question that is not specifically discussed herein was: "Are petaflops systems technically feasible or will they be in the next 15 to 20 years?" The reason for this is that opinions expressed verbally or in writing by computer experts were 100% of the opinion that "..... there is no question that petaflops systems are technically feasible and will be introduced". Frequently this reduced to "No doubt, what-so-ever!"

Another factor, that is nearly universally accepted now, is that the computer business has become a commodity business — with all that that implies. Thus, it will be driven by economic factors rather than technology as it has been until about five years ago. This means, in particular, that the semiconductor products developed and offered to the market, will be aimed at the broad market. This has already meant that new supercomputers use commodity processors and memories rather than the custom chips that previously were designed and produced for a specific computer. "Will these commodity chips meet the demands of future supercomputers and specifically into the petaflops era?" remains an unanswered question. This also applies, in some part, to other computer components.

Another way to look at the question raised in the previous paragraph is: "Will commodity chips and other computer components be used in petaflop computers?" This study fails to settle this question. The majority opinion seems to be that for the chips, economic considerations will constrain custom special supercomputer chips from being developed. This fits the present trend but will it in the year 2007? Also the majority of experts (and non-experts for that matter) expect that prices for the next several generations (a generation equals about 3 years) will continue to fall at the present rates and that this will not increase by much the cost of computer systems. This may mainly be wishful thinking. The minority opinion appears to favor some breakthrough. When, at what cost, and by what means is unknown at this time.

An optimistic view of effective rates, expressed by some, is based upon achieving efficiencies of from 25% to 50% — higher than now being demonstrated on scalable systems — on up to an order of magnitude more processors. To reach reasonable efficiency much software work will be necessary — algorithms, languages, and application programs.

Is There a Requirement?

This study does not directly address requirements for petaflops speed but there will not be petaflop systems unless there are requirements that furnish the funding. There is no doubt that real requirement exists. In many cases, important problems are being tackled and the analysis is being limited by the computational power available on the fastest computers now in service. Various compromises are necessary for these tasks to fit the memory and processing limits. Some future forecasts anticipate that commercial applications will become a significant factor in developing a petaflop market.

In the early 1990s, the scientific community initiated programs to address the "Grand Challenges" by making the step from gigaflops to teraflops. This initiative was endorsed and funded by Federal programs. Grand Challenge has been defined [Bibliography Wils89] as critical and important topics that are crippled by existing computers but that could be addressed by computers to be available within a decade. This led to the Teraflop Initiative and we are just now beginning to see computers that can run at these speeds.

There are numerous references available to document the requirements of the Grand Challenges. Some of these requirements extend into the petaflops and exaflops range — and even well beyond that. These have been reviewed in a prior report [Bibliography Pres94] by this author. The list of Grand Challenges changes with time but the "Official" list contains those most frequently cited and is reproduced in Table No. 1

The GRAND CHALLENGES	Ref. 5
CLIMATE MODELING	QUANTUM CHROMODYNAMICS
FLUID TURBULENCE	SEMICONDUCTOR MODELING
POLLUTION DISPERSION	SUPERCOMPUTER MODELING
HUMAN GENOME	COMBUSTION SYSTEMS
OCEAN CIRCULATION	VISION and COGNITION
ADDITIONAL "UNOFFICIAL" GRAND CHALLENGES *	* (Added later)
COMPUTATIONAL CHEMISTRY	PHYSICS (with various categories)
MEDICAL (with various categories)	

Table No. 1 The Grand Challenges for the Teraflop Era

References

Because of the way this study was done, references cited and additional information and sources indicated, are identified in two different ways.

For sources specifically cited within this report as well as on charts and graphs, references are identified by a citation to the Reference Number. These are shown either directly on the charts or in the spreadsheet where the data is reproduced. Where a reference number is given, these are listed at the end of this report in the Reference section [Example: Ref. 5)].

A separate Bibliography was prepared and augmented when each new source was studied. This Bibliography is included as Appendix A of this report. It has been annotated to assist anyone in reviewing the material as well as recalling, for this writer, pertinent items. The annotations are for the writer's benefit and not a scholarly review of the paper. Many of the papers reviewed were omitted from the Bibliography because they were deemed not pertinent. To supplement the Bibliography, each entry in the bibliography was given Keywords and listed under these as Categories to assist in searching for related material. (The Keyword Index and Category Index are included within Appendix B and C.) Papers are listed chronologically in the Bibliography. The Category Index is keyed to the Bibliography using a four letter code and the year to identify the source. Within this report, the sources listed from the Bibliography use the four letter code [Example: Name95] when referring to the Bibliography; and constitute additional information and not as a cited reference.

STUDY AND ANALYSIS

Review of Applicable Source Data

Literature Search

The first step of this study was to initiate a search for applicable technical papers. This search was active primarily for a period of about six months (March to September 1996). During this time, references were studied and relevant material analyzed. References found subsequent to December 1996 have been added to the Bibliography to keep it up-to-date.

During the period when the Library was looking for material for me, I worked in my office independently on the Internet, using several of the search engines. Here, these mainly made access to references in specific categories, such as computers. Within these categories, the search engines went to various collections and ran through the papers looking for the topic being searched. The references found could, in most cases, be examined in total which was valuable and not always possible with the library search results. Despite being overloaded with irrelevant responses, some suitable material was collected. It is probable that had I had more experience with the various search engines, I would have been more successful. As an example of this, it took some time to discover that searching for PETAFLOP would not find items on PETAFLOPS (or vice versa) — despite what one would expect. References on the net frequently work out to be poor because they are changed, moved or deleted. Many of the ones found a year ago are no longer to be found under the address previously used.

For a study such as this, it is necessary that it be on-going. There is always new material being published so searches must be repeated. For this reason, the search and reference study must have continuity to be efficient and effective.

Because of the importance of literature searches to this type of technical report, this report would not be complete without mentioning that although the searches were very helpful, the limited material found constitute a restriction on the coverage of this subject. It is worthwhile to examine briefly the problems with such searches.

Search Problems

Prior studies on related computer technology forecasts had acquainted me with the general topic and much of the published literature. Earlier, very little material was found looking forward beyond five years. Therefore, it was anticipated that the search for petaflop computer references and the technology for computers, components, and costs would be difficult and constrain this study. This would be less true for a search made a year later because of the increased publications on petaflops performance.

As was expected, specific references to petaflops were meager so the search was broadened to include high performance computing, advanced systems, and components, etc.. Software status and forecasts were searched but it proved nearly impossible to focus this on the petaflops era. Because of the importance of economic issues, cost projections were sought. This brought out the typical problem with this of getting all sorts of references that were not applicable.

Libraries depend upon external sources for search engines and therefore generally do not prepare the material that serves as the reference to papers. In my experience, this source material most often uses the title of the paper, the abstract, and keywords. Thus, these sources often are ill suited to searches on specific topics. The author of the papers generally do not prepare the Abstract to make it suited to reference searches. As an example, the abstract may contain the words “computer” and “costs” yet the only reference to costs in the paper is a paragraph on “cost effectiveness” without any economic or actual cost data. To screen out old data, searches were limited to 1990 and thereafter.

The biggest problem, of course, was the scarcity of good relevant papers on petaflop systems, components, software, and costs. This study is anticipated to be followed by further analysis and additional literature searches. It is expected that these additional studies will seek to obtain more and better references — and keep up with new papers.

One very good source, covering two workshops in 1996 and not received until 1997, Reference 9 was not available until after the preparation of this report.

Source Search Results

References contain citations to other papers and these pathways lead to trails that can be followed. In general, this study did not go very far down these leads. One needs to be an expert in many fields to understand and evaluate the reports. It is a difficult and a full time job these days to keep up with the literature in one specific area — and harder still in all the computer fields applicable to future systems. For those who wish to study petaflops further, the Bibliography can serve as a good starting point.

The tangible results of the literature searches can be assessed by examination of the study analysis that follows. Further study and analysis can be expected in work to follow this.

Petaflops System Requirements

System Sizing

Experience over the past 15 or so years has shown that supercomputers need the proper balance between speed, memory, local storage, internal communication rates, and the bandwidth of the communication interface to external systems. Looking at this for today's supercomputers, some of the system architecture has changed with the introduction of scalable systems, with a large number of processor chips which are almost self-contained computers. For these chips, the chip design defines the internal communication performance. Memory and storage is likely to be distributed and located at nodes. Communication between nodes (and processors) must be scalable and currently is by message passing — frequently via a switch. For this study, it appeared that the sizing analysis should mainly address storage and memory requirements.

The wisdom of supercomputer architects in the 1980's was to recommend a memory size of one WORD per FLOP. (8 Bytes or 64 bits per word) Only the most well endowed facilities have been able to afford this is the past, and many operated with less than this but they suffered. With the recent reduction in memory cost, the situation has improved. Memory requirements vary with the applications and different ratios are appropriate in specific cases. Current practice appears to size the memory in the order of one Byte per FLOP — a factor of eight less than the earlier goal. Extrapolating from one gigaflop to one petaflop (factor of one million) is a very large step and represents a huge impact on system costs. Experience with teraflops systems will provide a basis for informed estimates for petaflops systems.

For petaflop systems, several studies have been reporting that memory cost will be a large fraction of the total cost. Some architects continue to support the need for a linear relationship between memory and processing speed. Others, driven by the huge estimates for memory cost, have reported that they see the need for memory size to grow at a lower rate than processing speed – such as to the three quarter power. This will be a function of the application and the projects are using the supercomputer simultaneously.

Insufficient memory results in slower processing. Memory access rates are already failing to keep up to the growth in processing rates and this can be expected to worsen. To get around this, more use is made of cache, and more levels of cache and virtual memory are introduced. It seems best to size the very fastest computers using the ratio of one Byte per FLOP. Of course, a petaflop machine devoted to one discipline, such as molecular chemistry, could adopt a different ratio, determined by experience in that particular field. Accordingly, this study and report, adopts the ratio of one BYTE per FLOP for memory sizing. This memory (8 bits per FLOP) must be "close in access time" to the processor that uses it.

By the time petaflops systems are designed, alternate memory architectures may be utilized to curtail the latency. These include; Processor-In-Memory (PIM), or All-Cache designs, or other design strategies, as yet unknown. These alternative strategies would necessitate a different wisdom for sizing memory equivalents.

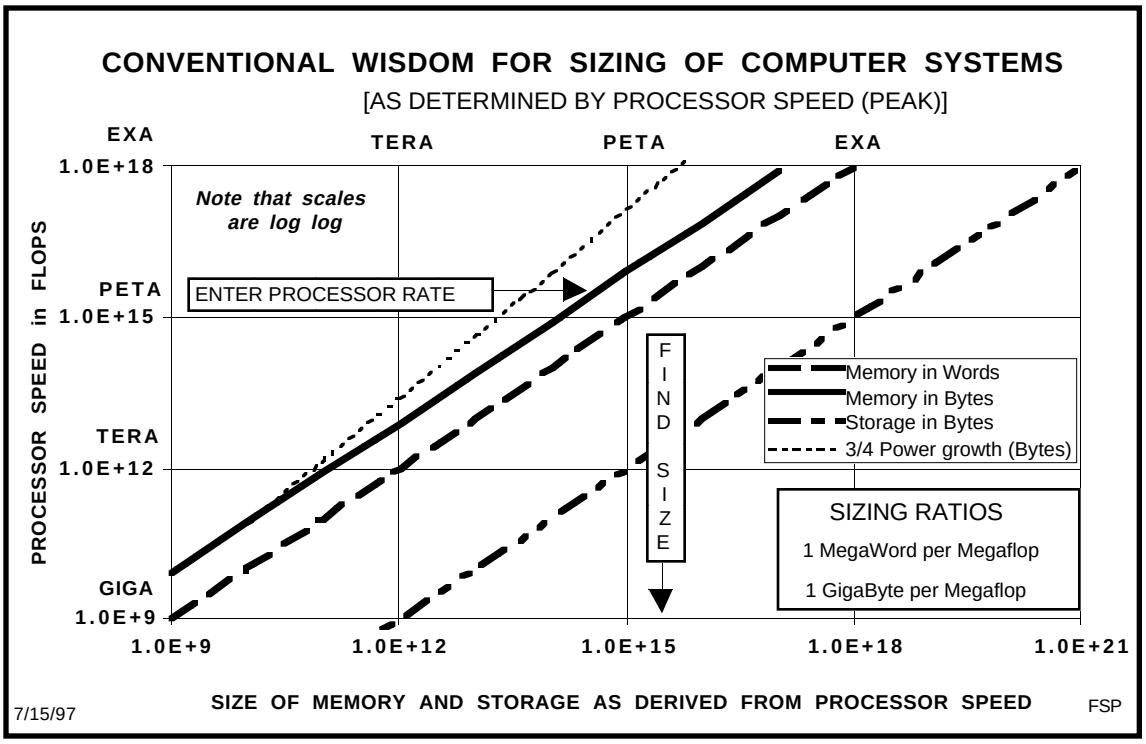


Figure No. 1 Memory and Storage Requirements

For storage, the conventional rule has been 1,000 BYTES per FLOP (or one GigaByte per MegaFLOPS). The storage is not as critical a problem as memory because it is not as expensive, and there is more freedom to distribute its location. For a supercomputer, some local storage is needed for work in process, programs, and system software. Permanent storage is supplied by a mass storage system and is in addition to the local supercomputer storage. A hierarchy of media is available to meet different performance requirements on a cost effective basis. For the local storage, the current access requirement is only satisfied with disks. This study and report is only concerned with this local storage. A proper system design can transfer some of the storage function from the local to the mass storage.

> ***NOTE:*** *For convenience in reading, small figures have been included within the body of this text. Full page figures are available in the on-line version of this report (in Appendix D) for the convenience of those who need to refer to larger illustrations.*

Figure No. 1 shows the memory and storage requirement for various processor speeds. This is drawn so that you enter the chart with the desired processor speed on the Y axis and pick off the memory and storage requirement on the X axis. For one petaflops, this shows One PetaBytes for memory and 1,000 PetaBytes of local storage (10^{18} Bytes). These values (and the ratios shown in Figure No. 1) will be used in cost estimates.

For comparison, the chart also shows the memory based upon a growth rate of the 3/4 power of the processing speed ($10^{0.75}$) above one gigaflop. {Since one gigaflop is one millionth of a petaflop, the assumption of a linear relationship below one gigaflop has a negligible impact on the size at one petaflops.} Since $10^{0.75} = 5.62$, the difference between the 3/4 power growth and a linear growth is only about a factor of two for a ten times faster

processor. However $(10^6)^{0.75}$ is 31,623 and this represents a factor of 31.6 less memory than the linear rate. This could represent a large difference in system cost. This kind of a cost saving could apply to specific applications and should be re-examined in further studies. A risk of this approach could be introduced by component and system designs that do not allow for the easy expansion of memory that today's designs permit.

When will Petaflops Systems be in Service?

To forecast the technical feasibility as well as the costs of systems and components, it is essential to define the time scale. Actually, the present rate of technical progress and costs could be used to determine the time when such systems are technically possible and economically viable. The general consensus is that a petaflops system could be constructed with today's technology — without any new invention.

A historical view of the progress in scientific computers is shown in Figure No. 2. This shows reaching a petaflops capability by about 2013, using the present trend. The data and the computers for Figure No. 2 do not represent microprocessor progress which has grown at a faster pace. (See Figure No. 5.) The dates forecast by experts in petaflops papers for petaflops computers employing commercial microprocessors generally fall before 2010 with some at 2007 or a little earlier. As will be shown subsequently herein, this depends upon cost and the power requirements more than the microprocessor speed.

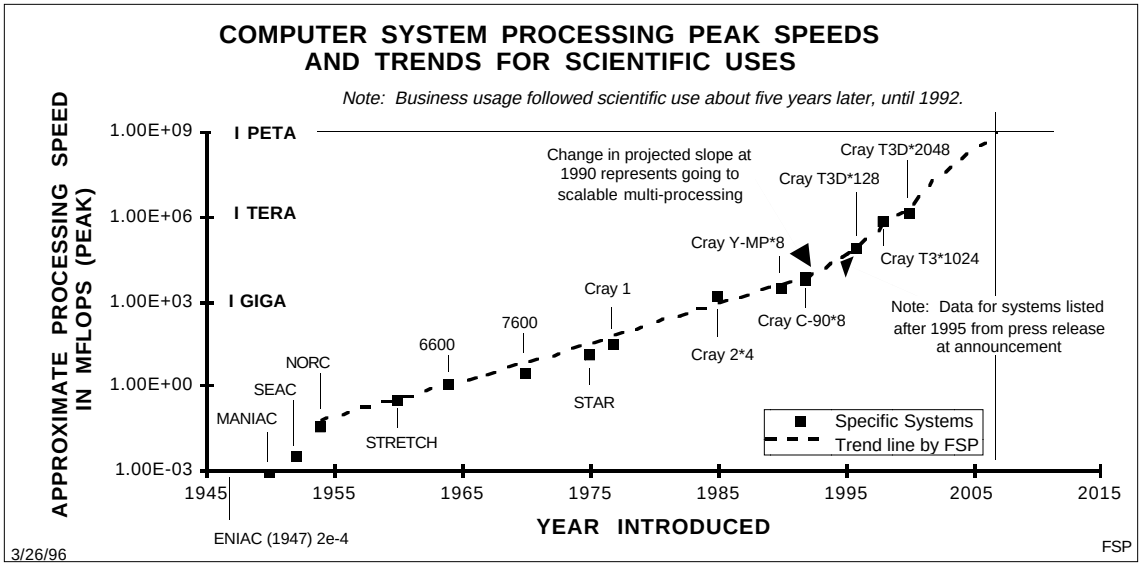


Figure No. 2 Possible Target Date for a Petaflops System

To get a rough estimate on putting together such a system with today's components, it is conceivable that the costs would scale directly with the processor speeds — using the size ratios from Figure No. 1. This results in a cost in the tens of billions of dollars ($50,000,000,000.) which approaches the gross national product or other outlandish measures, as shown in Figure No. 3

All this demonstrates is that whereas we can build such a system with current technology, we can't produce it at a cost that anyone could afford, nor would it operate within acceptable electric power consumption. *(See section on Power Requirements)* Thus, it

could be some time before a petaflops system would be available for use. In addition, as will be shown later, the number of processors would be so large that the effective rate would approach zero. Another way to look at this is that, so-far, progress on processors and memory has closely followed Moore's Law — a factor of four times the performance every 3 years or new chip generation. From now to 2010, this permits at the most five generations or a factor of about 1,000 times increase (maximum) in performance. This probably is not enough time to achieve affordable petaflop systems.

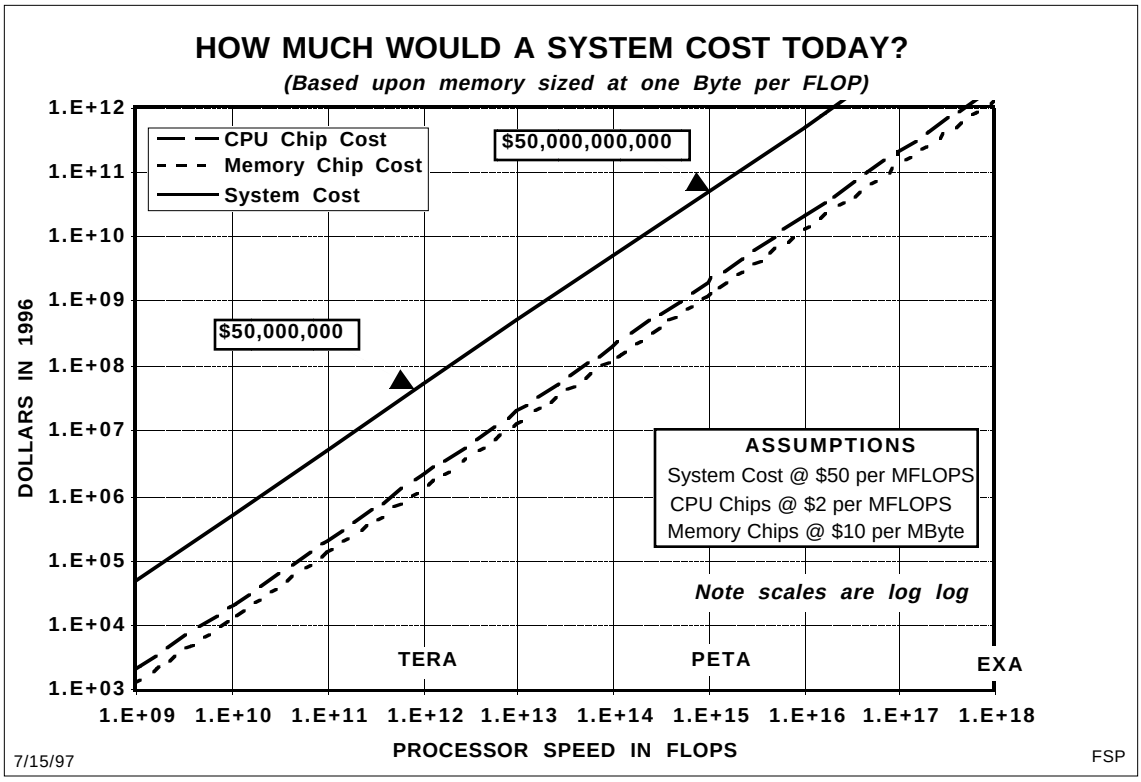


Figure No. 3 Potential Cost for a Petaflops System Today

A different way to assess the probable time before a petaflops system would be available is to extrapolate from where we are now, using the present growth rate. This projection points to the year 2014 whereas many of those working in this field say from 2007 to 2010. These dates differ by less than the probable error. It is most likely that the date will be determined by system costs and funding more than "pure" technology. Of course, it will take a great deal of technology to get the costs down to the region being forecast by those who pick 2007 to 2010.

Petaflop Component Characteristics Projections

Memory Projections (DRAMS)

The Semiconductor Industry Association (SIA) periodically issues a projection of goals which they label as a ROADMAP. This is as good a reference on future performance as any published. [SIA_95] It has the advantage of inputs from a wide coverage of semiconductor experts, with emphasis on manufacturing. It deals with the current

technology and projects the continued use and growth along these lines. It contains one major assumption that is stated in the text but is not brought out as an assumption when charts and summaries are generated. Thus, many references to their reports fail to consider the limits on their projections.

The basis for their assumption is the continued growth in performance (and lowering of costs) following the historical trends experienced over about the past 20 years. This is best expressed by Moore's Law which states that there is a factor of improvement of two every 18 months {factor of four every three years}. (Moore is one of the founders and an officer of Intel.) It is likely that this improvement goal can be continued for about seven to ten years or a factor of from 16 to 32 times the present number of elements per chip — provided the economic climate will support the huge costs associated with the necessary new production facilities.

Moore's Law has been met because of ever increasing density of memory elements, possible by increasing the resolution. The factor of four has been achieved each chip generation about every three years so the Law has also become a goal for the next step. About the year 2007, it is likely that the density will have reached the limit imposed by X-ray resolution. No published references were found that forecast a specific technology that permits higher resolution. Experts in the semiconductor industry recognize this limit and don't publicize what will transpire beyond that. It is possible to achieve higher resolutions by "tracing" out the pattern individually for each chip, but this is prohibitively expensive. Some experts take the position that when the time comes, some new technology will be found and introduced. Semiconductor chip performance will continue to improve, even if (or after) the smallest feature size limit is reached.

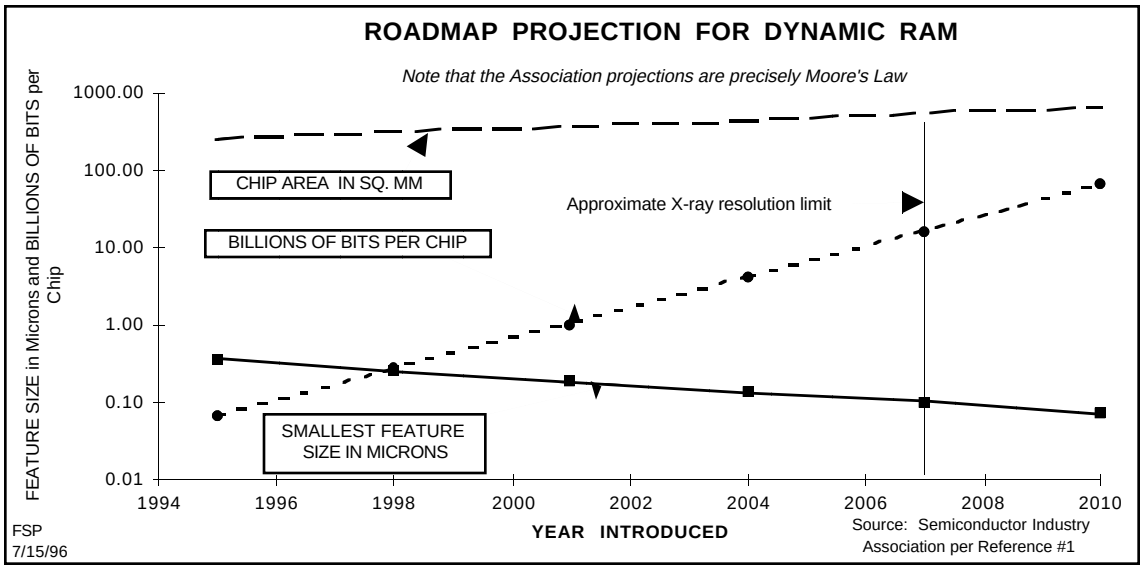


Figure No. 4 SIA Roadmap for Performance Growth for Dynamic RAMs

Figure No. 4 shows the SIA projected goals for feature size and bits per chip. To their data I have shown the present X-ray resolution limit on Figure 4. Improvements in performance can be anticipated to continue, even if the resolution is not increased.

Processor Projection

Figure No. 5 shows the SIA projected goals for feature size and bits per chip for microprocessors. (The SIA values are for high performance chips, and three values have been modified slightly to fit a smoothed line.) I have added the present X-ray resolution limit on Figure 5. Improvements in performance can be anticipated to continue, even if the resolution is not increased.

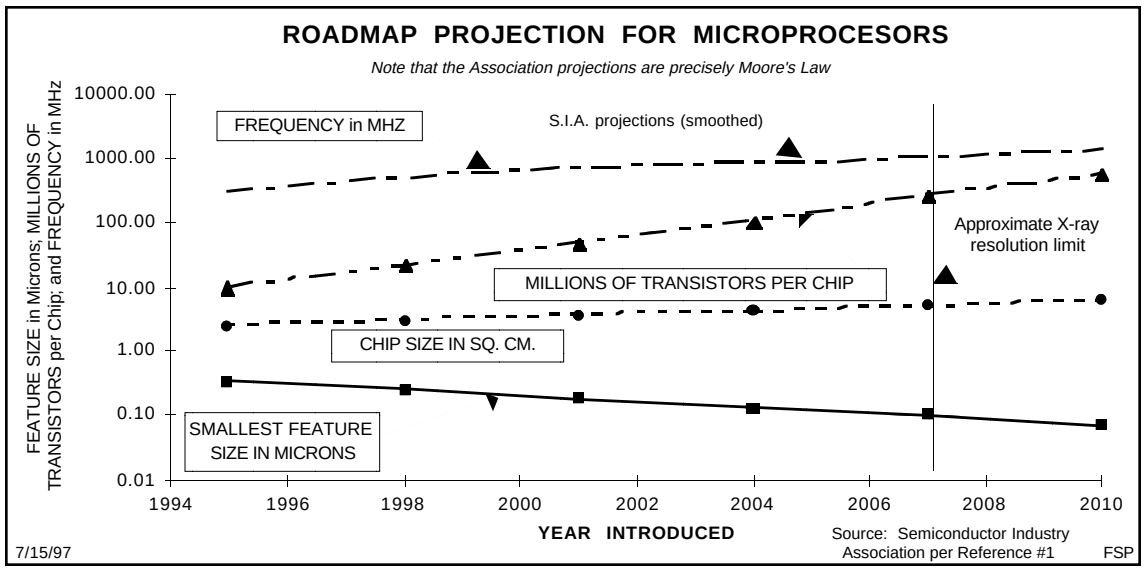


Figure No. 5 SIA Roadmap for Microprocessors

In the case of memory chips, performance has not keep up to the growth that processor chips have made. It is likely that in the future, processor growth will continue to exert a strong incentive – technical and economic – on memory chips. For processor chips, there continues at this time strong economic forces toward higher performance. Once costs have reached the level that heavily constrains improved performance measures for CPU chips, some possible gains may be continued and some dropped for cost reasons. This factor may slow up processor improvements before the level necessary for chips needed in petaflop systems is reached. That would have a profound effect on the projections. For this report, it was assumed that only the resolution limit would be reached before the petaflop era.

Other Components and the Petaflops System

The requirements for cache memories may be significant and will impact on the design and the amount of memory needed. Depending upon the architecture, this will take the form of cache within the processor chip (and/or PIM) as well as additional cache at the node with the processor and its memory. It is premature to try to estimate this now but this topic will need to be re-visited and data generated to complete the picture.

Likewise, other portions of the petaflops system must be defined to generate a complete forecast. The costs of the communications within the processing system will be significant and must also be considered after the architecture has been better defined.

Petaflop Component Cost Projections

Cost projections are even more difficult to forecast than the performance. In addition to the uncertain about the technology and the performance variables, additional unknowns are introduced by the economics, market forces, and the direction the commodity component market will take. The assumptions completely dominate the projections. Never-the-less, cost forecasts are necessary to help define what is and what is not likely for petaflops costs, timing, and capability.

Memory Projections (DRAMS)

Several different projections were made, based upon different assumptions. Certainly, none of these represent the probable future — the petaflops era is just too far away for good realism. These projections serve several useful functions. First, they show that the costs vary widely, depending upon the assumptions. Second, that even optimistic assumptions yield quite high figures. Third, the most reasonable assumptions generate exorbitant costs, confirming the opinion of some experts that memory costs will be the largest fraction of a petaflops system expense. Fourth, if the present trends for cost reductions can be maintained for the next ten years, the costs will reach a manageable sum. Finally, a petaflops system is likely to be affordable within the proposed petaflops time frame, **ONLY** if its sustained speed requirements can be meet with the mainstream commodity memory chips.

Several different schemes were employed to forecast the memory cost for a petaflops system. These will be described below but only one will be illustrated. The different approaches could yield useful models, provided the assumptions are solidly grounded. By combining the projection for feature size and bits per chip with the scaling requirement for memory (1 million billion Bytes) from Figure 1 for a petaflops system, the number of chips can be calculated. Then if some figure for the cost per chip (or MByte) is available, the cost for a petaflop system can be derived. This was done using various assumptions for deriving the cost per chip.

The first method assumed that high speed memory chips, like currently used in shared memory processors, would be needed. The number of chips was reduced as higher resolution fabrication technology was introduced. Because this type of memory would be limited in production, no reduction in the price per Megabyte was assumed. This resulted in nearly a flat cost over the years totaling in the billions of dollars.

The second method assumed that today's commodity memory chips will evolve by about the year 2010 into chips that are suitable for a petaflops computer. The technology is that defined by the SIA Roadmap and shown in Figure No. 4. The costs were assumed to start at 1995 figures and to decrease at an annual rate of 10% per year. Today's rate is faster than that and any value can be used in these spreadsheet models. This forecast starts out measured in billions of dollars for 1995 and in the order of $100 million in the time frame 2010 to 2015.

The third method uses in addition to the SIA performance projection, the SIA cost per bit (in millicents) for large volume production. Their figures go out to 2010 and to a resolution finer than the current X-ray lithography limit, which they show being reached in 2007. The memory cost for a petaflop computer was calculated from their cost per bit for one petaByte. These values are show in Figure No. 6 and amount to $40 million at the possible time for a petaflops system to be placed in service. If this is achieved, memory

costs can be manageable. Also shown on Figure 6 are two other curves for comparison. All these start from the same value. There are some that believe that semi-conductor prices will fall at the same rate as performance increases — currently Moore's Law. Assuming the cost reduction is inversely proportional to the performance increase, this results in each year being 63% of the previous year. The SIA costs are equivalent to 76% and the recent experience is about 80%. The Department of Commerce maintains a Computer Price Index and a curve fit to their values is 84%. For special applications that do not require as much memory — say for example a growth at the 3/4 power rate, an additional factor of up to 32 less memory, and cost, may be anticipated.

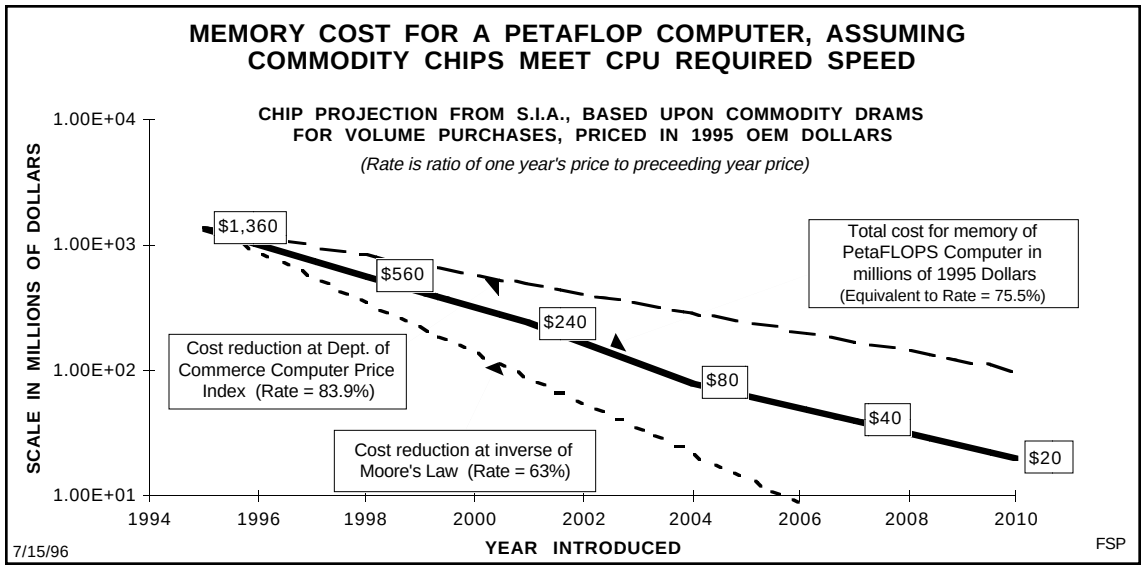


Figure No. 6 Memory Cost for Petaflops Computer System
(Derived from SIA Roadmap costs for DRAMS)

One final note: Cache is not included within these memory costs. Cache that is on the processor chips should be incorporated within that cost. Cache that is off the CPU chip could be included within the processor costs, or separately; but for this report has not been included with these estimates. Multiple levels of caching are proposed for several of the NSF point design studies.

Processor Projections

A similar methodology to that used for memory projections was employed for processors. Two different projections were made, based upon different approaches. Certainly, none of these represent the probable future — the petaflops era is just too far away. These projections serve several useful functions. First, they show that the costs vary widely, depending upon the assumptions. Second, that even optimistic assumptions yield quite high figures. Third, the most reasonable assumptions generate exorbitant costs. Fourth, if the present trends for cost reductions can be maintained for the next ten years, the costs will reach a manageable sum. Finally, a petaflops system is likely to be affordable within the proposed petaflops time frame, **ONLY** if its speed requirements can be meet with the mainstream commodity processor and memory chips.

Several different schemes were employed to forecast the processor cost for a petaflops system. These will be described below but only one will be illustrated. The different

approaches could yield useful models, provided the assumptions are solidly grounded. By combining the projection for feature size and transistors per chip with the requirement for one petaflops (See Figure No. 1), the number of chips can be calculated.

The calculation of the number of CPU chips is not as straight forward as for memory, where chip projections are defined in bits per chip. Some method of conversion from the active elements on the chip to processor speed performance must be used. I assumed that the performance was proportional to the number of transistors and the operating frequency (and in some other cases other forecast parameters). This assumption implies that all of the growth in capability is converted into a higher processing speed. It is likely that new functions will take up some of the increased component capability.

The first method assumed a chip with the SIA characteristics for 1995 a cost of $1,000 (OEM) and delivered 100 MFLOPS. This was ratioed to generate future performance based upon directly proportional to frequency, chip area, and resolution improvements. From this the number of CPUs was calculated as a function of years. This gave a figure of $3 billion for 1995. A 90% learning curve rate was applied, resulting in costs of about $400 million for 2007 and $200 million for 2010.

The second method relied more directly on the SIA Roadmap for costs. They give the number of transistors and the cost per transistor (millicents per effective transistors) so the cost per chip can be calculated. Their values result in nearly a constant value over the years of about $100 per chip. This value for 1995 is about a factor of 10 less than used in the first method and considerably less than the OEM values that were experienced in 1995. Thus, this method results in significantly lower costs for processors than the first method. I again assumed the 1995 chip to yield 100 MFLOPS per chip and then calculated the number of chips for one petaFLOP as a function of time. The number of transistors forecast by the SIA increased by a factor of 56 in the year 2010 and the operating frequency by a factor of 3.3 resulting in an overall factor of about 185. This then produces a great reduction in the number of processors required and thus the costs over this time span. This is plotted in Figure No. 7 and show costs for the processors to be about one third of that for memory (Figure 6).

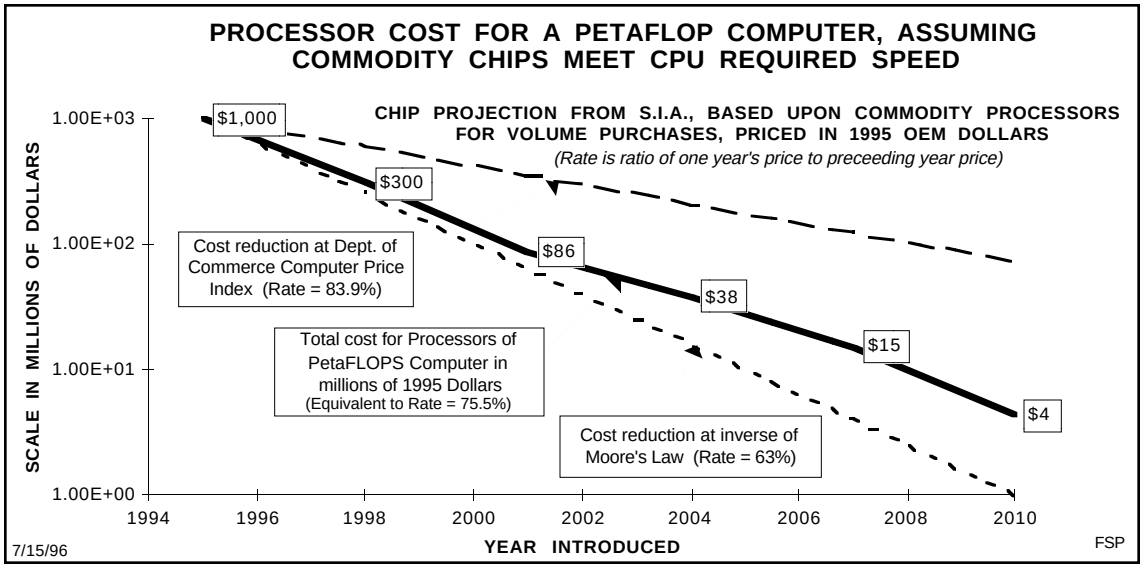


Figure No. 7 Cost for Processors, Assuming Commodity CPU Chips

Also shown on Figure 7 are two other curves for comparison. All these start from the same value. There are some that believe that semi-conductor prices will fall at the same rate as performance increases — currently Moore's Law. Assuming the cost reduction is inversely proportional to the performance increase this results in each year being 63% of the previous year. The SIA costs are equivalent to 76% and the recent experience is about 80%. The Department of Commerce maintains a Computer Price Index and a curve fit to their values is 84%. Certainly, if the SIA cost projections can be achieved, a petaflop computer may well fall within a manageable price range.

Component Performance

The charted data in Figures 4, 5, 6, and 7 do give a good picture of the significant change from today's chip performance to that of 2007. Below are given a few of the resulting values and numbers needed to reach one petaflops (PEAK). For both memories and processors the feature size would be 0.10 microns — approximately the highest possible with X-ray lithography, the best now known to be feasible.

For the memory chips, the area is 960 Sq. mm or 3.4 times today's. The chip would provide 16 Gigabits per chip or 62 times that of today. The cost would be $0.04 per MegaByte or about one 250th of today's cost (in 1995 dollars).

For processor chips, the area is 520 sq. mm or 1.7 times today's size. The frequency projected is 1,000 MHz or about twice today's. Each chip could accommodate 260 million transistors or about 12 times today. The cost per chip is forecast at $130 for 8.7 gigaflops per chip. This results in 115,000 chips for one Petaflop.

Power

This study (and report) has not addressed the electric power requirements for a petaflops system. This decision was based upon the assumption that until the architecture and components were better defined, it was premature to make such estimates. This was probably a poor decision. At the very least an estimate can be made for the power required with today's components. As a first approximation, the power needed might be in the order of 1,000,000 times that used by today's one gigaflops system. Another estimate would be to assume 1,000,000 processor chips each using 25 watts for a total of 25 Megawatts. With an electric rate of ten cents per Kilowatt ($ 100 per Megawatt), this is $2,500 per hour or $ 60,000 per day (or about $ 20 million per year). This very limited discussion does not include the problems associated with cooling. This demonstrates that the power requirements are a very significant factor and must be addressed in the architecture and design from the outset.

System Projections

With so little good hard data, it is futile to try to project the total costs of a complete petaflops system — processors, memory, cache, communications, storage, networks, etc.. One way to look at it is that high performance systems today run in the order of $250 per MFLOPS. The lowest figure being discussed today is about $100. per MFLOPS. Taking $50. as a convenient round number, this works out to $50,000,000,000 ($50 billion) which is the figure shown on Figure 2. THIS IS BIG BUCKS! The learning curves of 80% results in a cost reduction of a factor of 10. This generates a system cost of $5 billion in 2007. Further cost reductions of can come from increased performance. The SIA

forecast for chip cost reduction is about twice the reduction for the 80% currently experienced. This still results in a far higher system costs than petaflop advocates were talking about in 1996. It is from the basis of these figures that I state that the projection and actual employment of a petaflops system will be totally driven by economics. Perhaps, it will take a technology breakthrough to achieve costs that are acceptable. In this case, the date of introduction would probably slip well beyond 2015.

Certainly, the system cost projections must be based upon a more thorough look at these numbers. This should be the subject of additional analysis, after more of the architecture has been defined.

What About the Effective Rate?

System Efficiency

In today's view of achieving ever faster supercomputers, the only way to higher peak performance is with more software and hardware parallelism. This may change by the time petaflop systems arrive – say in 2010 – but can't be counted on. This parallelism appears in more processor nodes, with more processors per node, and more processors per chip. It also shows up in superscalar processors employing more instructions operating in parallel within the CPU. Further parallelism is achieved with long words to pass more data or instructions per cycle. All this parallelism imposes extra demands upon the other elements of the processing system. Higher bandwidth communications are needed, which comes at the expense of faster components, higher frequency and more parallelism in the channels. Higher access speed to memory imposes higher rates on the memory elements. Since memory speed technology already lags behind processor speeds, this is a major constraint on processor performance. To circumvent this constraint, either much more expensive memory is required, or a hierarchy of different rates — thus the growing use of caches. One level of cache is common with two or three becoming more widely used or considered in new designs. Where will this end up?

Computer system processing efficiency suffers unless the system design is balanced to avoid bottlenecks. To accommodate increasing system performance, scalable designs allow the customer to add processors and increase the memory and communication bandwidths proportionally. The computer efficiency is defined by the following equation:

EFFICIENCY = (EFFECTIVE RATE) / PEAK RATE

Even in a single processor, the efficiency is never 100%, although for some applications and situations, it can be nearly 100% if all of the data can be made available as fast as the processor can process it (input, all processing, and output per step within a single clock cycle). This becomes much more difficult with distributed memory and processing going on at many nodes. Data generated and stored at one node at one clock tick may be needed at multiple locations by the next clock tick. Message passing moves the data and takes resources at each node – using processor capability. Various topography designs are employed for this, with the highest performance being delivered by large high speed switches.

Today's best designs select the highest performance CPUs available and use the minimum number in order to keep the efficiency up as high as possible. The efficiency is also a function of the application. Data dependency varies from nearly zero; such as adding a

large set of pre-stored values; to very high; such as multi-dimensional, non-linear, partial differential, high-order, equation arrays that must be evaluated by iteration. As processor elements are added, the efficiency declines. It is possible on a single job, by going beyond the intended number of processors in the designed operating range, to reach a situation where adding processors actually decreases the total throughput. This is illustrated in Figure No. 8

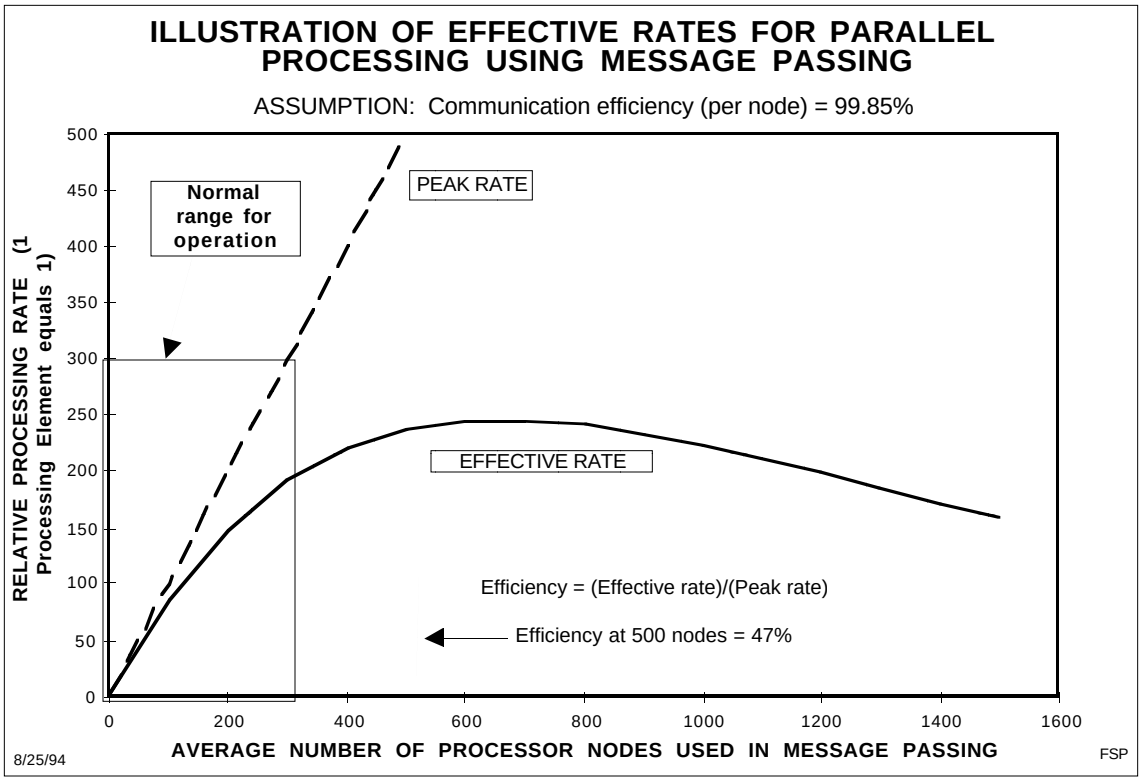


Figure No. 8 A Processing System Needs Efficient Access to Data

There may be a good reason for a system to be installed with more processors than can effectively be used on one job. Nearly all supercomputers are employed by multiple users on a variety of jobs. Many locations have far more users than are running jobs at any one time. Thus, the scalable system may be subdivided to process jobs simultaneously, with dynamic allocation of processors. This gives improved effectiveness of using these costly computers. Care in scheduling is desirable to avoid assigning an excess number of processors to a given job — even if not all processors are busy at the time.

Another constraint exists on the number of processors that can be effectively used. Some jobs can not be subdivided beyond a certain limit — the application can not use all of the possible threads of the system. This can become a real possibility when over 100,000 processors are available, as now being considered for petaflops computers.

How many Processors in One Petaflop?

Let us examine this question and look at one example. The question is easily answered if you know the speed of each CPU. Today's chips can deliver hundreds of Megaflops and

are approaching 1,000 (one gigaflops). Thus Peta divided by Giga is one million. (10^{15} / 10^{9} = 10^{6}). Now one million processors is certainly too many to consider today. What will be a reasonable number in the period 2007 to 2014? This will depend upon what they will cost and this in turn will be market driven. Therefore the question can better be posed, "What will be the processing rate of commodity chips available fifteen years hence?" No answer was found to this question.

Figure No. 7 was derived for the potential cost for processors. This assumed commodity chips and defined their performance by extrapolating growth trends. As an example, the specifications of the chips implied by the data used for Figure No. 7 results in a possible 8.7 gigaflops per chip corresponding to about 115,000 CPU chips. These values are derived from figures that assume a faster rate of growth than many experts have used in their forecasts. However, the value of about 100,000 processors is in the range estimated by others. It is likely that before that time, the business community will support a market for CPU chips in the range of one to ten gigaflops. There is no assurance that a ten gigaflops chip will actually be made by the time needed for a petaflops system.

What will be the Effective Processing Rate?

If we assume that the present processing efficiencies can be maintained for the hundreds of thousands of processors, some estimate can be formed for the effective rate. Traditionally, shared memory supercomputers have operated in the range of from 30 to 50% efficiency for well-developed programs. Since a number of applications are now only able to get a few percent efficiency on scalable, parallel systems using tens to hundreds of processors, it will be a real achievement to obtain even 5% for 100,000 or so processors. Special tuning of applications can improve the efficiency but COTS software or "dusty decks" will be worse. This will not be satisfactory, of course. For this analysis, a rather arbitrary figure of 5% was used for the parallel, scalable processors operating in the gigaflops range. This is assumed to rise to 13% by the time petaflop processors might be available in 2007. These were the assumptions used to generate Figure No. 9. The dotted curve was presented before in Figure No. 2. (Figure No. 2 represents my projection for when petaflop systems would be in service. Figure No. 9 shows a faster rate corresponding to the SIA Roadmap values which I find quite aggressive.) The solid curve numbers are these values multiplied by the assumed efficiency.

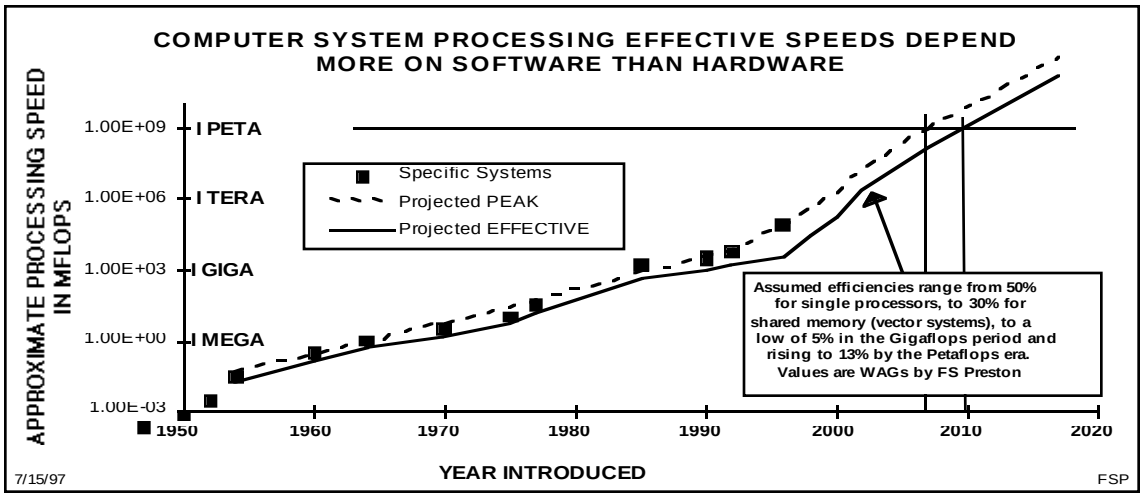


Figure No. 9 Effective Rates Could be Far Below Peak Rates

This doesn't look too bad on a semi-logarithmic plot scale. It looks as if the delay from one petaflops peak to one petaflops effective is only about three years. However, since the cost of a system is nearly directly proportional to the speed, the 13% efficiency translates into a factor of 7.5 times the cost to achieve one petaflops effective versus peak. Let's leave this at this point. It is not a question of how accurate is the estimate of the example, but rather that nearly any reasonable estimate will demonstrate that the difference is HUGE. It could be even worse than presented. The moral of this story is that something must be done to get the efficiencies into the usable range of about 20 to 25%. This is probably possible with large investments in software.

This ends the section on Study and Analysis. It would be nice to find more and better data on present technical and cost projections and specific data on the steps from teraflop to petaflop systems. Without more data, further analysis is probably of little value. There is considerable written on Software in general, but specific references on software for petaflop systems were not available until after 1996. *See the section on Software which refers to these*. The next section (Other Topics) will cover Software and other related subjects; and the following section (Summary) will address some Questions, Issues, and Problems — followed by some Conclusions and Recommendations.

OTHER TOPICS

Software

Developments in software must accompany progress in hardware. For a factor of 1,000 or more growth in system performance, much new and improved software will be needed. This study failed to obtain any specific petaflops software references up to the end of the literature search in December 1996. Subsequently, several papers became available, including one from 1996 and several from the beginning of 1997. These are Ref. 9, Ref. 10, and Ref. 11 in the References section of this report and are also included within the Bibliography. This report does not address the specifics covered in these three references and will leave for a later report a more detailed analysis of the software for petaflops systems. Rather, herein are discussed just some evidences of the issue that software progress continues to lag behind the hardware. This is particularly evident in the HPCCP computer programs..

There are abundant references available on software for scalable systems and the general software problem. The modern computer period, starting in the mid-1980s, commenced with software already lagging behind the hardware progress. Since then, software has failed to either close this gap — or even keep up. The hardware/software system balance has become worse. There is every reason to expect this trend to continue for the next fifteen years. There is grounds for some hope and evidence that this problem is being recognized and addressed by those pushing toward petaflops

A previous section on Efficiency has emphasized the major influence of this on performance and projections for petaflop systems. The degradation in performance is a combination of the data dependencies of the application as well as the computer architecture and design. These in turn are controlled by the hardware and software. For hardware, the

system's communication bandwidths and the latency in memory access are the major controlling functions. Software is a major contributor that impacts on the time needed to fetch, and later store results. If the computer is waiting for data and engaged in transferring data, it is not producing results and the efficiency suffers. With the advent of scalable and massively parallel computer systems, system architects have gone to more levels of cache to reduce the latency. The algorithms, operating system, and program software all influence how well the computer uses its time in processing. The problem is worse for petaflop systems than for teraflop or gigaflop systems. No satisfactory solution is available for jobs that can not use 100,000 threads that are to be run on systems with more than that.

It is conceded by supercomputer system architects, designers, and users that the algorithms now in use for computationally demanding jobs will need to be refined or replaced for the petaflops era and beyond. The same can be said for the operating systems and programming languages. This is a massive effort and should be part of the petaflops Initiative. Instead, the limited references that approached this subject expressed the opinion that software developments logically follow the hardware development rather than proceed in parallel. This could be a valid point if some revolutionary architecture and hardware is to be used. It is NOT valid for an evolutionary path – which the majority believes will occur. Certainly, as of 1997, we can be sure that software will not be available to match the hardware schedule. Will petaflops systems use WINDOWS '10 in 2011?

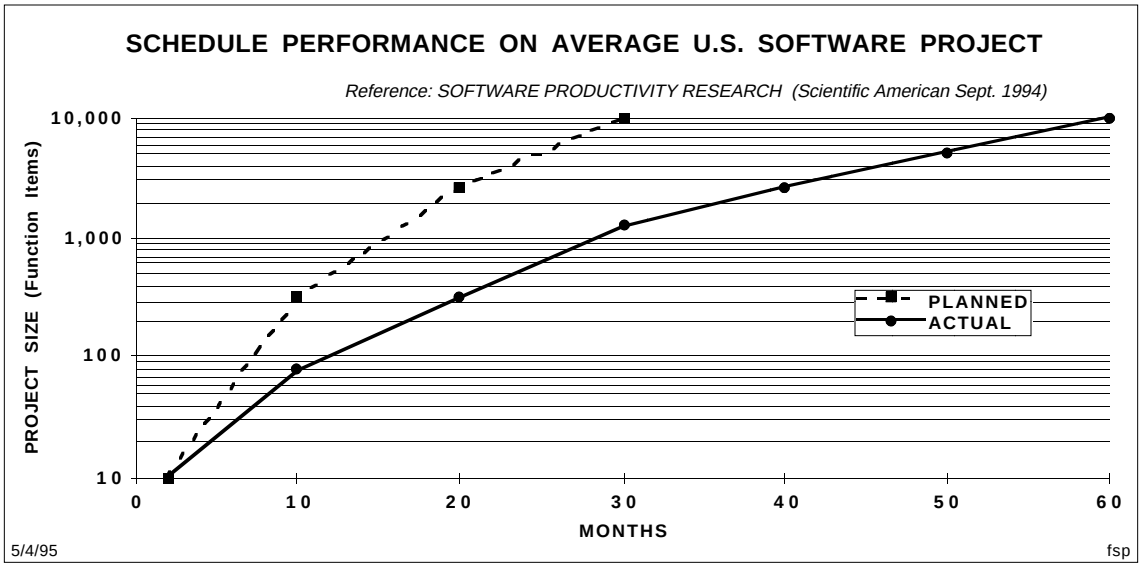


Figure No. 10 The Average Software Project Will be Completed Very Late

A look at the recent performance in the software arena is warranted. Figure No. 10 shows how well – or poorly – the average software development project in the U.S. meets its delivery date. Note that the bigger the project, the greater the lag. Supercomputer software and operating systems are BIG projects where the lag is more than two years. Certainly an actual delivery that takes twice as long as the plan is a very poor record.

The record is worse than shown in Figure 10. In the size range of supercomputer software projects, from 25% to 50% of the projects started are canceled and not completed. This is shown in Figure No. 11.

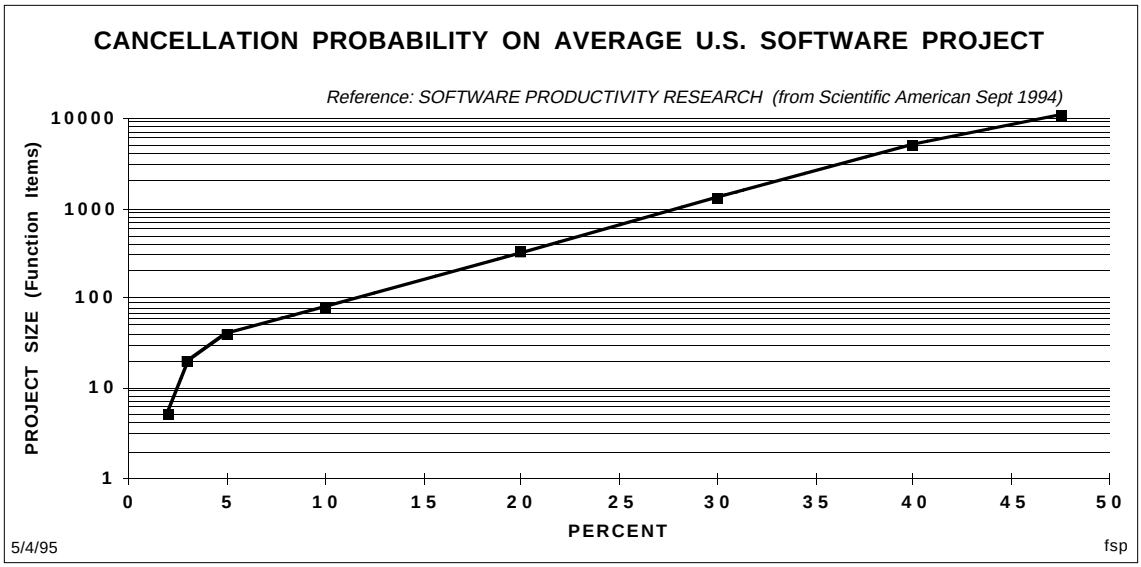


Figure No. 11 The Bigger the Program, the Less Likely it Will be Finished

The software picture is not all bleak; there are some favorable trends. Some architecture and chip developments have somewhat eased the load on the programmer. Schemes have been introduced that allow distributed memories to be addressed as if they were shared memories. More use of superscalar designs with more on-chip parallelism is in vogue. Out-of-sequence processing is now possible, as are very long instruction words. Compilers have improved. Some of these improvements aid the user/programmer community at the expense of more complex chips and compilers — transferring some of the software problems to the manufacturer. It is not clear how these trends will be applicable to the type of hardware that petaflop computer systems will employ.

The software now in use is often poor — full of errors, mistakes, and bugs that go undetected for long periods. As an example to conclude this software section, the French Ariane rocket was recently destroyed at launch as a direct consequence of a software error. This cost tens of millions of dollars as well as set back the programs that were dependent upon the payload. This also had a negative impact on the Ariane program. (The problem was traced to inadequate testing of a program. [Ref. 8])

This section on software should not conclude on a negative tone. Software, particularly algorithms but also programming technology, have contributed about the same improvement and performance growth over the years as has hardware improvements. The factor of about 100,000 increase needed to achieve petaflops is not depended solely on hardware. Software advances can be expected to contribute a portion — specifically in improving the effective speed by better efficiency.

FRONTIERS Conference

Up through the end of 1996, there have been very few open, technical conferences on petaflops, although there have been conferences on the HPCCP Initiative program which is indirectly related. There was a conference, FRONTIERS 95 [NASA95], sponsored by NASA-Goddard. (Proceedings were gathered into an informal publication.) In 1996 this conference was followed up by FRONTIERS 96 [IEEE(b)96] which combined two topics:

1) Massively Parallel Computing; and, 2) Frontiers 96 Symposium and Workshop. Up through 1996, this was probably the most public conference that addressed petaflops computers. Preston attended this meeting and prepared a report [Ref. 6]. Because of the significance of the meeting, some observations brought back will be discussed below.

The petaflops emphasis was on the status of the Point Design Studies, about eight projects funded by the National Science Foundation (NSF). These designs and studies are scheduled for about five to seven years (but funded annually). Their goal has been limited to 100 teraflops on the basis that "... none of them will mature to one petaflops." NSF expects that the designs will then merge. One of the studies involves superconductor computers to operate in the 100 to 500 GHz range. They say that they "... have a petaflops computer prototype operating". The projection of about 66% of the attendees was that petaflops would be reached by an evolution of present systems (i.e. now 1 teraflops); and, the about 33% of the remaining forecast this by revolution. *[Preston expects to see it by revolution by 2015, achieved by following commercial developments and markets that develop evolutionary.]*

Some conclusions were stated at the meeting, or may be drawn from round-table and other discussions. "Memory will be the largest cost item." "Software {development} will be the largest system cost." "petaflops systems will require users to have teraflops performance in local workstations." "There is no question about achieving a teraflops capability."

Alternate Computer Technologies

If petaflops computer systems arrive via a revolution, it will come out of alternate approaches with different technologies. This is discounted by the majority. References to alternate technologies did not couple their efforts and progress with petaflops. Most experts state that a different technology will take more than 20 years to be introduced. This is then consistent with the viewpoint of an evolutionary development on a time scale of from ten to fifteen years.

There is considerable research being devoted to alternative computer technologies. Nearly all of this is basic and does not appear to be ready to move into the development or prototype stage. These will just be mentioned here as the start of several "threads". Optical computing has been "just over the horizon" for about ten years. Its early promise is waiting for good optical-to-electric interfaces. Superconducting computer components and computers have considerable promise and some workable elements. This promise seems to be held back by high costs. What it needs is a super-application that can pay for the early introductory costs. Perhaps petaflop systems could furnish the support — the same way previous supercomputer generations subsidized the introduction of new technologies into the main stream.

Without necessarily requiring new technology, there are a number of special purpose computers in operation and a few under development. At least one of these is funded under the NSF Point Design Studies program. *[I believe that there should be more effort along these lines. I have considered and recommend looking at the Digital Differential Analyzer (DDA) approach. This technique is particularly powerful for difference equation solutions which are hogs of computer resources. This has possibly two advantages: 1) it can operate in near real time; and, 2) much of the data is retained within the processor, with less need for retrieval or memory.]*

This study and analysis program looked for alternate technologies but did not attempt to go into these in depth or analyze their potential. Thus, this report contains only sparse references to this area. Further, follow-on study and analysis could plunge deeper into these fields. This may be productive but it would also be difficult. Good basic and general papers on the various areas are hard to find. Further, many widely different technologies are involved. To properly review and report of these many approaches, one should be working in the field. Therefore, a proper review would best be handled by experts in each area, who already keep up with the papers in their respective fields.

If the supercomputer community is serious about achieving petaflops, it should endorse and support parallel paths of evolutionary and revolutionary designs and prototype use.

SUMMARY

Questions

When this study and analysis was started, there were several questions that immediately came to mind. The literature search and discussions have not developed the answers — so far.

1.1) Can (or will) the commodity memory chips keep up with the speed requirements of commodity processor chips that will be used in the 2010 time frame? ..... or will this be a limit for supercomputers?

1.2) Will commodity processor chips be suitable to use in petaflops systems? (Examples of this are power requirements and cost.) ..... or will this be a limit, or will the supercomputer arena need to develop its own custom chips?

1.3) What will the software environment be in the 100 teraflops systems and faster era? What will be the operating system? What will be the programming language that is suitable? How will the software effort evolve and how funded?

1.4) What follows petaflops? ..... will exaflop systems evolve from petaflops, or is it probable that petaflops efforts will not evolve to exaflop systems?

Problems Which Become Issues

Naturally, the road ahead is strewn with problems, that will have to be settled in time. A few were exposed that are not settled for existing supercomputer systems and no resolution of these appears eminent. Some are listed below:

2.1) NASA's (and the Federal) budgets for the next ten years are projected to decrease. Where will the funding come from for petaflops systems? ... will they develop from the commercial area? ..... for what applications? ..... will these systems be suited to the scientific requirements?

2.2) Related to the above, existing systems are compromised by inadequate support personnel and peripheral systems. How will this ongoing constraint be resolved?

2.3) The electric power requirements of a petaflops system must be addressed as a part of the system architecture and the selected component technology.

Conclusions

This study has only reached an intermediate milestone and is incomplete. Nevertheless, some conclusions come to mind that are worth emphasis at this time. They must be re-examined in the future after further study.

3.1) Economics will define what is available and possible for the foreseeable future. Computers are now in and likely to remain a commodity business. Scientific computers led the commercial until recently. This is in the process of swinging in the opposite direction.

3.2) Petaflop computer systems are technically feasible with no new inventions. Such computers will not be affordable nor operate within practical power limits, so technology improvements will be needed to keep improving the cost-effectiveness of components and systems.

3.3) Software and algorithms will present much more of a challenge than hardware.

3.4) Systems must be balanced, which will entail increases in support.

3.5) Workstations must be sized in some proportion of the supercomputer (petaflops) system that serves as a "co-processor". Certainly the current workstations are not fast and powerful enough to work effectively with a petaflops computer. As an approximate size, we must consider workstations in the teraflops range. By the time petaflop systems are ready to be deployed, such workstations – or something equivalent – must be available. These will probably not be exorbitant in cost because the same cost economies that apply to the petaflops system components will be applicable.

3.6 Development toward petaflops systems are not following the tried and true method of new developments that are faced with problems that demand breakthroughs or face delays and/or unbalanced systems. More parallel attacks on the bottleneck problems has been the mode for success in the past.

3.7 If not now, soon the computer field must awake to the realization that our present digital computer approach has absolute speed limits. If these limits are not reached by petaflops, what about exaflops or the next step beyond that? A portion of the computer science effort needs to be directed at a radically different basis for calculations.

Recommendations

The author makes the following recommendations:

A) The Petaflops Initiative should be achieved in three stages rather than considered as one step. These stages should be as follows — on the assumption that we are currently moving into the one teraflops zone:

Stage 1) 10 teraflops
Stage 2) 100 teraflops
Stage 3) 1,000 teraflops = 1 petaflops.

B) NASA should produce and publish a 15 year plan consistent with recommendation A above. This plan should show how to get from now to the petaflop era. This plan may need to be revised and re-issued, as circumstances warrant. The purpose is to get users and management all moving together. Efforts not consistent with the plan should be curtailed. Algorithm and application development need to be emphasized.

C) NASA, and the supercomputing community at large, should initiate alternatives to an evolutionary path toward petaflops — as a parallel program justified because of the risk of failure without something revolutionary.

D) The software and hardware total effort should be considered as a total package and allocated to achieve the maximum from a cost-effectiveness standpoint. Based upon past experience, this division should result in about equal funds for hardware and software.

E) This study and analysis program at Langley should involve more personnel than the one or two so far assigned. Much more contact and working with similar efforts at Ames and other NASA centers should be established and maintained.

ACKNOWLEDGMENTS

Because the author works on a part-time schedule, he largely worked alone on this analysis. This work was carried out under the general direction of Geoffrey Tennille, Technical Monitor. To get a better review of this report, about four preliminary copies were submitted to scientific supercomputing specialists. He gratefully acknowledges the comments and constructive criticism received from Thomas Sterling (NASA JPL and Cal. Tech.) and from David Bailey (NASA ARC) — both of whom are active on the petaflops initiative. David Adams of the Langley Library assisted with search for source material.

Personal Notes

I am more pessimistic about the time scale for and the economic viability of petaflop systems than most of the experts in this field. I have tried to present the majority viewpoint and only interject my pessimism as a cautionary comments.

The writer works on a two-day per week work schedule. During most of the time that this study took place, he had additional tasks. Therefore, this analysis which took place over about nine months, represents only a fraction of this in time available for this report. This part-time schedule did allow for the necessary waiting time for searches and documents. It also spread out the working time enough to allow time for thinking and review.

ACRONYMS

EXAFLOPS	10^{18} (a billion billion) Floating Point Operations per Second
FLOPS	Floating Point Operations per second
GFLOPS	Giga (1,000 Mega) FLOPS, or 10^9 FLOPS
HPC	High Performance Computing (synonym of supercomputing)
HPCCP	High Performance Computing and Communications Program
LaRC	Langley Research Center (NASA)
MFLOPS	Mega (1 million) FLOPS, or 10^6 FLOPS
NSF	National Science Foundation
PETAFLOPS	Peta (1,000 Tera) FLOPS, or 10^{15} FLOPS
PIM	Processor-In-Memory (memory and processor on same chip)
SIA	Semiconductor Industry Association
TFLOPS	Tera (about 1,000 Giga) FLOPS, or 10^{12} FLOPS

DEFINITIONS

The terms below need definitions because differing usage is in vogue in the computer industry. Usage of some of these have been changing and could change further in the future. The definitions are the author's who has tried to reflect current understanding. Some current use of these terms is wrong or misleading.

Cluster	An aggregation of processors which are interconnected and operated as a shared resource capable of distributing the processing of tasks among the cluster elements or with processors individually dedicated to separate job. Clusters may consist of either homogeneous or heterogeneous elements. Cluster elements may be comprised of more than one processor. Clusters generally may be employed in a scalable parallel fashion on problems with loosely coupled data dependencies.
Effective Rate	The speed of the system for a specific application or benchmark, which must be stated with the rate to be meaningful. (The effective rate changes with the application.)
Efficiency	Ratio of Effective rate to peak rate. This efficiency is made up of at least two major factors: 1) the efficiency of the computer system; and, 2) the effectiveness and efficiency of the language, program, and algorithm used. The efficiency of the computer

system depends upon the access time (latency) for data, the amount of data to move, and the communication rates.

Peak Rate — The operation instruction rate based upon the number of operations per clock cycle, the clock rate, and the number of processors involved. *This rate is the value that the manufacturers guarantee will not be exceeded.* The value is expressed in floating point operations per second, and now most often in MegaFLOPS.

Scalable — *Also sometimes spelled SCALEABLE*. To be scalable, the processing power, memory, local storage, and communication bandwidth must all increase proportionally with increased numbers of processing elements. Scalable use requires program code that is scalable. Scalable also requires scalable operating system software. To be useful in a scalable mode, the external communication rates and ports and support functions should be sufficient to not limit overall performance.

Applied to computers and systems, this implies that a number of computers may be assigned to work on a single problem. The number shall be variable and selectable at run time. In some cases, the number can be dynamically changed during different steps in the run.

The problem (application) must be scalable. This implies that the number of processors assigned varies linearly with the growth of the problem. Speed up should be nearly proportional to the number of processors assigned. A scalable computer system also implies that the system can be subdivided to serve a number of separate users.

Super Scalar — If the Instruction count (or operations) per clock cycle is greater than one, the processor is classed as super scalar. Values of up to about 3 per clock tick are now in use in systems.

Supercomputer — The latest and fastest (scientific) computer available — or more probable: the next generation beyond the computer you now have.

Workstation — A processor with a display monitor and software suited to the tasks submitted to it. The monitor is located at the workplace of the user. The processor may be located at the user's station or remotely. Workstations can be operated stand-alone but are more generally attached to a network with access to support services. The user nearly always maintains control over the workstation files and resources.

REFERENCES

For this report, references are listed in the approximate order that they were used and therefore not in any organized sequence. They are listed by number which is used to designate the reference on charts and copies of the data (within spreadsheets). Not all data used is covered by these specific references. Since data was used wherever available, as long as it appeared to be significant, some of the references are not readily available from the typical library sources. The reader is referred to the annotated BIBLIOGRAPHY for this study for additional information on these and other references. The sources in the Bibliography will be found under the date published, and by author, as explained in the section on References in this report. (Example: Name95)

1) The National Technology Roadmap for Semiconductors Semiconductor Industry Association (S.I.A.) (See Bibliography SIA_95)
Quoted in IEEE SPECTRUM Jan. '96 page 53 (See Bibliography Semi96)

2) Future Technologies and Economic Prospects for VSLI H. Homya IEEE ISSC '93 (from FSP Report #23 Bibliography)

3) Computer Trend Analysis (unpublished Report No. 23) FSP 1994 Figure No. 5

4) Office of Technology Policy Technology Administration, U.S. Dept. of Commerce 1994

5) U.S. Office of Science & Technology Policy 1991

6) Report on FRONTIERS '96 Conference of Oct. 27-31, 1996 at Annapolis, MD
Memorandum of conference by F.S. Preston dated Nov. 5, 1996

7) Software Productivity Research Scientific American September 1994

8) Design by Contract: The Lessons of Ariane IEEE Computer January 1997 pp 129-130

9) The Petaflops Systems Workshops Proceedings by Cal. Tech. of two workshops as listed below. (See Bibliography Fost96)
PetaFlops Architecture Workshop "PAWS '96" held April 1996
PetaFlops System Software Summer Study "PetaSoft '96" held June 1996

10) First Workshop on Hybrid Technology Multithreaded Architecture {HTMT} for Very High Performance Computing Sterling, et al Conference Notes: Cal. Tech and JPL for meeting of Feb. 24-25, 1997 (See Bibliography Ster(a)97)

11) 1997 Petaflops Algorithm Workshop (PAL '97) (See Bibliography Vari97)

APPENDICES

Note: The appendices listed below are included within the on-line version of this report; but not within the hard copy version.

A: BIBLIOGRAPHY (Annotated)
B: KEYWORDS
C: CATEGORIES
D: REPRODUCTIONS OF FIGURES (Full Size)
E: SPREADSHEETS (used for charts)

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 1998	Contractor Report

4. TITLE AND SUBTITLE
A Petaflops Era Computing Analysis

5. FUNDING NUMBERS
C NAS1-20048
WU 509-10-21-01

6. AUTHOR(S)
Frank S. Preston

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Computer Sciences Corporation
3217A North Armstead Avenue
Hampton, VA 23666

8. PERFORMING ORGANIZATION REPORT NUMBER
Report No. 25

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Langley Research Center
Hampton, VA 23681-2199

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
NASA/CR-1998-207652

11. SUPPLEMENTARY NOTES
Langley Technical Monitor: Geoffrey M. Tennille
An electronic version of this report can be found at: http://techreports.larc.nasa.gov/ltrs/ltrs.html

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified-Unlimited
Subject Category 62 Distribution: Nonstandard
Availability: NASA CASI (301) 621-0390

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

This report covers a study of the potential for petaflops (10^15 floating point operations per second) computing. This study was performed within the year 1996 and should be considered as the first step in an on-going effort. The analysis concludes that a petaflop system is technically feasible but not feasible with today's state-of-the-art. Since the computer arena is now a commodity business, most experts expect that a petaflops system will evolve from current technology in an evolutionary fashion. To meet the price expectations of users waiting for petaflop performance, great improvements in lowering component costs will be required. Lower power consumption is also a must. The present rate of progress in improved performance places the date of introduction of petaflop systems at about 2010. Several years before that date, it is projected that the resolution limit of chips will reach the now known resolution limit. Aside from the economic problems and constraints, software is identified as the major problem. The tone of this initial study is more pessimistic than most of the Superpublished material available on petaflop systems. Workers in the field are expected to generate more data which could serve to provide a basis for a more informed projection. This report includes an annotated bibliography.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Supercomputing, High Performance Computing, Scientific Computers, Future Computers, Massively Parallel Computers, Supercomputer forecast	38
	16. PRICE CODE A03

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	