Stanford University

C I S A C

Center for International Security and Arms Control

The Center for International Security and Arms Control, part of Stanford University's Institute for International Studies, is a multidisciplinary community dedicated to research and training in the field of international security. The Center brings together scholars, policymakers, scientists, area specialists, members of the business community, and other experts to examine a wide range of international security issues. CISAC publishes its own series of working papers and reports on its work and also sponsors a series, *Studies in International Security and Arms Control*, through Stanford University Press.

Center for International Security and Arms Control
Stanford University
320 Galvez Street
Stanford, California 94305-6165

(415) 723-9625

http://www-leland.stanford.edu/group/CISAC/

Building on the Basics: An Examination of High-Performance Computing Export Control Policy in the 1990s

Seymour Goodman, Peter Wolcott, Grey Burkhart

A Report of the
Center for International Security and Arms Control

Stanford University

November 1995

The Center for International Security and Arms Control, part of Stanford University's Institute for International Studies, brings together Stanford researchers from several scholarly disciplines with senior specialists from around the world and pre- and postdoctoral fellows for research projects, seminars and conferences, and international scholarly exchange. The Center publishes its own series of reports and papers on its work and also sponsors a series, Studies in International Security and Arms Control, through Stanford University Press.

The Center is grateful to the Carnegie Corporation of New York for supporting this project. The opinions expressed here are those of the authors and do not necessarily represent positions of the Center, its supporters, the United States government, or Stanford University.

Center for International Security and Arms Control
Stanford University
320 Galvez Street
Stanford, California 94305-6165
(415) 723-9625
Douglas Peckler, Editorial Associate

ISBN 0-935371-39-7

Abstract

This paper reports the results of a study of the special export control regime for high-performance computers. The history and purpose of this export control regime are reviewed, and a framework for analysis is established, which can be used to test the basic premises on which the control regime rests and to suggest viable control thresholds. The fact that the export of certain computer systems cannot be effectively controlled is established, and the limits of controllability are defined. U.S. government applications for high-performance computers are reviewed with respect to the requirement for and criticality of such computing for national security. Finally, judgments are made as to the levels of control that are possible, and the desirability and feasibility of maintaining Such controls. Near- and intermediate-term problems that may erode the liability of the basic premises underlying high-performance computer export controls are identified.

Acknowledgments

This study was sponsored by the Assistant Secretary of Commerce for Export Administration and the Deputy Assistant Secretary of Defense for Counterproliferation Policy, and funded by a contract from the U.S. Department of Commerce. Support was also received from the Carnegie Corporation of New York, the Center for International Security and Arms Control, Stanford University, and the University of Arizona, Ann Danowitz, of the University of Arizona, and Diane Goodman, of the Center for International Security and Arms Control, provided administrative and research assistance. Cameron Binkley, also of the Center for International Security and Arms Control provided research assistance and also drafted much of the first chapter. Lauren Rusk and Douglas Peckler provided editorial support.

The Authors

Seymour E. Goodman is Professor of Management Information Systems and Policy at the University of Arizona and Carnegie Science Fellow at the Center for International Security and Arms Control, Stanford University. His research interests include international developments in information technology and related public policy issues.

Peter Wolcott is an assistant professor in the Department of Information Systems and Quantitative Analysis at the University of Nebraska, Omaha. His research interests include high-performance computing developments throughout the world, particularly in the former Soviet Union and People's Republic of China.

Grey Burkhart is the president of Allied Communications Engineering, Inc., a consulting engineering firm in Fairfax, Virginia.

Table of Contents

List of Figures

List of Tables

Executive Summary

Export controls on dual-use technologies and products have been part of U.S. national security policies since the 1940s. Since the end of World War II, they have been applied with remarkable consistency, regardless of the makeup of Congress or who is in the White House. Export controls on high-performance computing (HPC) systems are implemented by determining a threshold definition of a "supercomputer," based essentially on a measure of processing power-Composite Theoretical Performance (CTP), measured in millions of theoretical operations per second (Mtops). Extraordinary licensing and safeguard conditions may be placed on the sale or transfer of any machine at or above that threshold.

Three basic premises underlie this policy:

1. That there are problems of great national security importance that require high-performance computing for their solution, and these problems cannot be solved, or can only be solved in severely degraded forms, without such computing assets.

2. That there are countries of national security concern that have both the scientific and military wherewithal to pursue these or similar applications.

3. That there are features of these computers that permit effective forms of control.

If the first two premises do not hold, there is no justification for the policy; without the third, no effective implementation is possible. A strong case can be made that all three premises held during the Cold War, and that export controls on computing were an important and effective element of U.S. national security policy.

Since the end of the Cold War, the environment for this policy has changed in three significant ways:

i. The nature and extent of foreign threats to security have changed. With the demise of the Soviet global superpower, the threats to national security are individually smaller in scale, but have become more numerous and varied.

ii. There have been dramatic changes in computing technology. In particular, powerful microprocessor, workstation, and networking technologies have developed rapidly over the last half dozen years, greatly expanding the forms of and accessibility to HPC.

iii. The uses of HPC have expanded significantly within the U.S. national security community. In particular, applications to the development and operation of advanced, high-performance, conventional military systems have been growing rapidly.

The purpose of this study is to analyze the continued viability of the three basic premises under the conditions of the changing environment, and to use this analysis to produce a threshold value that satisfies the premises. More generally, this study represents a further step in the evolution of the export control regime toward a more factual, objective, and repeatable process of policy formulation.

The basic premises can be tested by deriving a range of candidate threshold values for which the premises are true. If no such range exists, then the premises do not hold and the viability of the export regime must be questioned. A lower bound on the CTP threshold value can be derived from an analysis of the computers available in countries of national security concern and the factors that make it essentially impossible to control effectively the transfer of certain levels of computing power. These factors include computer power, scalability, size, numbers manufactured, number and forms of the primary sources of the technology, number and forms of distribution channels, and product development times. We believe the resources that government and industry can effectively bring to bear cannot control the international diffusion of computing systems with performance beneath this level. Furthermore, the premium paid in time, effort, money, and know-how by countries seeking to circumvent the controls diminishes rapidly. Attempts to control beneath this level would become increasingly

ineffectual, would harm the credibility of export controls, and would unreasonably burden a vital-sector of the U.S. computing industry.

Our analysis produces a lower bound mid-1995) of 4,000-5,000 Mtops-- which is likely to rise to approximately 7,500 Mtops By late 1996 or 1997 and exceed 16,000 Mtops before the end of the decade. In addition to the factors listed above, these figures take into account the time needed for mature markets to develop, currently no more than two years from product introduction. Computer aggregation technologies (e.g., networking, clustering), other than the Massively Parallel Processor (MPP) and shared memory Symmetric Multiprocessor (SMP) architectures, were only minimally considered beyond what is currently covered in HPC export controls. These are rapidly developing technologies that should be given more attention in any longer term analysis of controls.

An upper bound is determined by national security applications requiring the use of HPC systems whose performance lies minimally above the lower bound. Setting the export control threshold above this level would de-control the computer technology needed to carry out these applications. Four categories of applications were considered: nuclear weapons; cryptology; the development (usually design) of advanced, high-performance conventional weapons (e.g., stealth aircraft, armor-piercing projectiles);

and the use of HPC in operational systems (e.g., C 4 1, defense against anti-ship cruise missiles [ASCM]). The first two categories are the traditional applications that justified HPC export controls during the Cold War. There seems to be a group of research and development applications starting roughly at the level of 7,000 Mtops, and a group of military operations applications at 10,000 Mtops.

The latter includes such applications as weather forecasting with a precision and over a time period useful for military operations, the design and testing of acoustic sensor systems, defense against ASCM, and battlefield surveillance.

The framework for this analysis could be used to update both bounds in light of changing technologies and applications. Given shortening product cycle times in important parts of the computing industry, we believe this should be done no less frequently than every twelve months. In the past, the policy has been reviewed infrequently, forcing the continuation of outdated threshold values on industry. Reviews tend to be put off by the government until a great deal of contentious pressure builds up from industry.

Time constraints were such that we were not able to do a comprehensive review of applications. Nevertheless, enough information was collected to make the conjecture that, for various reasons, the majority of national security applications of HPC are already possible (at least from the standpoint of the necessary computing) at uncontrollable levels, or will be so before the end of the decade. Furthermore, many of the most important U.S. applications exist at or are gravitating toward that level of computing for a number of reasons. Many of these applications are essentially uncovered by export controls because the measure used, namely CTP processing power, does not reflect key computing

requirements (e.g., inter-computer communications needed for C 4 1 applications). It is questionable whether many of these applications could be effectively covered by any form of export controls, but at least the problem should be explicitly examined and the U.S. government should not delude itself that this policy is retarding such applications in countries of national security concern.

A more comprehensive examination of this longer term conjecture should be made as soon as possible. Serious thought needs to be given to the national security consequences, if it should prove to be true. Such thought should start with, but not be limited to, the continued long-term viability of the basic premises of export controls.

Key Findings

* The basic premises underlying the export control regime continue to be viable, at least in the short term, although less strongly than was the case during the Cold War.

* The premises and changing technical, geopolitical, and applications environments can be incorporated into an analytical framework that can be used to determine whether or not a viable
"supercomputer" threshold value exists and, if so, to derive the lower and upper bounds of a range of viable threshold values.

Applying the framework produces a current lower bound of controllability of-41000-5,000 mtops, which is likely to rise to 7,500 Mtops by late 1996 or 1997. The upper bound is determined by the performance requirements of applications of

national security concern. There is a cluster of such applications starting at approximately 7,000 Mtops, and another starting at approximately '0,000 Mtops.

While the basic premises will hold over the near term, preliminary analysis suggests that the efficacy of the current control regime will weaken significantly over the longer term. The principal reasons for this decreasing effectiveness are the rapid rate of technological development and diffusion, the changing nature of HPC usage in the U.S. national security domain, and the increasing difficulty of using the Composite Theoretical Performance (CTP) as a basis for distinguishing those systems that can and should be controlled, from those that cam-lot or should not.

Recommendations

Short-term

Perform annual reviews of the export control regime, applying a methodology that is open, repeatable, and based on reliable data.

* Use the analytic framework developed in this study to determine upper and lower bounds, based, respectively, on militarily important applications and uncontrollability, for a new threshold definition for export control on HPC. Depending on the level of caution deemed most prudent with respect to the applications, choose the new threshold to be near one or the other of these bounds.

* Use this framework to update and review the bounds no less frequently than every twelve months.

Longer term Recommendations

*Significantly enhance the analysis of applications of national security interest. Look closely and comprehensively at the national security applications of HPC that matter most to the United States, and which are most reasonably within the capabilities of countries of national security concern. In particular

*Identify applications of national security interest, assess the importance of preventing their proliferation to countries of national security concern, and determine the likely impact of failing to do so. It remains unclear whether the kinds of applications that have been examined for this study are as compelling a justification for export controls as were nuclear, cryptologic, and anti-submarine warfare (ASW) applications during the Cold War.

*Seek to distinguish those applications that cannot be performed in a satisfactory, cost-effective fashion on uncontrollable technology from those that can, at present and in the future. Maintain a list of the former, deleting applications when it can be demonstrated that they can be performed cost-effectively on uncontrollable technology, or adding applications as new uses of controllable HPC technology arise. The applications on such as list should provide the basic rationale for the existence of the export control regime.

*Significantly improve the quality of data related to applications of national security interest. Hard data on the relationship between the applications, computational methods, algorithms, and computer architectures and configurations is inadequate and often-nonexistent. Data about HPC usage and requirements in the national security community should be gathered more rigorously. Better data about the actual distribution of HPC technology in the United States and throughout the world would improve the quality of analysis.

*Conduct further study of the trends in HPC usage in the national security community and the implications for the export control regime.

*Conduct a study of the implications of networked computing systems on the export control regime. These systems do not lend themselves to easy classification using a single metric like the CTP, are not easily controlled, and will continue to be a problematic element in export control policy formulation.

*Cultivate comparative advantage through means other than control of hardware exports. Although export control policy has emphasized denying potential adversaries certain computational capabilities, national interests are also served when potential adversaries are forced to acquire technology and know-how at greater cost, effort, delay, and uncertainty than their American counterparts. Close working relationships between U.S. practitioners and systems developers should he encouraged to ensure that the former have access to advanced technologies well before their foreign counterparts. Additionally, the extensive experience of U.S. practitioners is a strategic asset, not easily duplicated abroad, which should be cultivated and preserved.

CHAPTER 1. ORIGINS AND PURPOSE OF HIGH-PERFORMANCE COMPUTER EXPORT CONTROLS

Export controls on dual-use technologies (technologies with both military and civil applications) have been part of U.S. national security policy since the 1940s. Restricting the export of high-performance computing (HPC) I technology has been one of the most consistent elements of U.S. national security policy since World War 11. The need to reduce the ability of potential adversaries to carry out applications of national security concern by controlling their access to computer technology has been embraced by each Congress and Presidential Administration since the dawn of the nuclear age over a half century ago. In 1949, the United States and its Allies established the Coordinating Committee for Multilateral Export Controls (CoCom) to coordinate the efforts of member countries in preventing Western goods, services, or technology from contributing to the military potential of Eastern Bloc countries. 2 The application of this policy to high-performance computing has been effective beyond reasonable expectations. During the Cold War, there were few examples of the successful Covert acquisition and use of Western HPC by countries of national security concern.

Since this time, computer technology has developed at a rate without precedent in history. Adapting export control policy to accommodate the rapid changes in the design, development, distribution, and use of computer technology is necessarily an ongoing and difficult process. 3

In early 1995, as a result of interagency discussions regarding the current U.S.-Japan bilateral supercomputer export control arrangement, the Interagency Working Group on Nonproliferation and Export Controls, chaired by the National Security Council, directed that a study be conducted to evaluate the effectiveness of the bilateral arrangement and make recommendations for changing it, if necessary, in time for the bilateral talks on supercomputers later that year. This study is a contribution to that effort.

The Recent Evolution of HPC Export Control policy 4

Through the mid-1980s, the export of nearly all computer systems to countries with communist governments was controlled. While the stringency of the policies increased or decreased with the prevailing political climate, 5 exports required explicit approval by the Department of Commerce through the authority granted it under various versions of the U.S. Export Administration Act (EAA) (e.g., 1969, 1977, 1979, etc.).

The rapid growth of the personal computer industry 6 during the 1980s forced policymakers to grapple with limitations in the ability to enforce export control policy. Consumer demand fueled the production of millions of systems by manufacturers located throughout the world. Their small size, large installed base, innumerable distribution channels, and international production made it impossible as a practical matter to control their diffusion to countries of concern.

The success of the personal computer industry forced the government to distinguish between at least two types of computers-low-end systems like PCs and

workstations that were geared towards mass markets, and high-end systems, characterized by HPC computers serving a much smaller market with more specialized and demanding needs.

The distinction was important because the globalization of computer manufacturing, (one of several important trends affecting the utility of export controls) meant that it would become increasingly difficult to regulate low-end computers effectively, regardless of their value to foes abroad. Recognizing this reality, Commerce decontrolled the first wave of PCs in January 1985, making such computers as the IBM PC-XT freely exportable. Further liberalization did not occur until August 23, 1988, after Congress passed the Omnibus Trade and Competitiveness Act of 1988, which both -amended and re-authorized the EAA of 1979.

During the early 1980s, potent competition to U.S. high-end manufacturers emerged in Japan. Japanese HPC developers forced the U.S. government to realize that the United States would not indefinitely remain the world's sole source of this technology. U.S. officials initiated negotiations that led, in 1984, to a U.S.-Japan bilateral arrangement to regulate jointly the export of high-performance computers. The accord is known as the Supercomputer Control Regime.

For the first seven years of the arrangement, the two governments coordinated the export of a specific list of the ten or so highest performing computers. The U.S. did not formalize the accord by publishing written regulations. Instead, Commerce worked informally with the few HPC manufacturers in imposing a supercomputer definition and the security safeguard requirements for a particular sale. 7 Newly developed computers were subjected to HPC controls whenever they exceeded 100 Mflops (millions of floating point operations per second). 8 Although the accord did help to ensure that Japanese and U.S. HPC vendors were operating under comparable export control guidelines, it was not without industry criticism, especially as new licensing burdens were imposed on HPC exporters. Moreover, because the tern-is of the accord were not published, manufacturers come to feel that government licensing decisions were arbitrary and that Japanese producers were not held sufficiently to
similar standards.9 Movement toward greater transparency began in December 1988 when Commerce first published, as
required by the Omnibus Act of 1988, a proposed supercomputer definition. 10 The definition established a threshold performance level above which a computer would be considered a supercomputer for export control purposes. The initial definition was set at 160 Mflops 11 and was intended to be subject to periodic review. The definition also contained specific technical guidelines and a standard formula for12 measuring a computer's performance. Public comments were solicited.

In January 1990, Commerce published a revised proposal defining the term supercomputer. 13 To allay industry concerns that the definition was set too low, Commerce reinforced its intention to review the regulation, now on an annual basis. The revised rule proposed to define a supercomputer by three distinct

thresholds, set at 100, 150, and 300 Mflops respectively. The different thresholds would correspond to the level of required security safeguards, based mainly on the country of destination. The publication of a final supercomputer definition and safeguards requirements, however, awaited the outcome of other events.

By late 1990, the bilateral accord with Japan was in great need of modification. The U.S. government faced growing pressure from manufacturers to modify the agreement, especially as East-West relations continued to improve and several European firms appeared ready to enter the industry while not bound in any way to the accord. Moreover, rapid technological change was simply undermining the agreement's established framework.

Between March and June, 1991, U.S. and Japanese officials renegotiated the seven-year old accord. Both countries adopted a supercomputer definition requiring security safeguard arrangements at 195 Mtops 14 (millions of theoretical operations per second), ending their previous practice of identifying machines by specific name. Each country was allowed 30 days to review the other's license applications. Strict safeguards were required on HPC exports to states of national security or proliferation concern, such as those having failed to sign the Nuclear Non-Proliferation Treaty. Licensing and safeguard requirements for less risky sales to Western allies were eased. 15

In his announcement concerning the agreement with Japan, Press Secretary Fitzwater stated that "[b]oth Governments share the view that supercomputers are of strategic concern, particularly for the development of nuclear weapons ,and missiles, and that great importance should be attached to export controls on supercomputers for the purpose of preventing the proliferation of such weapons." 16 Months later, in formal remarks to Congress about the U.S.-Japan agreement, President Bush further noted that high-performance computers are controlled for their relevance to cryptology, strategic defense, and anti-submarine warfare activities. 17

Originally, exporters appeared pleased with the U.S.-Japan accord. 18 Some systems, such as the Cray XMS, were freed from controls, some threshold levels were set higher than first proposed, and the government committed itself to a more open regulatory process. Yet computer manufacturers often expressed frustration with the failure of government decision-making to keep pace with the rapid technological advances in their industry. Workstation producers were beginning to find themselves classified as supercomputer companies subject to the more rigorous and costly HPC-level security safeguard requirements. 19

In September 1993, the Clinton Administration published a report of the Trade Promotion Coordinating Committee (TPCC), a major interdepartmental study backed by both the National Security and the National Economic Councils. The report, which outlined a series of actions designed to promote U.S. exports, proposed that the U.S. begin negotiations with Japan to raise the supercomputer threshold

limit from 195 Mtops to 2000 Mtops while reviewing and updating safeguard requirements. 20

To implement this proposal, the U.S. had to renegotiate the terms of the U.S.-Japan HPC agreement. In October 1993, U.S. negotiators met with their Japanese counterparts to discuss raising the accord's threshold level. (At the same time the U.S. moved both unilaterally and within CoCom to liberalize

low-end computer controls up to the supercomputer boundary of 195 Mtops. 21) By December 1993, U.S. negotiators had successfully concluded their discussions with the Japanese, 22
although the Administration did not reach its avowed goal of raising the supercomputer level from 195 to 2000 Mtops. Instead, the new (and current) level was set at 1500 Mtops. Commerce amended the Export Administration Regulations accordingly in February 1994. 23
In February 1995, twenty months after the TPCC's first assessment, the Clinton Administration began a

further review of both low- and high-end computer controls. 24 The present study is one result of that effort.

Study Objectives and Structure

A brief review of export control history underscores a number of points:

* There has been considerable consistency in the ultimate purpose of the policy: inhibiting potential adversaries from carrying out nuclear, chemical, and conventional weapons development, cryptology, and other applications of national security interest.

*The policy has continually needed to accommodate rapid technological change.

*Changes to the policy have come about as a result of considerable discussion, even confrontation, among interest groups (e.g., industry, various government agencies, other countries), at considerable cost in time, effort, and expense.

*There has been a trend toward making policy formulation more open and objective, but much discussion has taken place without the quality of data or the clear analytical framework needed for transparent and defensible decision-making.

During the 1990s, the policy has had to take into consideration a number of other complexities. Not only is computer technology advancing faster than ever, but the world's geopolitical structure is changing dramatically, and with it, the nature of the national security threats. Following the break-up of the Soviet Union, the threats from regional conflicts and "terrorist countries" have grown relative to those from a global superpower. Threats have become more numerous, but individually smaller in scale. These changes have important consequences for the computer-based applications that constitute threats to U.S. national security, or are employed by the U.S. national security community to protect U.S. interests.

As part of an effort to establish an updated export control threshold, this study evaluates the importance of high-performance computing systems in applications of national security concern, and the

availability of comparable technologies in some countries subject to control restrictions. It presents a broader framework for analysis, based on three basic premises that provide the foundation for the existence of export control policy. It is the authors' hope that the analytical framework employed here will provide a basis for deriving future control thresholds that is transparent, objective, defensible, and repeatable. Such an approach is increasingly needed in today's highly dynamic technical and geo-political environments.

The basic premises and the analytical framework are presented in Chapter 2. Chapters 3, 4, and 5 apply the framework to derive a viable control threshold that is consistent with the basic premises. Chapter 3 discusses trends in domestic and foreign HPC development and distribution to establish a lower bound for a viable control threshold. Chapter 4 discusses the role of HPC in national security applications in order to evaluate the need for, and efficacy of, export controls. It discusses four broad categories of applications: nuclear weapons development, cryptology, conventional weapons programs, and military operations. The first two categories in particular have traditionally been offered as the primary justification for the existence of the export control regime. Chapter 5 integrates the results of the earlier chapters into specific recommendations for control thresholds. Finally, Chapter 6 examines export controls in light of key trends in technology development and use, identifies some weaknesses in the current control regime, and makes recommendations for future action.

Methodology

This study relies heavily on data about the computing requirements of applications of national security concern, HPC technology trends, and HPC industry trends. These data were gathered with the help of many individuals from four principal sources:

* On-site interviews with applications practitioners. The most detailed and insightful data about the computing requirements and trends in key application areas were gathered through personal interviews with practitioners. Applications targeted for such examination were selected because they were (a) historically important in HPC export control policy making, (b) exhibited unique or otherwise significant computational requirements, or (c) were representative of broad categories of applications.

* Application requirements databases. The U.S. Department of Defense (DoD) High-Performance Computer Modernization Office (HPCMO) has compiled data about the current and future HPC needs of approximately 700 DoD HPC applications. The data are not as precise or consistent as one would like, but are useful in their scope of coverage.

*Industry representatives. Considerable data about current systems, technology trends, and system usage were obtained through communications with industry representatives, who provided corporate data as well as their insights into industry trends.

*Open literature and public domain sources. The open literature (including electronic sources) contains a wealth of information about developments in HPC not only in the United States but throughout the world. It varies considerably in quantity and quality, but collectively, and in conjunction with other data sources, provides a considerable amount of useful information. The authors have been gathering such information, augmented with field research abroad, for several years.

The study has been subject to ongoing verification. The authors have repeatedly presented the analytical framework to a wide range of applications practitioners, industry representatives, and government officials. Their feedback has been incorporated in subsequent refinements of the framework. Some data, such as those found in the HPCMO databases and Top5OO Supercomputer Sites listings, could not be verified exhaustively, but key data elements have been checked against original sources and, when necessary, corrected. However, time constraints were such that follow-on research and re-visits with key personnel were not possible. Only about three months (late April through late July) were available for conducting this study, of which research and visits were only one (albeit major) component. Thus, inevitably, some errors remain. One of the major purposes of this study, however, has been to establish a framework for an ongoing, public dialogue on the basis for export control policy that

encourages interested parties from government and industry to come forth with concrete data regarding systems, markets, applications, and the relationships between them. We hope that whatever errors remain will over time surface and be corrected.

Chapter I Notes

High-performance computing system is a term usually loosely applied to the technology, or collection of technologies that make it possible to address the most computationally demanding problems. In the past, supercomputer was used to describe the most powerful systems available. These systems were characterized by the highest computational performance at a given point in time, small production runs, and high cost (tens of millions of dollars). During the late 1980s and 1990s the term high-performance computing has become more popular. This term recognizes that a computing system today often depends on not only a powerful computational engine, but also high-speed networks, advanced storage systems, sophisticated graphics, collections of less than the most powerful computers, etc. Most attention is usually paid to the individual computers that are at the heart of such a system. Throughout this study, HPC will refer to these as computational engines, unless otherwise noted.

2 For a review of the entire export control regime, see R.W. Schmitt, et al., Finding Common Ground: U.S. Export Controls in a Changed Global Environment, NAS/NRC (Washington, D.C.: National Academy Press, 1991).

3 John R. Harvey, Cameron Binkley, Adam Block, and Rick Burke, A Common-Sense Approach to High-Technology Export Controls (Stanford, CA: Center for International Security and Arms Control, 1995). See also, National Academy of Sciences, Committee to Study International Developments in Computer Science and Technology, Global Trends in Computer Technology and Their Impact on Export Control (Washington, D.C.: National Academy Press, 1988), which was the most extensive single study ever done on export controls and computing.

4 Thanks to Cameron Binkley, of the Center for International Security and Arms Control, who drafted this section.

5 For example, during the Ford administration, Control Data Corporation was permitted to sell a number of Cyber 17x systems to the Soviet oil and gas industries, and the export of a number of other systems was approved for use in Soviet air traffic control systems and the Kama River Truck Plant. In contrast, during the late 1970s and early 1980s, in response to Soviet human rights policies, the invasion of Afganistan, etc., the Carter and Reagan administrations took steps to tighten restrictions, cancelling a number of high-end systems sales. See T.J. Richards, "An Examination of the Issues Affecting ADP Technology Transfer to the Soviet Union," Ph.D. Thesis, George Washington University, Washington, D.C., Nov. 1980.

6 Apple Computer, Inc. introduced the first commercially viable personal computer in 1977, followed four years later by IBM's PC.

7 Supercomputer security safeguards are any of various restrictions, such as 24-hour surveillance, reviewing the records of computer activity via special software audit programs, or limiting personnel access, designed to prevent or uncover recipient uses of an HPC unauthorized by the terms of the exporter's license. "US in Early Stages of Supercomputer Control Talks," Export Control News (March 26, 1991). Mflops are used to rate the speed of computers that can perform complex scientific calculations based upon floating-point arithmetic.

9 See, for instance, Michael M. Phillips, "Supercomputer Talks Try to Balance Commerce, Security," States News Service (March 15, 1991).

10 53 FR 48932 (December 5, 1988); Commerce intended to include the supercomputer definition required by Section 5(a)(6) of the EAA by revising § 776.10 of the Export Administration Regulations (EAR).
11 The theoretical peak performance of the Cray-1.
12 See, 53 FR 48932 (December 5, 1988) for specific technical details of the proposed regulation.
13 55 FR 3017 (January 29, 1990).
14 In June 1990, as a result of a major reform effort, CoCom adopted a new standard for evaluating computer performance. The new standard is called CTP (for Composite Theoretical Performance) and is measured in Mtops. Mtops are roughly equivalent to Mflops, but take into account non- floating-point computation, account for variations in word-length between systems, and are able to rate the performance of low- as well as high-end computers. The CTP formula and supercomputer definition were published in the Federal Register on February 6, 1992 (57 FR 4553).
15 There are five tiers of security safeguard levels, determined by country. Between supplier states (currently defined as the U.S. and Japan) no controls are applied, minimal requirements are imposed on major U.S. allies (e.g., Britain, France), a somewhat larger group of states requires a safeguards plan (e.g., South Korea, Sweden), while still others must further have certification by the government of the importing country. Finally, licenses for restricted countries require all safeguard levels, but will generally be denied (e.g., Iran). See, 57 FR 20963 (May 18, 1992).
16 Statement by Press Secretary Fitzwater on Supercomputer Export Controls," in Weekly
Compilation of Presidential Documents (June 7, 1991), p. 736.
17 George Bush, "Message to the Congress Reporting on the National Emergency With Respect to Export Controls" (March 31, 1992), in Weekly Compilation of Presidential Documents (April 6, 1992), pp. 561-563.
18 Eduardo Lachica, "U.S., Japan to Ease Licensing Burden on Export of High-Powered Computers," Wall Street journal (June 10, 1991), p. B3.
19 Indeed, Bill Clinton faced the issue on his first post-election visit to California's Silicon Valley. During a nationally covered visit to Silicon Graphics, Inc., where the special graphics effects for the motion picture Jurassic Park were created, an employee challenged the President to lift supercomputer control levels now that her company was building workstations that exceeded them. Tom Abate, "The Politics of Supercomputers," San Francisco Examiner (February 28, 1993), pp. E1-E7.
20 Trade Promotion Coordinating Committee (TPCC), "Toward a National Export Strategy: U.S. Exports = U.S. Jobs," (September 30,1993), p. 57.
21 See, Dan Cook, "TPCC Announces Major Decontrols," The OEL Insider (U.S. Department of Commerce, December 1991), pp. 1, 4, for a listing of the various low-end computer decontrols.
22 Conclusion of Negotiations to Define New Supercomputer Threshold," Fast Facts (U.S. Department of Commerce fax information service, January 14, 1994), p. 2.
23 59 FR 8848 (February 24, 1994).
24 Administration to Assess Computer Controls," Export Control News (February 28, 1995).

CHAPTER 2. BUILDING ON BASIC PREMISES

The 'Basic Premises" Behind Export Control Thresholds

HPC export controls have been implemented by determining a definition of supercomputer based on some measure of processing power. Extraordinary licensing and safeguard conditions are placed on the sale or transfer of any machine at or above this processing power threshold.

The policy has been successful in part because it has been based on three premises that were largely true for the duration of the Cold War:

1. There are problems of great national security importance that require high-performance computing for their solution, and these problems cannot be solved, or can only be solved in severely degraded forms, without such computing assets. Secondary assumptions are that the most important parameter of computers for these problems is computational performance, and that computational performance can be adequately measured by a single metric, such as the Composite Theoretical Performance (CTP), which is applied to individual computer systems.

2. There are countries of national security concern that have both the scientific and military wherewithal to pursue these or similar applications.

3. There are features of high-performance computers that permit effective forms of control. Implicit in this premise is that countries of national security concern do not already have sources of HPC technologies that are outside the control of the participants in the export control regime (e.g., CoCom prior to 1994).

If the first two premises do not hold, there is no justification for the policy. Preventing the acquisition of this technology is not imperative if all problems of national security importance can be solved with computers that are not considered high-performance systems, or if countries of national security concern are unable to use computational results effectively. 25

If the third premise does not hold, an effective export control policy cannot be implemented, regardless of its desirability. If the technology diffuses so that it becomes possible for recipients of national security concern to circumvent the export restrictions and acquire the technology through covert or alternate channels at modest cost and effort, the policy becomes ineffective. It no longer accomplishes its objectives, and one may question its continuance.

A strong case can be made that all three premises held during the Cold War, and that export controls on computing were an important and effective element of U.S. national security policy. Since the end of the Cold War, however, there have been significant changes in the context for this policy:

1. The nature and extent of foreign threats to security have changed. During the Cold War, the principal threats were from a small number of countries with advanced scientific, technological, and military capabilities, most specifically the

Soviet Union, the German Democratic Republic (East Germany), and the People's Republic of China. Regional conflicts took place within the overarching bi- or tri-polar tensions. With the dissolution of the Soviet Union as a global superpower, the threats to national security have become individually smaller in scale, but more numerous and varied. On the one hand, adversaries are less likely to have highly developed domestic R&D programs (vis-A-vis the United States); on the other, their arsenals consist of substantial amounts of foreign technology, purchased on the international arms markets. Under these circumstances, involvement of U.S. forces in possibly numerous, rapidly developing, and distributed conflicts has increased the role of command, control, communications, computing, and intelligence (C 4I) systems, and advanced conventional weapons systems. All of these have very demanding and specialized computing requirements.

2. There have been dramatic changes in computing technology. Microprocessor, workstation, networking, and storage technologies have developed rapidly over the last half-dozen years. At

the same time, the relative importance of the federal government as a supporter of the high-performance computing industry has declined since the end of the Cold War. Developments in these technologies are today driven primarily by the needs and requirements of commercial markets.

3. The uses of high-performance computing have expanded significantly in scope and nature within the U.S. national security community. In particular, applications related to the development and operation of advanced, high-performance conventional military systems are growing rapidly.

The purpose of this study is to examine the continued viability of the three basic premises under the conditions of the changing environment, and to use this analysis to recommend a threshold definition of supercomputer. If the premises are valid, it should be possible to derive a control threshold from an analysis of the premises in a way that is explicit, justifiable, and repeatable. A methodology based on this approach could provide the foundation for periodic reevaluations of the threshold in a consistent fashion. If the premises are not valid, then the analysis should clearly illustrate why no effective control policy based on the premises is possible.

Selecting an Export Control Threshold: A Dynamic Framework

Elements of an Analytical Framework

The basic premises described in the preceding section incorporate four concepts: applications (i.e., the application of computers to problems of national security importance), computer systems, a performance metric, and the controllability of computer systems. The premises explicitly require that there be applications of national security importance that require high-performance computing systems for their solution. Premise three touches on the issue of controllability, or the degree to which the sale, transfer, and use of computer hardware can be monitored and regulated by government agencies and industry vendors. The current export

control regime relies on a performance metric to evaluate the performance of all kinds of computer hardware. Measured in millions of theoretical operations per second (Mtops), the Composite Theoretical Performance (CTP) is a function of a system's execution rate for individual computational elements of various types, word length, and, to a lesser extent, internal bandwidth. The CTP depends only on a computer's hardware and is only loosely affected by differences in computer architecture. It was designed to be simple, software- and application-independent, and
applicable to a wide range of hardware architectures. 26

The current export control regime assumes that a metric like CTP can be used to capture the aspects of a computer system's hardware that most directly determine its usefulness to applications of national security importance, that is, its computational performance. If the premise that there are problems of national security importance that require the use of high-performance computing systems holds, then, in principle, there should be a way to relate computer performance to an application's computational requirements. The current export control regime leaves little choice but to use Mtops to describe the computational requirements of applications as well.

Change over time is a hallmark of today's environment that is not adequately captured by the basic premises. They are static in nature, and their validity is tested for individual points in time (e.g., "today"). It is crucial, however, to understand the dynamics of change in the applications, computer systems, performance, and controllability. Our analytical framework tries to capture time as a fundamental element.

Before presenting the analytical framework in its entirety, we discuss applications and technological trends in more detail.

The Applications Stalactites

The first premise postulates that there exist applications with high minimum computational resource requirements. If computing power above some minimum threshold is not available, it is effectively impossible to perform the application in a useful fashion. Often the minimum threshold is determined by timing considerations, by a minimum time-to-solution. For example, the minimum computing power needed to crack Digital Encryption Standard (DES) encryption codes might be determined by a need to be able to break keys within 24 hours. If less computing power is available, a solution may be obtained in more than 24 hours, but the result has less or no operational value. Timing considerations vary greatly among application groups. Real-time applications such as missile and submarine detection require solutions in minutes, seconds, or fractions of seconds. In aircraft design, systems that provide computational solutions overnight make much more efficient use of engineers' time than those that require longer times for processing. While the output of a single run of an application might be the same regardless of the time required, rapid turn-around permits engineers to maintain their concentration on a single problem and iterate more frequently, and leads to qualitatively more effective design solutions and shortened development times.

To establish the minimum performance requirements in this study, people who work directly with the applications were asked to identify the minimum computer configuration that they would need to carry out the application. The minimum bound for this application was quantified in Mtops by looking up the CTP of the system used to achieve this level of performance.

Over time, the minimum requirements for a given application (or more precisely, the minimum requirements for an instance of that application of a given size) tend to drift downward. As algorithms,
models, and systems software improve, the number of computer cycles and amount of memory needed to 27
achieve the same results declines. But for a given problem and problem size, they do not increase.

At any point in time, the maximum computing resources that can be applied to a particular application is the performance of the most powerful system available. At least three definitions of this maximum apply. The purely technical maximum is the most powerful system that can be constructed if are possible if money is not a limiting factor. However, budgets are constrained and "the most powerful system" may mean the most powerful system that can be acquired for a fixed amount of money. A third possible definition is that of the most powerful computing system installed anywhere at any given point in time. Systems more powerful than this exist only in a theoretical sense and, obviously, perform no useful work. During the height of the Cold War when the vector-pipelined architectures (e.g., Cray) were the most powerful systems available, these three definitions frequently coincided. This is no longer necessarily true. There are numerous massively parallel systems (e.g., the Cray T3D) that have not been installed in the theoretically maximum configuration because customers have not been willing to pay for such a system.

Between the lower bound and the maximum computing resources available lies the system actually used for the application. Figure I illustrates these concepts for the F-22 aircraft design application. The first time the application is successfully performed, the actual system may coincide with the lower bound or the maximum (usually the latter). Over time, as the maximum computing power available rises, the performance of the actual system used for the application also rises.

Figure 1. [Omitted] Range of Computational Power for the F-22 Design

From an export control perspective, the most important bound is the minimum computing performance necessary to carry out the application effectively. If all applications of interest have, like the F-22 design, a minimum bound that is low enough that it can be done on other than supercomputers, then restricting the export of high-end systems does not deny potential adversaries the computational capability to solve this problem. The first basic premise requires that there be

applications with high minimum computational requirements. Chapter 4 explores the validity of this premise.

The Technology Curves

The third premise postulates that there are characteristics of high-end systems that make it possible to control the export of these systems such that they are not delivered to countries where they would be used for applications that are of national security concern to the United States. There are clearly computers that are not controllable. Millions of personal computers are manufactured annually, and sold without knowledge of their end use around the world. There are dozens of independent vendors of this technology. Anyone anywhere in the world can easily acquire it.

If some systems are clearly controllable (e.g., traditional supercomputers) and some systems are clearly uncontrollable (e.g., personal computers), then between these extremes lies the most powerful uncontrollable computer system. There may not be broad consensus on what this system is at any given time, for controllability is a partly subjective quality. It is also possible to specify several "maximum uncontrollable systems": e.g., from each vendor, or in each architectural class, and so on. Chapter 3 discusses the issue of controllability at greater length. What is clear is that the computational performance of the most powerful uncontrollable system(s) rises over time. As it rises, it overtakes the minimum computing requirements of individual applications. When this happens, it is no longer possible for an export control regime to prevent potential adversaries from acquiring the computing hardware necessary to perform that application.

The computing power available to potential adversaries is a function not only of Western uncontrollable systems, but also of the availability of computing systems from domestic or other non-Western sources. A final important trend is the performance of the most powerful systems, domestically produced or imported, in use in countries of national security concern. By definition, these systems are beyond the control of Western export control regimes, and should be considered in determining which countries have the computing power necessary to perform specific applications. In defining this trend, we do not include one-of-a-kind installations that are under a Supercomputer Safeguard Plan.

The chart shown in Figure 2 illustrates the dynamic over time of the three technology trends mentioned above. For the sake of clarity, the figure shows only a single "stalactite," representing the computational requirements of a single application. It is located on the X-axis according to the year in which the application was first successfully performed. In reality, there would be a stalactite for each application (and, potentially, for each qualitatively different problem size within a type of application) of national security concern.

Selecting a Control Threshold

The HPC applications requirements and technology trends discussed above provide an analytical framework that can be used to guide the selection of an export

control threshold at any given point in time. There are two basic questions to address. First, what is the range of valid thresholds that do not violate any of the three basic premises? Second, on what basis might one valid threshold be preferred over another?

The lower bound for a valid threshold should be defined by the greater of the lower technology curves, that is, the performance of the most powerful uncontrollable systems and the most powerful systems available in countries of national security concern. If the threshold is set below the level of controllability, then export control policy will try to control the uncontrollable and will be unsuccessful. The inability to achieve its goals plus the time, effort, and expense expended by government and vendors trying to comply with the law will reduce the credibility of the policy and further erode its effectiveness.

The theoretical maximum of the threshold is the performance of the most powerful systems available. Between the maximum and the lower bound, however, is a range of performance levels at which computer systems are controllable, and satisfy the third basic premise. The question remains, on what basis should the selection of one of these performance levels be made?

Figure 2. [Omitted] HPC Applications and Technology Trends

There are three perspectives from which a selection can be made. From the first perspective, the control threshold is set as close to the lower bound as possible. This perspective takes the position that that which can be controlled should be controlled, regardless of the computing requirements of actual applications. An alternative approach is to set the threshold in light of the performance requirements of actual applications. This perspective says to examine the set of all applications whose minimum performance requirements lie above the uncontrollability threshold, and set the threshold just below the minimum of all the minimum requirements. This threshold might lie significantly above the controllability threshold, but from a national security perspective does not matter; all the applications that can be controlled, are controlled. If the minimum of all the minimum requirements lies close to the uncontrollability threshold, these two approaches suggest the same threshold.

A third perspective takes into account both national security and economic factors. It is possible to estimate the size of world-wide markets for machines at various performance thresholds. If the market for machines at some level just above the uncontrollability threshold is large, an argument might be made that the economic gain to U.S. industry from setting a threshold above this level outweighs the cost to national security of effectively decontrolling the technology needed for a few applications just below this level.

Although the purpose of this report is not to weigh economic and national security considerations against each other, a discussion of an ideal scenario might clarify the approach that takes them both into account. If we were to collect data on the

minimum computational requirements for a representative sample of applications of national security interest, past and present, we could plot how man applications have requirements that fall within a certain range of performance, measured in millions of theoretical applications per second (Mtops). Similarly, at some instant in time, one could plot the number of installed computers at each CTP rating. Figure 3 shows a purely hypothetical distribution of the minimum computational requirements of applications, mapped against a graph of the distribution of computer installations at various CTP levels.

Figure 3. [Omitted] Hypothetical Distribution of Applications and Computer Installations

Figure 3 is a snapshot that reflects the distribution at one point in time. The CTP levels of the uncontrollability trend-line and the CTP of the most powerful computer systems available can therefore be drawn as lines A and D. Any threshold set between lines A and D should be enforceable, since systems at these CTP levels are controllable. Lines B and C illustrate how different thresholds potentially can have significantly different national security and economic effects.

If a threshold were set at B, the relatively few applications falling between A and B would be given up. This loss might be outweighed, however, by the economic gains from additional sales of computer systems falling between A and B, for which a robust market is known to exist. If the threshold is to be set anywhere above A, B is a reasonable choice.

In contrast, C is not a reasonable choice. The economic benefit gained from decontrolling systems between
B and C is likely to be minimal since the market, reflected by current installed base, is not large. 28 On the other hand, the cost to national security could be very significant, since there are large numbers of applications with minimum computing requirements between B and C. Under these circumstances, B could be considered a reasonable upper bound for a CTP threshold that satisfies the three basic premises and, at the same time, balances the national security and economic interests of the country.

In summary, if anything other than a "control what can be controlled" approach is used to set the threshold, an analysis similar to the preceding one should be employed. Ideally, a control threshold should be set below a "hump" in the applications distribution and above a hump in the computer installations distribution. There is no guarantee that a threshold satisfying both of these conditions exists, or that there will at all be an obvious choice. It should be clear, however, that thresholds just above a hump in the applications distribution should be avoided.

This snapshot approach can be used to select a threshold at a single point in time. Over time, the following changes occur to the graph in Figure 3:

*The shape of the curve reflecting numbers of existing applications remains relatively stable, although there is likely to be some drift toward the lower end of the X-axis, reflecting improvements in software and algorithms.

*New applications will emerge, potentially creating new humps, particularly at the high end of the X-axis. The applications will arise as the available computing power (a) makes it feasible to address old applications with larger problem sizes that result in qualitative changes in the results that can be obtained, and (b) makes fundamentally new applications possible.

*The distribution of installed systems will change dramatically. Humps reflecting large installed bases will move toward the right of the X-axis. The most powerful systems commercially available (line D) will shift right, as will line A, the performance of the most powerful uncontrollable systems.

The rate of change of these trends may be difficult to project over the longer term, but at any given point in time, it should be possible to establish a snapshot, and project trends in year or eighteen months into the future. Figure 11 in Chapter 5 provides such a snapshot, mapping the current (6/95) distribution of high-performance computing systems against the performance requirements of key applications of national security importance. This analytical framework is best used in the context of regular reviews. At each iteration, the applications and systems curves should be reevaluated, and an export control threshold selected that will be in place until the next iteration.

Basic Premises and Future Scenarios

As long as the three basic premises hold, a case can be made for establishing an export control threshold. All three held strongly during the Cold War, but it is not clear that they will indefinitely. There are a number of scenarios under which one or more of the basic premises could fail, and result in the collapse of the export control regime as it currently exists. These scenarios can be modeled using the analytical framework.

I Premise one, that there are applications of national security concern that can only be performed using high-performance systems, might fail if there are no applications stalactites with minimum computational requirements above the level of the most powerful uncontrollable system. This could happen if the capability of the most powerful uncontrollable computing system exceeds the minimum computational requirements of all applications of national security concern. Since the capability of the most powerful uncontrollable systems rises steadily over time to overtake the minimum computational requirements of current applications, such a scenario might take place if new applications with very high minimum computational requirements do not emerge. It is important to note that the reasons behind this scenario might be largely economic, rather than technical or application-related. Applications practitioners operate under budget constraints and must make decisions about the allocation of funds. The wisest allocation might not be the purchase of a single, "most powerful" system. The applications

actually performed may be those which are possible with a collection of more cost-effective systems. Such systems often are among, or comparable to, uncontrollable systems.

2. Premise two, that there are countries of concern with the military and technological wherewithal to pursue applications of national security importance, could fail if the global geopolitical landscape changes such that countries with the necessary knowledge and technological expertise are no longer considered to be of national security concern. This consideration is beyond the scope of this study.

3. Premise three, that there are characteristics of high-performance computing systems that make export control possible and feasible, could fail -if the gap narrows between the most powerful systems available and the most powerful uncontrollable systems. The viability of the policy depends on the existence of a sufficiently large difference between lines A and D)in Figure 3 that a threshold can be drawn between them with some confidence. If A and D lie close together, there is no meaningful range of controllability. Such a scenario could come about if there is a shift in the computer industry from the construction of powerful individual systems based on proprietary technologies to the construction of basically uncontrollable building blocks that can be combined in powerful configurations.

25 For example, the ability to design an advanced fighter aircraft using a supercomputer is of little benefit if a country lacks the manufacturing technology to build a usable aircraft.

26 See A. James Ramsbotham and Greg C. Miller, Composite Theoretical Performance (CTP)- Development and Evolution of a Computer Performance Model for Export Control (Washington, D.C.: Institute for Defense Analyses, September 1994).

27 What does increase over time is the size of the problem an engineer works on. As more powerful computers become available, an engineer is likely to increase the size of the problem by reducing the time-step used in a simulation, increasing the resolution of the application domain grid, etc. But once a problem of a given size is successfully solved with a particular computer configuration, the results are repeatable- it can always be solved by that same, or a comparable, computer configuration.

28 The conclusion that the current installed base is a good indicator of the future market should not be accepted without question, however. Within a given (e.g., U.S.) market, demand at given performance levels can rise rapidly as price/performance ratios change. It is likely, however, that markets at these particular levels would grow more slowly in countries of national security concern than in the United States. Estimates of foreign markets for systems at the performance levels in question should complement the current analysis.

CHAPTER 3. ESTABLISHING A LOWER BOUND

Art appropriate control regime for high-performance computing systems must reflect not only what computing capability is possible, but also what is practical. The practicality of control is

a function of a. number of factors. Two of the most important are the inherent controllability of HPC technologies

available from traditional supplier countries (e.g., the United States, Japan, certain Western European countries) and the availability of domestic or imported HPC technology in countries subject to export control restrictions.

In an effort to establish a reasonable lower bound for export control thresholds, this chapter examines the relationships between indigenous HPC developments in selected restricted countries-Russia, the PRC, and India-and the trends in what might loosely be defined as "uncontrollable" HPC systems. These are systems having certain combinations of characteristics that make it difficult, or impossible, as a practical matter, to restrict their export only to known and approved destinations.

The next section discusses trends in HPC development in Russia, the PRC, and India. We then address ,,he factors driving HPC development in the United States and the implications for the controllability of systems. The third section examines the likely future relationship of HPC efforts in countries of control interest and "uncontrollable" systems developed in traditional supplier countries, and is followed by our conclusions.

Trends in Foreign Indigenous HPC Systems

Russia 29

Russian computer scientists have developed multiprocessor computing systems since the early 1960s. Faced with a perceived need to maintain the appearance of computing parity with the West, a weak domestic microelectronics industry, and the rich research opportunities afforded by parallel processing, Soviet scientists and engineers developed a variety of indigenous multiprocessors. They felt that they could compensate for the country's inability to manufacture individual processors that matched the world-wide state of the art by combining numerous processors with modest performance. The HPC sector was notable for the breadth of approaches taken, ranging from shared-memory coarse-grain systems to programmable architectures and dataflow systems, but not for its ability to develop usable machines that could meet the needs of the domestic computing community. The most powerful machine put into series production, the 10-processor Elbrus-2, was a 94 Mflops system introduced during the mid-1980s, created at the Institute for Precision Mechanics and Computer Technology (ITMVT). The institute was the principal developer of high-performance computing systems from the early 1950s until the end of the Soviet era. The head of the Elbrus program, Boris

Babayan, is currently doing research and development in Moscow under contract with Sun Microsystems.

The reform process in Russia and the resulting economic and political upheavals have devastated much of high-performance computing there. Nearly all the Soviet era programs and projects have ended, or are limping along with insufficient funding to build concrete systems. The most powerful, completely indigenous system to pass state testing (around 1990) was the Macro-Pipeline processor (MKP), developed at ITMVT. In a dual-processor configuration, this system had a peak performance of N..2 Gflops. Four units were manufactured during the early 1990s, but for lack of paying customers, production has ended. With the exception of ITMVT employees who are working with Babayan on Sun-related projects, ITMVT has largely atrophied and is no longer developing high-end systems.

During the early 1990s, a growing number of 'institutes became involved in developing or using systems based on Western microprocessors, principally transputers. Activity in this area increased as microprocessors like the transputers and 1860s became commercially available, and certain prohibitions against the use of foreign technologies for military-related applications were relaxed. In 1990, a Soviet (later Russian) Transputer Society was formed to cultivate domestic support for distributed memory multiprocessors and help disseminate information about foreign developments in this area to domestic scientists.

There are numerous examples of small (7-32 processor) T800 transputer configurations (including some imports from India and Bulgaria) in Russia, and several vendors of boards based on transputers, i860s, and other microprocessors. Few organizations appear to be putting together configurations of the more powerful microprocessors, however. The Kvant scientific research institute is perhaps the leader among those that are. As Table 1 shows, Kvant has constructed systems integrating transputers and i860 processors. The configurations currently available have 32 i860 processors and will reportedly soon be upgraded to 64 processors. The architecture is said to be scalable to 512 processors, however. Kvant has built systems using other processors as well, including the TMS32OC40 digital signal processing chips, and is relatively well prepared to incorporate other advanced microprocessors in the future.

Table 1. [Omitted] Russian High-Performance Computing Systems

In the short term, domestic high-performance computing will be severely hampered by the poor economic state of traditional HPC customers-the military, geophysics, and aerospace-and the market preference for imported workstations and high-end personal computers.

People's Republic of China 30

Over the last decade and a half, the Chinese government has initiated a number of national programs to cultivate a computer industry capable of supporting the

government's plans for modernization and economic development. During the 1990s, these policies have had a pragmatic, export-oriented nature. The government has concentrated resources in computing niches in which Chinese industries could be internationally competitive, and has promoted the importation of technologies it needs, but cannot adequately produce domestically. Since their export to China has been restricted, high-performance computing systems have not fit neatly into government programs. Chinese industry has neither been able to develop internationally competitive systems, nor to import the advanced systems it needs. Consequently, high-performance computing has been supported in a rather piece-meal fashion under a variety of programs, usually as basic research, or as work on tools needed for research that is the direct focus of the programs. Consequently, modest-scale HPC programs are scattered about at many leading Chinese institutes.

The Chinese have concentrated on developing traditional vector-pipelined and distributed-memory multiprocessor architectures. Efforts to build a system analogous to the Cray-1 began in 1978 at the National Defense Science and Technology University (NDST) in Changsha, Hunan Province. The Galaxy-1, a 100 MIPS machine, passed state testing in 1983. Its successor, a 400 Mflops system with four tightly-coupled vector-pipelined processors called the Galaxy-11 passed state testing in 1992, and reportedly has been used productively since "hen. A new system currently under development, the Galaxy-111, supposedly integrates shared memory and massively parallel distributed-memory architectures. 31

Table 2. [Omitted] High-Performance Computing Systems of the PRC

At least a dozen other institutes have undertaken multiprocessor development projects. The most prominent of these are listed in Table 2. The table shows that commercially available western microprocessors are being used extensively in Chinese projects. The most popular are the transputers, thanks to their built-in communications capabilities, which make them relatively easy to combine into multiprocessor configurations. Other Western processors, such as the i860, 486, and Texas Instruments TMS34020 digital signal processing chip, are used as well. While the available data indicate a considerable lag between the appearance of a microprocessor on the world market and its incorporation into Chinese machines, at least one project, the T9000-based SmC at Quinghua University indicates that this is not universally true.

India

In 1992, India attracted considerable attention in the high-performance computing community when it introduced the first supercomputer developed in a third-world country. The Param 8600, based on commercially available i860 and T800 (transputer) processors, boasted a peak performance of 1.5 Gflops in a 64-processor configuration. In the following years, development teams at half a dozen Indian institutes have vigorously pursued development of more advanced models,

illustrating that a strong domestic microelectronics industry is not a prerequisite to the development of advanced computing systems.

Parallel systems have been built in India for nearly 10 years. The first Indian multiprocessor, constructed in 1986, was probably the MH1, a four-processor system based on Intel 8086/8087 processors. 32 Since then, a dozen or more models, the most prominent of which are listed in Table 3, have been constructed at a variety of institutes throughout India. Of these, the Param systems are the most "commercial." More than 30 Param systems have been installed within India and in Canada,
France, Germany, Russia, and the United Kingdom. 33

Table 3. [Omitted] Indian High-Performance Computing Systems

The efforts to build serious parallel systems began during the late 1980s. A Cray X-MP had been installed at the Indian Weather Bureau in 1986, but was installed with safeguards that made it inaccessible to the scientific community. Disenchanted with the high cost of foreign supercomputers and extensive constraints imposed by export control conditions, Indians pursued indigenous development based as much as possible on commercially available technologies. Lacking suitable microelectronics development and manufacturing facilities, they had little choice.

The current emphasis is on developing new models based on an open, processor-independent architecture using internationally accepted standards for inter-processor communication (e.g., PVM, MPI), input/output, networking, miss storage, and programming environments. While transputers and the established i860 microprocessors offered the most feasible route to high-performance multiprocessing, Indians today appear to be ready to take advantage of other microprocessor lines, including SPARC,
Alphas, and PowerPCs. 36

Applications and tools developing, and porting applications to the parallel architectures remains a principal concern, particularly at the Center for Development of Advanced Computing (CDAC), where over a hundred software developers are employed

Common themes

The examples of Russia, China, and India hive common themes. Each country has viewed parallelism as the most viable route to high-performance computing systems. Parallelism has been seen as a means of overcoming the shortcomings of domestic microelectronics industries unable to manufacture high performance, high-reliability components in the needed quantities. Figure 4 shows trends in the most powerful domestic systems found in these countries, together with the definition of supercomputer used in the export control regulations

Although Russia and China have pursued the development of completely indigenous machines in the past, the number of examples of development and use of systems based on Western, commercial technologies in all three countries is growing. The transputers have been popular systems because of their availability and the ease with which they can be combined into larger configurations. The

TMS32CxO digital signal processors from Texas Instruments have been popular with some researchers for many of the same reasons.

The i860, the earliest 64-bit microprocessor to become widely available, has also been a common computational engine. Several systems in India and Russia in particular have combined the computational power of the i860s with the communications abilities of transputers, integrating these two processors to form a single node in a parallel system.

Although the i860 and. T800 are now considered "old" technologies, 37 each country seems poised to take advantage of commodity technologies as they become available (e.g., processors such as the Alpha, the Pentium/P6, members of the SPARC family of processors, etc.). Given adequate time and financing, it is likely that they eventually will do so. A key question, however, is how quickly they can assimilate the new technologies relative to vendors of "uncontrollable" systems in the Western industrialized countries, chiefly the United States. This question is addressed in the sections that follow.

Figure 4. [Omitted] HPC in Russia, PRC, and India

Trends in "Uncontrollable" HPC Systems

A number of trends in the development of computing are particularly relevant to the export control dialogue. After discussing these trends-performance increases, the commercial market as a driver of HPC, scalability and ease of use, shortening development cycles, and alternatives to direct-sale distributions-we explore the notion of "uncontrollable" systems, and suggest HPC systems whose export might today be very difficult or impossible to control.

Dramatic increases in performance

Improvements in computer hardware that lead to high system performance

*include reducing the time needed to execute an instruction

*increasing the number of instructions executed concurrently

Performance improvements are obtained as processor clock frequencies increase, the number of instructions executed per clock tick within each processor grows, and multiple numbers of processors are combined in integrated systems.

Improvements in these parameters, as well as in the auxiliary technologies (interconnects, memory, storage, software) necessary to create balanced, efficient, and usable systems have been particularly dramatic at the high-end of the workstation market. Because progress has been so rapid and the cost-effectiveness of the technologies is so great, nearly all large-scale, massively parallel systems now build on technologies developed primarily for the workstation market.

The Rate of development of microprocessors

As Figure 5 indicates, microprocessor performance has increased exponentially during the 1990s. Today, single-processor workstations match or exceed the performance of single-processor supercomputers such as the Cray Y-MP from the late 1980s. Performance increases are a result of faster clock periods (40 MHz vs. 150-300 MHz), superscalar execution of more than one instruction per clock cycle, and the use of advanced cache memory management to support rapid instruction execution. Commercial microprocessors such as these are now the basic building blocks of nearly all commercial parallel systems.

Figure 5. [Omitted] Advances in 64-bit Microprocessors

The shift to parallel processing in mid-range systems

The workstation vendors have adopted multiprocessing relatively recently, but extremely dramatically. Silicon Graphics, Inc. was the first workstation vendor to introduce a multiprocessor system, in 1988, but was not joined by other vendors until the introduction of Sun Microsystem's SPARCstation 10 multiprocessor, introduced in June 1992. 38 In the three years since then, all workstation vendors have introduced multiprocessor models. Such systems, in which processors are tightly coupled via shared memory, are commonly referred to as symmetrical multiprocessor (SMP) systems. While most SMP installations sport 4-16 processors, the maximum number of processors in non-clustered (shared-memory) systems is 12 in HP's T-500 and DEC's AlphaServers, 36 in Silicon Graphic's Challenge series, and 64 in Cray Research's CS6400.

Memory access constitutes an inherent bottleneck in shared-memory systems, and it will be difficult to increase the number of processors in shared-memory configurations beyond today's levels. However, vendors are pursuing hierarchical architecture's that would enable shared-memory systems to be combined in an integrated, yet distributed fashion, allowing the number of processors to grow further to hundreds or thousands of units. Convex's Exemplar system is based on this principle. Each node in this distributed memory system consists of up to eight RISC-RA microprocessors, arranged in a shared-memory, symmetrical multiprocessing configuration.

Although the transition from shared-memory to distributed-memory systems has historically been a very difficult one for the industry, there are strong incentives to make the investments needed to

negotiate it smoothly. If Convex's experience is any indication, the prospects are promising. 39

The Commercial market as a driver of parallel systems

The Mid-range parallel systems market

According to Superperformance Computing Service (SCS), a market research firm specializing in parallel processing, the parallel and high-end SMP market achieved $2.5 billion in sales in 1994. 40

The commercial parallel computing market alone is expected to grow to $5.2 billion by 1998. 41 While still smaller than the $75 billion PC market or the $30 billion low- and mid-range workstation market, the SMP and low-end Massively Parallel Processing (MPP) segment is the fastest growing, experiencing sales increases of over 40% per year, according to SCS.

Although the demise of the mainframe market has been overstated, parallel SMP systems have made inroads into corporate data processing, competing with mainframes for traditional applications, such as on-line transaction and batch processing. These systems offer extensive processing power in a machine that is, in many cases, less costly to purchase, operate, and maintain than a mainframe. 42 Furthermore, since most parallel SMP systems are based on industry standards, companies are less likely to find
themselves locked into a proprietary system. 43

Commercial applications of massively parallel systems

High-end, massively parallel systems are also penetrating the commercial market, but not nearly with the force of the SMPS. The commercial market is in its infancy, but has considerable room to grow. AT&T Global Information Systems, IBM, Tandem, and ncube have established a presence in this market, and Unisys recently unveiled its new OPUS massively parallel systems, based on Intel Pentium processors and an interconnect licensed from Intel's Supercomputer Systems Division. 44

Highly parallel systems are best suited for applications that involve very large data sets and sophisticated computation. Commercial applications in this category include data mining and buying-pattern analysis, advanced decision support involving cross-company database analysis, real-time pricing and merchandising management, and others. --

While the commercial market for MPP systems is growing, it will remain significantly smaller than the SMP market. According to industry analyst Gary Smaby, the number of applications that can Lake advantage of MPP systems is a small friction of commercial applications; SMP is more appropriate than MPP in 90% of commercial installations. 46 Although they are easier to use now than in the past, considerable expertise is still needed to program, tune, and administer MPP systems. Wary of the difficulties, companies have tended to purchase entry-level systems with fewer than 16 processors, limiting their initial investment, while leaving open a clear upgrade path.

Scalability and ease of use

Scalability has always been desirable in parallel systems, but has become a competitive necessity in the commercial markets. Scalability allows users to "start small" and increase the performance of their systems incrementally as their needs and experience grow, while preserving their investments in applications software. SMP architectures, based on a shared-memory model with a single, integrated operating system managing processor activities, are the easiest parallel systems to program, and the most transparent to applications. In most SMPs today, the number of processors can be increased and used effectively with no changes in the applications. Upgrading a system involves turning the machine off, inserting new processor boards, and then re-booting the system. The Cray CS6400 is even easier to upgrade; processors can be inserted while the machine is running. The system automatically takes the new processors into account, with no interruption in an application's execution.

Massively parallel, distributed-memory systems permit a greater range of scalability, allowing increases by a factor of a hundred or more in the number of processors. They are, however, significantly more difficult to use than SMPS. Advances in systems software now permit existing applications to run without modification on different configurations of a massively parallel system, but extracting maximum performance is difficult. Users spend considerable effort tuning a system so that it runs efficiently for their applications, and often need to rewrite applications to adapt them to the distributed architecture. Massively parallel systems vary greatly in the ease with which they can be upgraded. Large configurations usually require extensive vendor involvement.

Short development cycles

A development cycle is the time that elapses between the introduction of one model and a successor model at a comparable price. The development cycle for many high-end, multiprocessor workstations today is 1.5-2 years. For lower end models, it is less. For high-end, massively parallel or vector supercomputers the development cycle is more variable, but typically in the range of 2-4 years. Companies may continue to manufacture a given model for longer than the average development cycle, but except for replacement parts, production of most workstation models is discontinued 3-4 years after the model was introduced.

Use of VARS, OEMS, systems integrators, and dealership networks

High-end vector and massively parallel systems vendors (e.g., Cray, Convex, Intel, IBM) usually rely on direct sales. High-end systems are sold in relatively small numbers and require vendor involvement during installation and operation. Workstation vendors, including SMP vendors, however, rely on very different distribution channels. For example, Digital Equipment Corporation (DEC) sells all of its models through an extensive network of value-added re-sellers, original equipment manufacturers,, systems integrators, and dealerships. Only large, corporate accounts are handled directly by DEC. Other vendors similarly sell significant fractions of their products through third parties. Consequently, there is

no single company that always has detailed oversight over a product from the time it leaves the factory until it is installed at a user location.

The "controllability" of commercial parallel systems

Export control regimes are based on the assumption that certain products are controllable, that is, their export and operation can be regulated by export control authorities and vendors, and regulated at some tolerable cost in time, effort, and money. It is difficult to establish a precise definition of uncontrollable for at least two reasons. First, tolerable cost is a subjective term, and depends in part on the perceived risks of an undesired export and (mis)use of a system. It is unlikely that all vendors, government officials, and users will have the same definition of tolerable cost. Second, controllability is a continuous function, not a binary condition as the term seems to imply. Some systems are more controllable than others.

Even without a precise definition of uncontrollability, however, there is general agreement that the following qualities affect the ability of export control authorities, in concert with vendors, to track the location of a given computer system, monitor its operations, and enforce appropriate use. In random order, these are:

*Size. Small systems are easier to move than large systems. For computer systems, the larger systems also tend to require greater infrastructure support: liquid cooling systems, special-purpose power supplies, and so on. Smaller, rack or desk-top systems can be run without building special rooms to house the units.

*Age. Given the rapid rate of obsolescence of computer equipment, several-year-old systems are replaced on a regular basis by newer systems. One high-end HPC vendor claims that half of their systems are de-installed within four years of installation, and nearly all machines are taken out of service within 8-10 years of installation. When product cycles are 1-2 years in length, secondary markets for such systems are extensive 2-3 years after a product's introduction. As older systems are de-installed, resold, or otherwise removed from their original premises, vendors may not have accurate or current information about their location and use.

*Scalability. Scalability is the ability to increase the computational power of a system incrementally, by adding processing elements, memory, interconnects, and so on. There are two dimensions of scalability: the range in performance from a small system to the largest system to which it can be upgraded, and the ease with which computational resources can be added. When systems are extensively at-id easily scalable, it is difficult to prevent a small, unrestricted system from being upgraded in the field into one that, if sold as a large configuration, would be subject to export controls. When such an upgrade can be accomplished without the involvement of a trained vendor representative, the vendor may not be aware that a more powerful system has been created. In practice, HPC vendors rely extensively on their "eyes and ears" in the field, the trained personnel that service

installations, to monitor a system's use. Users who perform upgrades themselves without vendor involvement can undercut this mechanism.

In cases like these, the principal barrier to high performance is financial, not technical. Today's SMPs may cost over a million dollars in a maximum configuration, but an entry-level version (below current control thresholds and easily upgradable to maximum configuration) may be obtained for a few hundred thousand dollars or less. 47

*Number of Units in the field. It is easy to know the location of a dozen units. It is virtually impossible to know the location of tens of thousands of units. While vendors maintain customer databases that identify the supposed location of systems, at some point it becomes economically infeasible for companies to monitor and verify this information on a frequent basis. It is difficult to define a precise threshold at which the number of installations' becomes large enough to make tracking them difficult. Company estimates vary from about 200 to several thousands of units.

*Dealership network. When systems are manufactured by the vendor in only a few locations and shipped directly to customers, the vendor is able to track the system easily from construction to installation. When a vendor sells systems through a network of dealerships, value-added re-sellers (VAR), systems integrators, and original equipment manufacturers (OEM), no single company has full and complete oversight over the entire delivery process. When a vendor sells systems in bulk to a dealer, rather than through a dealer to a predetermined customer, the vendor may not be able to track which systems are installed where. For larger systems, end-use confirmation is currently required for all sales, but this only partially eases the difficulty of tracking them.

*Cost of entry-level systems. The cost of entry-level systems is closely related to the size of the potential market which, in turn, is related to the size of the installed base. Measured both in dollars and numbers of units installed, the market for $1+ million systems is smaller than the market for workstations, which is smaller than the market for personal computers. In computing, approximately half a million dollars represents a crucial marketing threshold, for systems below this threshold lie within the budgets and purchasing authority of many universities and corporate departments. Thus systems with entry-level prices below this level enjoy significantly larger potential markets than more expensive systems. Similarly, systems with entry-level costs in the $100-200,000 range enjoy still larger potential markets.

Table 4 illustrates the "controllability" of several systems available today.

In this report, systems like the Cray CS6400 and Silicon Graphics Challenge series represent the most powerful uncontrollable systems available in mid-1995. Several thousands of chassis of the latter have been installed which, given the money and inclination, could be upgraded easily into a maximum configuration. The systems are sold through a sizable network of third parties such that the vendor loses direct control over the systems. While Cray and Silicon Graphics are confident they still

know the location of nearly all chassis, the ease with which the systems can be upgraded in the field makes it impossible as a practical matter to know exactly what the configuration is.

Table 4. [Omitted] Controllability of Selected Commercial HPC Systems49

One of the principal justifications for export controls is that they slow, or otherwise handicap the acquisition of HPC technologies by countries subject to the restrictions. When controls are effective, these countries pay a premium in time and expense to acquire the systems, lack crucial vendor support and training, run a high risk of detection, or are forced to pursue their goals using much less desirable technological approaches. The systems identified here as uncontrollable are not likely to suffer from these handicaps to a degree that would seriously impede efforts to circumvent controls. Occupying what may be viewed as the high end of the workstation market, the SMP systems not only are available at moderate cost, but also are designed to function for years with little or no vendor support. As the secondary markets develop, they may be sold and relocated without attracting much attention.

The performance of "uncontrollable" systems will rise quickly in the next few years, as shown in Figure 6. Since some time is needed for an installed base to build and a secondary market to emerge, we feel that such systems become uncontrollable as they reach the end of their product cycle, approximately two years after they are first shipped. Figure 6 reflects this two-year lag between introduction and uncontrollability, so that systems considered uncontrollable in 1997 are being introduced in 1.995.

Figure 6. [Omitted] Performance of 'Uncontrollable" Symmetrical Multiprocessor Systems

What about clusters of workstations?

The discussion thus far has concentrated on individual computing systems that are viewed as single products by vendors. Increasingly, however, computing installations consist not of a single high-performance system and attached peripheral devices, but of a network of interconnected systems, working cooperatively and concurrently on one or several tasks. A particularly popular arrangement is a cluster (or "farm") of workstations connected by non-proprietary networking technologies such as Ethernet, Fiber Distributed Data Interconnect (FDDI), High-Performance Parallel Interconnect (HiPPI), or Asynchronous Transfer Mode (ATM). The activities of workstations are coordinated by specialized distributed software such as Parallel Virtual Machine (PVM), Linda, Express, and others that facilitate the distribution of tasks to workstations (load balancing) and communications between tasks running on different processors.

Clustered workstations have attracted attention because they potentially can provide a great deal of computer power at modest costs and risks. Most installed workstations are greatly underutilized. Outside of normal work hours (e.g., nights and weekends), the machines are idle. Clustering workstations offers the hope of applying those idle machine cycles to useful work. In principle, it is possible to combine the computing power of numerous networked workstations to achieve a performance comparable to higher-end mainframes or traditional supercomputers. Hard data comparing the performance of clusters with vector-pipelined and massively parallel systems are difficult to come by, but some studies demonstrate that for a considerable number of applications, configurations of up to 50 approximately 16 workstations can offer performance levels comparable to more integrated systems.

To what degree should clustered workstations define the upper performance boundary of uncontrollable systems? A common argument is that since workstation clusters are based on commercially available, uncontrollable technologies, export control thresholds should be no lower than the performance offered by such conglomerations.

There is little question that most workstation clusters are uncontrollable. A collection of computers is only as controllable as its most controllable component. Based on conventional workstations, commercially available networking technologies, and public domain software (e.g., PVM) that is obtainable gratis over the Internet, most clusters have no easily controllable elements. However, clustered workstations have certain drawbacks that make a direct comparison with more tightly coupled systems problematic.

Table 5. [Omitted] Spectrum of HPC Architectures

Current computer architectures can be placed along a continuum, as shown in Table 5. Ad hoc clusters are based on an existing collection of workstations connected via a conventional network such as Ethernet or FDDI (100 Mbps). The machines are usually physically distributed throughout an organization. Dedicated clusters are more tightly coupled. They usually incorporate workstations of the same model that are physically placed close to one another (often in a single rack, without monitors), and connected by a high-speed interconnect such as FDDI, HIPPI, or a proprietary technology. As one moves down the continuum from more to less tightly coupled systems, it becomes increasingly difficult to harness the potential computing power of the computing elements for broad categories of applications. In other words, all other things being equal (usually not the case!), a machine with a more tightly coupled architecture is preferred to a loosely coupled system of comparable power. Traditionally, Cray supercomputers have been more attractive than massively parallel systems because they have been easier to program to run efficiently, and offer good performance on a more diverse set of applications.

In distributed-memory systems, data are distributed among the processors. Making sure that the right data elements are at the right processors at the right

time is crucial to efficient execution. When processors are functioning on parts of the same problem, data elements often need to be sent from one processor to another. The amount of computation relative to the amount of movement of data between processors is referred to as the granularity of the application. Bandwidth (through-put) and latency (transmission delay) are crucial parameters of the interconnect between processors. The lower the bandwidth, the higher the latency, and the less scalable the interconnect, the more of a bottleneck the interconnect becomes. The more the interconnect is a bottleneck, the more coarsely grained an application must be to run effectively on the system.

Clustered workstations are usually connected by networks with bandwidth and latency that are]--2 orders of magnitude inferior to the interconnects used in more tightly coupled systems. Consequently, they are most useful for applications or tasks in which the ratio of computation to inter-processor communication is high. Clusters have been used with excellent results primarily when used to improve system through-put, or to tackle replicated problem applications. When clusters are used to improve through-put, completely independent processes are farmed out across the cluster in a manner that balances the load on each. 52 Replicated problems are those in which the same problem must be solved many times on varying data sets to produce the final result. Individual problems can be solved independently, and the results combined later. Examples include ray tracing, some flow problems, and image analysis. In each of these cases, inter-processor communication is very low, if not zero.

When workstation clusters are applied to executing the parallelized code of a single application to reduce the time to solution, experiences appear to be quite mixed. 53 In particular, clusters did not appear to be competitive with vector-pipelined systems for shallow-water modeling, weather prediction, or problems involving the difficult-to-parallelize solutions of sparse linear systems of equations. 54 In general, clusters can have competitive performance with applications in which the granularity of the algorithm can be made sufficiently coarse to map well onto the computation/communication ratios offered by a cluster of workstations. This accounts for some, but certainly not all, applications of national security importance. This point will be discussed further in the next chapter.

Loosely coupled distributed systems often provide a more cost-effective solution than tightly coupled systems for highly parallel applications with high computation/communications ratios. However, it is not the case that a loosely coupled system is as broadly useful as a tightly coupled system with comparable computing power. An export control threshold established on the basis of a system listed high on Table 5 can therefore be applied to a system lower in the table. The reverse is not necessarily true. While a threshold based on machines with an SMP architecture can certainly be applied to distributed-memory systems and workstation clusters, a threshold based on workstation clusters should not equally be applied to shared-memory systems. For this reason, it is wiser to allow SMP architectures to set the lower bounds for an export control threshold, even though a workstation cluster may provide significantly greater performance on particular applications. 55

The Future Relationship of "Uncontrollable' and Foreign Indigenous HPC Systems

Common drivers of HPC development world-wide

Throughout the world, Computing platforms in all performance categories are being constructed on the basis of commercially available technologies developed principally for the workstation and server industry segments. Most notably, processors designed initially for workstations are being used as the building blocks of parallel processors. The high-end Cray T3D parallel system is based on the Alpha processor from Digital Equipment Corporation; IBM's SP2 incorporates the same microprocessors found in its RS6000 workstations. Intel's Paragon systems incorporate the i860 processors, and Unisys. Commercial "data mining" systems rely on Pentium microprocessors. HP RISC-PAS, INMOS' transputers, Sun SPARCS, and the IBM/Motorola PowerPCs are all used in workstations, parallel servers, and mid-range and/or high-end parallel systems.

The underlying reason for this phenomenon is that it is much more cost effective for parallel systems developers to use commercially available microprocessors than to develop proprietary processors of comparable performance. Such processors cost billions of dollars to develop and manufacture. Only personal computers and workstations offer markets that are large enough to recoup these outlays.

With the exception of INMOS' transputers, all of these microprocessors come from American companies or from foreign licensees of U.S. designs. The combination of a highly competitive concentration of firms involved in microprocessor development, a large and demanding domestic market, and a strong supporting infrastructure of software and manufacturing technology suppliers has given the American companies an overwhelming position in the microprocessor market. The enormous sums and extensive expertise needed for research and development make barriers to entry very high. Computer engineers throughout the world who seek to build high-performance systems have little choice but to ride this wave of technological development.

Barriers to development of foreign indigenous systems

While workstation and personal computer microprocessors can be considered commodities, there are significant difficulties in combining them into a functional parallel processing system. The main difficulties are in board construction, interconnect construction, software development, and overall system architecture design.

Successful use of microprocessors requires detailed, complete, up-to-date knowledge about the exact specifications-and deviations from specifications-for given chips. This information is usually obtained through close contacts between the microprocessor factories and systems developers using the factory's products.

When a new microprocessor is first introduced, manufacturers tend to give priority attention to customers with the largest potential commitment in the new product, or the most established relationships. Small or less established systems developers

may find themselves starved for the information needed to build computer systems in a timely fashion.

At present, virtually no systems developers in countries of national security concern have the type of relationship with American or European microprocessor developers that would give them early, in-depth access to such information. These problems diminish over time as information becomes more widely disseminated. Nevertheless, the lack of rapid access to information early on in a microprocessor's life cycle means that foreign developers are likely to lag behind Western counterparts using the same microprocessors.

The clock rates of microprocessors have increased dramatically during the 1990s, from 20 MHz for the Motorola 88000 Reduced Instruction Set Computer (RISC) processor (circa 1989) to the 200-300 MHz of today's Alpha, PowerPC, and forthcoming R10000 processors from Silicon Graphics' MIPS division. Other processors from Sun and Intel will reach these frequencies shortly. As the clock rate increases, timing issues throughout a printed circuit board become acute, and very high precision design and construction are needed to prevent timing mismatches, signal echoes, and so on. While these factors would normally greatly increase the difficulty of designing and building products using these components, chip manufacturers themselves are taking steps to shield users from this complexity.

To expand their customer base, microprocessor vendors must make their products as accessible as possible to a broad community of systems developers. They are making their microprocessors easier to possible work with by designing chips so that high frequencies are used only internally to the chip, while off-chip circuitry can function at lower clock rates. They are also providing, or building into the chips themselves, the support circuitry needed to connect the microprocessor with other components and devices. Board-level products proliferate. Under these circumstances, building systems out of ready-made boards manufactured in the West will become the fastest route to developing systems that are in any sense indigenous. Assembling boards is a much easier, lower-tech process than the construction of the boards themselves. Boards will be reliable, sophisticated, and available directly or indirectly from Western vendors.

As new generations of products enter the mass markets, building multiprocessors with modest numbers of processors (i.e., 4-16) will be a relatively straightforward process. The ability to construct larger Configurations is likely to vary considerably from country to country, however. Assembling the necessary hardware may not be the most serious problem encountered. Today, most of the investment of massively parallel systems vendors is focused on developing the software needed to extract maximum performance from the hardware. 56 The financing or ability to build such software is not something that can be controlled by U.S. export control mechanisms. The need to develop such software remains, however, a challenge that foreign developers must meet.

In summary, talented systems builders in countries of control interest are likely to continue to incorporate contemporary microprocessors and other commercial

technologies in their systems. They are likely to lag behind U.S. practice by at least several months, but probably by years for the more advanced systems.

U.S. uncontrollable systems encroaching on areas in which foreign developers could gain advantage.

In the past, most high-performance computing projects in countries of control interest involved parallelism. Lacking an ability to develop uni-processor systems with the performance of leading Western systems, engineers naturally turned to multiprocessing as a means of using relatively slower processor performance to achieve high system performance. As long as non-Western multiprocessors could be compared with Western systems with fewer processors, foreign developers could hope to "keep up" (at least theoretically) with the West.

Today, parallelism is found in the mainstream of commercial computing. During the 1990s, the number of processors in symmetric multiprocessor systems has grown from 2 to as many as 64 in the Cray CS6400. Future, hierarchical configurations of SMP systems will incorporate multiple times as many processors, and distributed-memory systems like IBM's SP2 are likely to find wide acceptance in the commercial sector.

No longer can foreign engineers achieve an advantage merely by combining processors. To maintain performance parity with Western SMPS, they will either have to use faster processors or greater numbers of processors, or improve performance through advances in software. The first option is essentially impossible; the second, increasingly unlikely; and the third, difficult. Leading Western MPP vendors are also committing substantial fractions of their R&D budgets to software development.

A second area of potential advantage for foreign engineers was the relatively low cost of trained talent. It was cheaper to support the in-house development of parallel systems than to purchase an expensive Western supercomputer, even when that was possible. Here also, foreign advantage is dwindling. The growing size and intense competition of the SMP market will continue to drive the cost of such systems (e.g., $/MIPS) down to the point where non-Western parallel projects become economically infeasible. While ground-tip domestic projects may continue to be supported because of the benefits of learning by doing, such projects are not likely to make sense on financial grounds alone.

Key Findings and Conclusions

Figure 7 shows the overlay of trend lines from Figures 4 and 6. This "spaghetti-like" mix illustrates several important points of considerable significance for the export control regime.

Performance of "uncontrollable" U.S. systems has increased dramatically, eclipsing most, if not all, non-Western HPC projects.

While several non-Western countries have in the past developed high-performance computing systems that exceeded the performance of Western "uncontrollable" systems, this is no longer true. One of the most significant developments in computing in recent years has been the evolution of the workstation industry in the direction of multiprocessing. The performance of symmetrical multiprocessor systems (SMP) has grown by two orders of magnitude in the three years since their introduction. This growth has been driven by advances in microprocessor design and, very significantly, the use of parallelism in an industry segment that formerly consisted solely of uni-processor systems. While there are limits to the degree of parallelism possible in shared-memory systems, the evolution of such systems is likely to be in the direction of hierarchical systems, consisting of shared-memory subsystems grouped together in a distributed-memory fashion. The degree of parallelism is likely to continue to increase for the foreseeable future.

This industry segment is being driven by market demands that make systems less and less controllable. Such features include ease of use, small size, reduced needs for regular and on-site vendor involvement, easy scalability, and distributed control over distribution channels. The number of systems sold in this market-thousands and, in the future, tens of thousands of units-significantly reduces the feasibility of tracking and monitoring individual units to ensure compliance.

Figure 7. [Omitted] Performance of Foreign and Domestic HPC Systems

The principal implication is that the most powerful systems readily available to foreign users are now likely to be of Western origin. Western machines usually considered part of the workstation market segment now dominate more than ever the discussion of a lower bound for export control thresholds.

Non-Western indigenous systems are likely to lag behind Western uncontrollable systems, both in raw computing power and practical utility, although the non-Western systems should be able to "keep pace" with Western systems over time.

While the performance of "uncontrollable" systems now exceeds that of non-Western systems, there are few technical reasons why non-Western systems developers couldn't develop systems that exhibit dramatic increases in performance as well.

The underlying reason is that multiprocessor systems developers throughout the world, in Western and non-Western countries alike, are riding the same waves of technological advance. They are all relying on commercially available microprocessors and storage devices. The interconnect technologies and systems software used to run distributed systems are, or are soon to become, commercially available and/or in the public domain. The forces of the marketplace are pressuring vendors of these technologies not only to improve their functional characteristics, but to make them easier to incorporate into finished systems. Western and non-Western developers alike will benefit from such developments.

Some lag between advances in Western and non-Western systems, on the order of months or years, is likely to persist, as Western systems developers, particularly Americans., have faster access to vendor information and early production runs than their non-Western counterparts, and are investing large amounts of resources in developing the software to extract performance from their systems.

Efforts to set export control thresholds within the envelope of non-Western and "uncontrollable" Western systems are likely to be problematic and ineffective.

The trends in Western "uncontrollable" and non-Western HPC efforts have become rather closely coupled. Figure 6, in particular, shows how the trend lines of the individual Western vendors themselves are interwoven in a dense "spaghetti" mix. Together, the numerous trend lines on Figure 7 create an envelope which is difficult to dissect in any meaningful way.

Thresholds drawn through the middle of the spaghetti mix are likely to be ineffective and have awkward political consequences. First, thresholds that make the most sense are those that are enforceable. Much of the discussion in this chapter has focused on the controllability of various individual product lines, arguing that restrictions on technology that are largely uncontrollable can be circumvented relatively easily by those who have the money and inclination. There is no point within the envelope that neatly divides the controllable from the uncontrollable.

Second, thresholds that make sense reflect changes 'in technologies or markets that make systems above the threshold quantitatively and qualitatively different from those below the threshold. In other words, presumably the definition of supercomputer is non-arbitrary, and has embodied within it a technological basis for distinguishing "supercomputers" from "non-supercomputers." There is no such technological distinction between systems within the envelope. Wherever the threshold is drawn, it will be possible to find at least one system lying below the threshold and increase its performance with little difficulty to make it lie above the threshold will changing the essential nature of the system or its mode of operation.

Third, the CTP metric is too imprecise to adequately distinguish between the deliverable performance of systems lying close together in the envelope. Actual performance is highly dependent not only on architecture, but also on the nature of the application, the algorithms used, and the maturity of the systems software. Differences of tens of percentage points in CTP rates of two machines may be compensated for by non-hardware means. Thresholds within the envelope that distinguish between systems with roughly comparable CTPs are not likely to reflect differences in the real utility of such systems.

Clustered workstations contribute to the density of the spaghetti mix, but should not by themselves be used to justify a lower bound for an export control threshold.

Because a system is only as controllable as its most controllable parts, clusters of workstations based on commodity technologies are inherently uncontrollable. For many applications, clusters offer an alternative, viable route to high performance.

While changes in technologies are narrowing the differences between distributed clusters and integrated systems, clusters have inherent weaknesses that limit their real usefulness (when applied to single applications) to applications using coarse-grain parallelism. While it is possible that clusters of systems lying within the uncontrollability envelope can deliver high performance on certain classes of applications, they should not generally be treated on an equal basis with tightly coupled systems of comparable CTP (assuming the CTP of a cluster can be determined).

It is all but inevitable that some day, if it has not already happened, adversaries will use American-made computer in the design or operation of a system that handles U.S. citizens and property.

The foregoing discussion has identified the practical limits of government's or industry's ability to regulate the diffusion of computer technology throughout the world. It is therefore likely that sooner or later an adversary will use uncontrollable American technology against American lives, property, and interests. However regrettable this may be, there is little an export control policy focused on hardware exports can do to prevent it.

Chapter 3 Notes

29 Peter Wolcott, "Soviet Advanced Technology: The Case of High-Performance Computing," Ph.D. Dissertation, University of Arizona, 1993.
30 Peter Wolcott, "High-Performance Computing in the People's Republic of China," working paper, 1994.
31 D. Kahaner, "Kahaner Report: HPCC '94 (Singapore)" (October 7, 1994), http://cs.arizona.edu/japan/kahaner.reports/hpcc.94a.
32 Developed at C-MMACS, the Center for Mathematical Modeling and Computer Simulations, Bangalore. D. Kahaner, "Computer and math modeling activities in India I," ftp:\\cs.arizona.edu\'apan.kahaner.reports\india.93a.
33 Official Promotes New 'Param 9,000' Supercomputer," The Hindustan Times (December 5, 1994), p. 11, as cited in FBIS-NES-94-237, Daily Report: Near East & South Asia (December 9, 1994), pp. 38-39; "Indian Technology is Powering Ahead," Electronics Times (January 12, 1995), p. 18.
34 The Hindustan Times (December 5, 1994), p. 11.
35 Pace-Plus' Supercomputer from Indian Defense Research Organization," HPCwire (May 4, 1995).
36 CDAC literature, 1994.
37 Intel Corporation never did develop a true successor to the i860, first introduced around 1989. In its own Paragon parallel systems, the company has until the present used i860S. The next generation Intel systems will use the P6 processors, more properly viewed as the successors to the Pentium.
38 Norris Parker Smith, "SGI Launches New Chip, New Systems; Targets New Markets," HPCU'ire (January 27, 1993); "Sun Unveils SPARCstation 10 Multiprocessing Workstation," Stipernet (June 1, 1992).
39 Norris Parker Smith, "Convex's Global Shared Memory: A Competitive Edge?" HPC7t7ire (April 21, 1995).
40 C. Babcock, "Into the Mainstream," Computer World MPP & SMP Special Report (March 27, 1995), pp. 1-3.
41 MRJ/IBM Team up to Take on Commercial Opportunities in HPC Market," HPCwire (September 28, 1994).
42 In many cases, at least. There exist horror stories about new systems not being able to support the high-availability legacy applications that used to run well on traditional mainframes.
43 Richard W. Sevcik, "Viewpoint: For Commercial Multiprocessing, The Choice is SMP," IEEE Spectrum (January 1995), p. 50.
44 Norris Parker Smith, "Unisys/Intel Launch New Parallel/Unix/Pentium OPUS," HPCwire (April 28, 1995).
45 C.G. Willard, "Let Your Needs Dictate Strategy," ComputerWorld MPP & SMP Special Report (March 27, 1995), p. 13.
46 Craig Stedman, "What You Don't Know ... Will Hurt You," Computerworld MPP & SMP Special Report (March 27, 1995), pp. 4-9.
47 For example, last year (March 1994) SGI's PowerChallenge in a maximum (18-processor) configuration had a price tag of $1.2 million. The entry-level (2-processor) version was priced at $128,000. (Norris Parker Smith, "Silicon Graphics Multiprocessors: Power and Price," HPCziiirc (March 11, 1994).) While price/performance ratios are changing rapidly, this price range continues to define

a very active market. Most SMP vendors continue to build and price their high-end systems to fall in this range.
48 While Cray is aware of where each chassis is located, it is difficult to prevent field upgrades by the user.
49 The systems listed are not the most advanced models currently offered by these vendors.
50 In T.E. Anderson, et al., "A Case for NOW (Networks of Networks)," IEEE Micro, 15 (February 1995), pp. 64-84, the authors describe the results of modeling the execution of the GATOR chemical tracer model on three types of systems: a 16 processor Cray C-90, a 256 processor Intel Paragon, and a cluster of 256 RS-6000 workstations. This is a highly parallel application tracing the flow of chemicals in the Los Angeles Basin. The simulation of the performance of the cluster showed that it exceeded that of both the Cray and the Paragon, but only when equipped with a high-bandwidth ATM interconnect, a parallel file system, and a very low overhead (e.g., not PVM) messaging system.
51 The SP2 actually straddles the dedicated cluster and tightly coupled distributed-memory categories. The SPI originated as a clustered workstation with the addition of a proprietary high-speed switch.
52 One of the rationales for the purchase of a supercomputer in the past has often been not just improved performance on individual applications, but the time and cost savings possible when an organization has many applications to execute. When an organization has enough applications to use a sizable fraction of the computing cycles available on a supercomputer, these machines in the past provided the lowest cost source of Mflops. The low cost per Mflops of workstations is making the use of clustered systems attractive for such high-volume computing environments. For a helpful treatment of clustered workstations, see Norris Parker Smith's columns in the January 21, January 28, February 4, and February 1994 issues of HPCwire.
53 See J. Mohr, "Clustered Workstations: The Dominant Parallel Architecture?" RCI, Ltd. (1994). Timothy Mattson of Intel's Supercomputer Systems Division has collected some data on the performance of workstation clusters: "Supercomputing on Workstation Clusters, Draft 2.0" July 5, 1995. He presents a number of examples of clusters of workstations (early 1990s vintage) out-performing the equivalent of a single processor Cray C90 (CTP: 1437). Clustered workstations worked well on applications involving ray tracing, molecular dynamics, seismic signal processing, etc. Some of these applications were embarrassingly parallel, but others only involved an algorithm granularity that matched the processor/communications capabilities of the cluster well. Clusters did not perform as well on other applications. For explicit finite-difference partial-differential equations solutions for modeling shallow water, weather prediction models, and sparse linear equation solvers (a very important, common, and hard to parallelize problem in technical computing), clusters were not competitive with integrated parallel systems like Intel's parallel systems (nor, presumably, with vector-pipeline processors). For most of the cases considered in Mattson's paper, reasonable speedups were often observed for clusters with up to 8-12 nodes, but few exhibited significant speedups for clusters of greater size.
54 Mattson, Ibid.
55 Another problem discussing workstation clusters in the context of export control is that there is no approved way of computing their CTP. It does seem

clear, however, that computations such as that in equation 1 in the Computer Systems Technical Advisory Committee (CSTAC) report "Recommendation on High Performance Computers," January 30, 1995, that assume workstations in a cluster operate at 75% efficiency are overly optimistic for all but the most coarsely grained and "embarrassingly parallel" problems.

56 See Norris Parker Smith, "Thinking Machines Seeks Rebirth, Plans Double Strategy," HPCzvi're, (May 19, 1995).

CHAPTER 4. NATIONAL SECURITY APPLICATIONS FOR HIGH-PERFORMANCE COMPUTING

HPC Applications

In discussing applications of national security concern, it is more common to discuss projects than actual "applications"-that is, computer software programs. For example, in discussing the design of a "stealth" aircraft, there is of course no such computer program; rather, the design project is facilitated by a variety of computer programs. Most often, this software is custom-developed, based upon the mathematical routines and algorithms commonly used by the particular engineering discipline involved. Programs developed to support particular technical areas generally have a common basis, and some generalities can be made regarding their computational difficulty. Table 6 lists nine computational technology areas common to U.S. Defense Department science and technology (S&T) projects. Computational areas common to the developmental test and evaluation (DT&E) community are listed in Table 7. The thirteen disciplines listed, which are described in the following section, are commonly employed in both nuclear and conventional weapons development programs and military operations. Cryptology represents a fourteenth distinct computational area.

CCM Computational Chemistry and Materials Science
CEA Computational Electromagnetics and Acoustics
CEN Computational Electronics and Nanoelectronics
CFD Computational Fluid Dynamics
csm Computational Structural Mechanics
CWO Climate, Weather, and Ocean Modeling
EN4 Environmental Quality Monitoring and Simulation
FMS Forces Modeling and Simulation/ C41
SIP Signal and Image Processing

Table 6. Computational Technology Areas (CTA) for Science and Technology Projects 57

DBA Database Activities
RTDA Real-Time Data Acquisition
RTMS Real-Time Modeling and Simulation
TA Test Analysis

Table 7. Computational Functions (CF) for Developmental Test and Evaluation Projects 58

Computational Chemistry and Materials Science uses chemical and molecular design tools to model and evaluate mechanical, molecular, and electronic structures and their interactions. New chemical systems are designed and optimized using methods of quantum chemistry and molecular dynamics.
Computational Elcctromagnetics and Acoustics involves the calculation of high-resolution, multi-dimensional solutions for electrical and acoustic fields in support of communications and surveillance technology development.

Computational Electronics and Nanoelectronics are used for the simulation of electronic devices, characterization of nanoelectronic devices, and determination of the electronic structures of new materials.

Computational Fluid Dynamics is one of the most frequently encountered families of applications in weapons design and evaluation. CFD is employed in the design of aerodynamic (e.g., aircraft, missiles) and hydrodynamic (e.g., ships, submarines, subsurface weapons) vehicles and represents a significant portion of the HPC performed in support of defense programs. The most computationally stressful applications include three-dimensional, high-resolution solutions of Navier-Stokes and Large Eddy Simulation equations.

Computational Structural Mechanics. Along with CFD, CSM is one of the most important national security applications for HPC, and one of the most computationally stressful. Significant applications include design and evaluation of advanced armor and armor-piercing weapons, and the design of deep penetration weapons ("bunker busters"). CSM is also used to model and evaluate the survivability of weapons platforms (e.g., ships, tanks, submarines, aircraft).

Climate, Weather, and Ocean Modeling. Refinement of models of the earth's climate, world and regional weather, and weather systems is important not only for scientific research, but for military operations. Ocean modeling supports refinement of anti-submarine warfare (ASW) sensors and signal processing.

Environmental Quality Monitoring and Simulation involves modeling of chemical and noise contaminant transport and effects in various ecosystems, in support of the development and effectiveness evaluation of preventive and restorative techniques. The EM area has not been evaluated for this study, as it is not an area of national security competition.

Forces Modeling and Simulation41. The goal of forces modeling and simulation is to produce faster than real-time simulations in support of operational planning and training. Combined with sophisticated sensor processing systems and high-speed communications, FMS will support development of battlefield decision support systems. Command and control are required to direct military operations, supported by communications and intelligence. FMS applications often require real-time processing.

Signal and Image Processing are used to organize and process raw sensor data (e.g., radar returns, satellite imagery) to produce useful information for target detection, identification, and tracking. Advanced signal processing techniques also facilitate high-speed, high-quality digital communications. Signal processing, and increasingly image processing as well, is often performed by special-purpose devices and processors in embedded, deployable (i.e., man-pack, airborne, mobile) systems. SIP applications often require real-time processing.

Database Activities. DT&E activities maintain and use very large relational databases of historical test data, requiring extensive on- and near-line storage

capacity. Efficient use of these data requires powerful servers and efficient communications links for high-speed data retrieval by remote users.

Real-Time Data Acquisition computational requirements include the ability to accept data at the rate being generated by the test under observation and the ability to process the data in a useful time frame. Remote systems and networks are generally incapable of supporting the large volume of data involved in developmental tests, due to limited data transmission speeds. The time scales of RTDA applications range from milliseconds to a few seconds.

Real-Time Modeling and Simulation is functionally similar to RTDA, but is computationally more stressful in that one or more of the activity's aspects are simulated. RTMS requires the generation and real-time manipulation of environmental or threat scenarios. Three types of simulations are conducted: "live," wherein only the environment is synthetic; "virtual," in which only the operator is real; and "constructive," where all aspects are artificially generated and monitored.

Test Analysis. Extensive computer modeling capabilities are used to support analysis before, during, and after tests. While turn-around time requirements vary, the shorter turnaround times facilitated by more powerful computers enable the conduct of more test runs and more detailed analysis.

Cryptology is the mathematical science which includes cryptography, the protection of communications using codes, and cryptoanalysis, the deciphering of encoded messages for which one does not have the key. Both involve the use of complicated mathematical algorithms and very large numbers, necessitating the employment of computers for efficient functioning.

Not every computational technology area defines a set of programs or missions of critical national security concern by the mere fact of pursuit by the U.S. defense establishment. Although every program contributes to national security, quite a few are rather prosaic and many are also in commercial use. While we have not attempted to prioritize or judge the national security criticality of specific programs, we have omitted from our research those programs, such as environmental monitoring, that are obviously not of export control concern. The sections that follow describe specific programs and missions that demonstrate the importance of HPC to our national security, and suggest the types of threats to our security that might be enabled or facilitated by the uncontrolled or insufficiently controlled and monitored proliferation of high powered computational capability.

The Collection of Data About National Security HPC Programs

About 700 different Department of Defense (DoD) HPC applications were reviewed in the course of this study. Information was collected on a relatively few applications through personal visits and interviews with project teams and lead engineers. The bulk of the data on scientific and technical (S&T) and developmental test and evaluation (DT&E) projects was derived from the databases compiled by the DoD High-Performance Computer Modernization

Office (HPCMO) in support of acquisition programs. There are important differences in these data.

The goal of reviewing as many applications as possible was to capture and illustrate the wide variety of requirements. Detailed examinations were required to determine the minimum computational requirements for as many applications as possible. However, most of the data that were readily obtainable reported current capabilities and future requirements. Personal interviews, sometime by telephone but usually in meetings, were necessary to derive values for minimum computational requirements. Time constraints-about three months were available for the entire study-were such that personal interviews could be conducted for only a small fraction of the applications.

Personal interviewing is a time-consuming process, not least of all because of the unique nature of this study. The question posed-"What is the least computational power that would be sufficient to execute your program?"-represents a significant departure from the traditional way people consider their computing resources. To a program manager, scientist, or engineer, today's computer almost always seems barely functional, and access to more powerful computing resources is continually sought. Additionally, most programs rapidly evolve to outgrow each new computing resource shortly after it becomes available. A more powerful computer is useful for running today's program more quickly, but it also allows the posing of more complex and detailed problems, resulting ultimately in an apparent level of performance inferior to that of the machine recently replaced. Thinking about the inverse situation, to estimate the least powerful configuration that would provide some minimal functionality, is more difficult, and the resultant answers may be rather subjective.

The information obtained from the HPCMO database was more empirical, but not optimal for the purposes of this study. It was, however, a good starting point in defining the critical national security computing requirements and in gaining a general understanding of the role of HPC in the Defense Department. Figure 8 depicts the number of S&T applications being run on machines at various performance levels. Figure 9 illustrates the current and projected performance requirements for IDT&E applications.

Figure 8. [Omitted] Performance Distribution of S&T Applications (1994) 59

Another problem with data collection and evaluation is the lack of a uniform metric for evaluating system performance. MIPS, MOPS, Mflops, and Mtops have all been widely used to reflect a system's performance. All have weaknesses. MIPS-millions of (fixed-point) instructions per second-is particularly problematic because different vendors use the same term differently. Some use the rate at which instructions are issued, not computed; other use the fastest instruction execution rate; still others use a weighting of the execution time of an assortment of different instructions. MOPS-Millions of Operations Per Second-is also problematic, because the relationship between operations and instructions varies widely. In some machines, an instruction is equivalent to an operation; in other

machines, several instructions may need to be executed to perform an operation, or one instruction may perform several operations. Mflops-millions of floating-point instructions per second-is widely used within high-performance computing circles to reflect a system's theoretical peak floating-point capability. This measure does not, however, reflect differences in word length between systems, nor does it reflect the capability of performing operations on fixed-point data. Mtops-millions of theoretical operations per second-was developed in part to overcome some of the inconsistencies and omissions of these other metrics, but is not widely used outside the export control community. None of the metrics reflect a system's differing capabilities in handling different classes of applications. Most of the data derived from HPC practitioners was in the form of Mflops. Wherever possible, the specific computer configuration was identified and its performance rating in Mtops ascertained. When this was not possible, an estimate of the equivalent Mtops was made.

The most important conclusion drawn from a review of the data was that the computational requirements for most of these programs fall well below the uncontrollability level; many are lower than current export control thresholds. The trend in DoD computing has been to acquire the most powerful systems available at any given time, within budgetary limits. As budgets have been squeezed, older systems have been retained for longer periods of time while fewer new systems have come on line. Coupled with the trend toward distributed computing on high-powered workstations, the result is that most of today's DoD HPC applications are being performed on relatively low-power machines.

Applications being run on the most powerful machines tend to be those whose criticality to national defense justifies the higher level of investment and those that absolutely cannot, usually for architectural reasons, be executed on distributed parallel systems, such as networks or clusters of smaller computers. A number of these applications were examined to see if their proliferation would present a national security threat and, if so, whether they defined a reasonable control threshold.

Figure 9. [Omitted] Performance Distribution of Current (1995) and Projected (1996) DT&E Applications 60

Additionally, estimates of future requirements, also contained in the HPCMO database, provided insight into computational trends. A large segment of DoD high-performance computing is migrating to small computers through the process of code conversion and "parallelizing" to take advantage of clustered and networked computers. Most of the applications amenable to parallelizing are loosely coupled applications currently being executed on massively parallel processor (MPP) computers. Some applications historically run on vector processors, however, also appear "parallelizable." There are important limits to the utility of parallel computing, and this is also reflected in the future requirements data. The processing demands of many projects will continue to grow, as the applications become more complex in response to the availability of more powerful computers. Some problems, such as tactical weather prediction, do

not parallelize well and will continue to require very high-performance computers. Other applications, such as imbedded computing in sensor systems, are subject to size, weight, and power consumption constraints that preclude the use of clustered or networked systems.

HPC Mission Areas

National security HPC applications can be categorized according to four broad groups for analytic purposes: nuclear weapons programs, cryptology, conventional weapons programs, and military operations.

The first two-nuclear and cryptologic programs-were traditionally used as the justification for the definition and control of high-performance computers. Indeed, during the Cold War, control of the proliferation of these capabilities, not only through the control of computational resources, was a bedrock of nuclear and missile anti-proliferation policy. At the time, computing power was a critical element in establishing effective advanced nuclear weapons programs and the conduct of cryptologic operations. The required computational power was a sophisticated tool in the hands of a few sophisticated adversaries. This situation has changed significantly, and today neither of these applications can be said to represent the most demanding requirements for computational resources.

The next applications area-conventional weapons programs-represents today's "bread and butter" of high-performance computing applications in the U.S. national security community. The design and development of advanced conventional weapons (ACW) has developed a symbiosis with high-performance computing: programs are often defined on the basis of the current or projected state of the art in HPC, and new computer hardware and software are frequently developed in response to program requirements, both with government funding and as independent research and development (IR&D) initiatives.

Finally, the use of HPC in direct support of military operations is a growing application area, driven by the dual development of small, powerful computers and embedded processors (i.e., highly mobile high-performance computing) and doctrinal requirements for increased use of a wide variety and large volume of data for the conduct of military operations. While there exist today few applications that require the use of very high-performance computers, numerous applications under development require or could make efficient use of very powerful computers. Concomitantly, the availability of similar resources to a capable country of national security concern could significantly degrade U.S. military operations.

The following sections will examine representative programs in each of the four applications areas, with a view toward identifying the requirements for HPC and describing the conditions that render HPC no longer critical. This examination is neither comprehensive nor exhaustive. A longer term study to rigorously review every identifiable application or program is required to support the definition of a viable control regime for the future. In the context of the current study and control regime, however, a more exhaustive examination is unlikely to alter significantly the key findings.

Nuclear Weapons Programs

The delay of nuclear weapons proliferation has long been one of the most important reasons cited for maintaining controls on the export of powerful computers. The "common knowledge" was that enormous computing power was required to design nuclear weapons. This is not and never was true. 61 Basic nuclear weapons design can be accomplished on a personal computer, which is significantly more powerful than the resources available to assist in the design of the first American nuclear weapons. "Fairly robust" nuclear weapons simulations can be executed on workstations in the 1,400 Mtops range when operating in a dedicated mode. 62

However, while the control of HPC will not affect fundamentally the proliferation of nuclear weapons to non-nuclear states and supra-national groups (such as terrorists), continued export controls will slow the exacerbation of existing nuclear threats. Control of HPC exports, by limiting those exports or imposing appropriate safeguards, to countries known to possess nuclear weapons will impede their development of improved weapons and reduce their confidence in their existing stockpile by limiting the opportunity to conduct simulations in lieu of live tests. Similar or more rigorous controls on HPC exports to countries with nuclear weapons development programs could impede their development of second-generation weapons.

The world's first nuclear weapons were designed with the assistance of mechanical calculators, as computers did not yet exist. This was sufficient, however, for the successful design of both first-generation gun-assembled and implosion weapons. This feat could be replicated by people with the appropriate expertise, again without the use of computers, although the calculations would be greatly facilitated by the use of a commercially common personal computer.

Live testing has been critical to the U.S. nuclear weapons program, and vast amounts of data have been collected. In fact, the availability of data from full- and limited-scale nuclear tests is more crucial than the availability of HPC. Computer models were partially based on test data, and as more data from nuclear detonation tests was acquired they were refined and expanded. The U.S. nuclear weapons program is a synergy between experimental and theoretical programs and-was one of the pioneers in the area of Computational Fluid Dynamics (CFD).

Computers that can execute nuclear weapons design codes based on the most fundamental physical representations do not yet exist. The advancement of a nuclear weapons program beyond basic weapons design requires both computational horsepower and empirical test data.

Key judgments-nuclear weapons programs

*First-generation nuclear weapons can be designed using systems below 1,500 Mtops.

*Second- and later-generation nuclear weapons design requires using computers of at least 1,500 Mtops and conducting tests to provide data for empirical model development. As pointed out in Chapter 3, computing power below 4-5,000 Mtops is no longer controllable. However, without nuclear test data and the resulting empirical models, computers at this level of performance-indeed, at any currently available level of performance-are likely to be insufficient to design such weapons.

*Confidence in existing stockpiles will be difficult to maintain in the absence of nuclear testing. Confidence erosion can be mitigated by extensive modeling and simulation, requiring the most powerful computers available. However, other countries can use different safety measures that could be simpler and require far less HPC support.

Cryptology

Cryptology is the second field that provided justification for controlling the export of HPC during the Cold War. At one time the exclusive province of HPC, today significant cryptologic capabilities can be achieved through the use of computers widely available commercially, eliminating many cryptologic applications as a Justification for continued HPC controls.

The two principal uses of computers in cryptology are for cryptoanalysis, the decoding of messages by an unintended recipient who does not have the key, and cryptography, the design and use of encipherment systems. Some of the software programs that implement these applications are loosely coupled and can be readily adapted for parallel processing environments.

A common-and the most stressful-application for computers in cryptoanalysis is the "brute force" attack, which involves testing vast numbers of potential combinations in order to discover the key to decrypt a message. "A brute force attack is tailor-made for parallel processors," since each processor, whether it is a single CPU in a multi-processor computer or a separate computer in a cluster or network, can be set to work on only a portion of the keyspace without reference to the activities of the other processors. 63 Thus, a country of national security concern with limited means (but also limited goals) could achieve significant successes in a narrow area, such as attempting to decrypt messages from one cipher system of one foreign country.

Keyjudgments-cryptology

*Cryptologic applications can be readily adapted for parallel processing.

*Significant cryptologic capabilities can be achieved through the use of widely available computer equipment, such as clustered or networked workstations or simple, massively parallel processors.

*Cryptologic applications can no longer be used as a basis for establishing an export control regime or defining a control threshold.

Advanced Conventional Weapons Programs

Advanced conventional weapons (ACW) programs involve research and development, test and evaluation, and transition to production of new weapons and weapons platforms. Significant programs in this area include low radar cross-section-"stealth"-aircraft, enhanced sensor systems to counter "stealth" and other low-probability-of-detection targets and to extend the battle force's surveillance range, and enhancing the lethality of conventional weapons, such as cruise missiles and bombs, by increasing the precision of their delivery. Table 8 summarizes the major mission areas examined for this study.

Table 8. [Omitted] ACW Functional Areas Aerodynamic vehicle design

This field involves the research and development to produce fixed and rotary wing aircraft and cruise and ballistic missiles. Programs in this functional area include the design and evaluation applications listed in Table 9. (The Computational Technology Areas-CTAs-are described in Table 6.)

Design Applications
Airfoils (wings) and airframe
Airframe structure
Signature reduction

Engines (turbines)
Rocket motors

Computational Technology Area
Computational Fluid Dynamics
Computational Structural Mechanics
Computational Fluid Dynamics, Computational Electromagnetics and Acoustics
Computational Fluid Dynamics
Computational Chemistry and Materials Science

Table 9. Aerodynamic Vehicle Design Functions

Threat

The threat to U.S. national security from the proliferation of significant capabilities in this field or the diminution of this country's technological lead comes from the production of more highly capable combat aircraft and deadlier missiles on the part of countries of national security concern. The most significant threat enabled or facilitated by possession of HPC capabilities would be from high-performance fighter aircraft with increased speed and improved maneuverability and controllability and cruise missiles with longer ranges, greater maneuverability, and increased speed. While the most significant potential threat to U.S. forces

might be the development of stealthy missiles and aircraft, computational capability is important but not a critical limiting factor for these programs.

Countermeasures

Were one or more countries of national security concern to acquire a sufficiently high-powered computational capability to pursue advanced aircraft and missile designs, U.S. forces would require countermeasures that are difficult and expensive to implement. Such countermeasures would include further enhancement of our own aerodynamic weapons systems and improved surveillance capabilities. Although HPC is not the limiting factor in the production of stealth aircraft or cruise missiles, introduction of either of these, even in small numbers, by a country of national security concern could have a serious negative impact on U.S. national security.

Typical applications

Design of a stealth aircraft is one of the threats often cited that could arise from the unrestricted availability of HPC. However, depending upon the design goals, HPC may not be necessary at all. Additionally, the availability of computational resources is not one of the most critical features of a stealth aircraft program; other problems, such as the use of advanced composite materials, represent a much greater concern for countries of national security concern. There is a myth that the F-117A is faceted because the computers available at the time of its design lacked sufficient power to optimize simultaneously the CEA equations for the aircraft's stealth and the CFD for the aerodynamic characteristics. The reason for the F-117A's faceted appearance is related to the electromagnetic properties of radar signal propagation in the frequency range of the radars to be avoided. Designed in the 1978-1980 time frame, the F-117A program did not employ the most powerful computers available. The aircraft was designed using an IBM 3090/250 mainframe (189 Mtops), but senior Lockheed computational scientists estimated that a DEC VAX-11/780 (0.8 Mtops) would have just met their requirements. 64

Similarly, the B-2's more refined appearance is unrelated to the performance of the computers used to design it. The B-2 is optimized to avoid detection by radars in a frequency range different from those to be avoided by the F-117A, which allowed for a more blended shape. The frequency range considered for the B-2 design not only changed the plane's appearance, but increased the computational difficulty of the task. However, the solution still did not require the use of the most powerful computers available. One competing Advanced Technology Bomber (ATB) design was produced on the same 189 Mtops IBM mainframe that was used to design the F-117A, and it was estimated that this was the smallest computer that could have been effectively employed.

The computational difficulty of designing the F-22 fighter was greatly increased by the more stringent operational requirements (e.g., unlike the F-117A, which operates like a light bomber, the F-22 is intended to be an air superiority fighter) and the greater variety of threat radars to be countered. Thus, not only are the

CEA and CFD calculations more difficult, but their simultaneous optimization has required the use of the most powerful computer available for solution within reasonable time scales. The F-22 was designed using a Cray Y-MP/2 (958 Mtops). Although the IBM 3090 could probably have been used again, use of the Cray was more economical.

The requirement to use the most powerful computers available may be more closely related to program economics than to feasibility. Long processing runs leave expensive people and facilities idle, making the purchase of a very expensive high-performance computer necessary to efficiently employ all the resources available. Absent the economic considerations, effective computational support could be provided by lesser machines, although the project would take significantly longer. The design of an F-22 type aircraft, however, could not be accomplished without HPCI because of the inability of less capable equipment to process the high resolution 3-D simulations at all. Some of the design functions, such as high-frequency (> I GHz) scattering analysis, have been adapted for processing on powerful workstations, but large high-performance computers are still required for low-frequency analysis of resonance and inhomogeneous wave effects.

One of the project teams designing a candidate joint Advanced Strike Technology (JAST) multi-service aircraft is using a 150-node Intel Paragon MPP computer (4,864 Mtops), significantly less powerful than the current state of the art, which exceeds 100,000 Mtops. The project originally employed a 128-node Intel iPSC/860 (3,485 Mtops), and that machine is believed to be minimally sufficient.

The design of stealth cruise missiles is less computationally difficult because the much smaller size of a cruise missile necessitates fewer calculations. However, as in the case of the F-117A, lack of the computational ability to design such a missile is not the primary factor preventing countries of national security concern from doing so.

Flight-test processing and simulation benefit from the application of HPC, but the requirement for computational power depend upon the complexity of the testing program and the time scales involved. Because these applications are readily scalable and run well on parallel systems, aggregate computational power is more important than the individual capabilities of any single computer. Funding (to acquire sufficient computers) and schedule (to determine how many computers are required) are the limiting factors. Real-time testing and image analysis supports a variety of assessment functions, including flight characteristics, ballistic trajectory validation, and store separation (weapons release). Powerful computing resources are required by U.S. DoD programs to enable real-time processing because of the short time scales of the events being observed or modeled and the vast amount of data to be processed. Less capable computers limit the number and accuracy of the variables that can be processed. For example, trajectory image analysis processing can be run on a six-node cluster of VAX-8600 mini-computers (about 6 Mtops) with very constrained results. However, one such project is now beginning to make use of a Cray T3D (3,439 Mtops), which enables the processing of a far greater number of sensor inputs in real time.

Store separation simulation is being conducted to determine which missiles currently in the inventory could be adapted for use with the F/A-18. As with most CFD applications, memory size is often more critical than processor performance. Computers in use range from the Silicon Graphics PowerChallenge (1,153 Mtops) and Power0nyx (2,124 Mtops) systems to the Cray C916 (21,125 Mtops) vector processor and a 352-node Intel Paragon (about 10,000 Mtops) massively parallel processor.

Submarine design

Submarine design programs emphasize enhancing the submarine's primary advantage: stealthiness, and improving its survivability should it be detected and attacked. Most extant programs involve the reduction of the acoustic and electromagnetic signatures of the platform and its weapons through the modeling of improved shapes and use of advanced composite materials. Programs within this functional area include the design and evaluation applications listed in Table 10.

Table 10. [Omitted] Submarine Design Functions

Threat

The threat to U.S. national security from the proliferation of significant capabilities in this field or the diminution of this country's technological lead would be from the resultant erosion of the U.S. combat advantage should a country of national security concern field submarines that are significantly more difficult to detect. This would present a direct threat to U.S. naval forces and operations. A country of national security concern might also be able to improve the survivability of its submarines, thus reducing the navy's ability to destroy them with conventional weapons once detected. The sum of these effects would be to severely limit the navy's ability to project U.S. power, especially in restricted waters (such as the Persian Gulf). However, possession of advanced computational capabilities is insufficient by itself to fully enable these threats. Concomitant developments in advanced materials and numerically controlled machine tools are also required.

Countermeasures

Were one or more countries of national security concern to acquire a sufficiently high-powered computational capability to pursue the design of an advanced submarine platform or weapon system, the U.S. would be required to refine further its own tactical sensor systems. This process would be uncertain, difficult, and expensive. The design of a more survivable submarine on the part of a country of national security concern would require countervailing improvements in U.S. submarine weapons technology or a renewed reliance on tactical nuclear weapons.

Typical applications

Typical design applications in use in the United States involve simulation of acoustic transmission through structures and in water. HPC is required for resolving equations to a useful level of detail in a reasonable time frame. One project is using a Cray C916 (21,125 Mtops) to run CSM programs and simulations. The applications are not readily parallelizable, however, so there has been only limited use of parallel processors. The efficiency of an MPP computer running a converted application is very low. Since a typical run takes 10-20 hours on the Cray, and must be repeated at least 2,000 times, there is little chance that a country of national security concern could replicate this program with computers not subject to export controls.

The increased requirement in the post-Cold War world for submarine operations in relatively shallow water has generated new design tasks of great difficulty. Modeling turbulent flows to determine design modifications that will decrease radiated noise levels requires vectorized processing and very large memories. A typical task, executed on a Cray C916 (21,125 Mtops), requires at least 128 million 64-bit words of memory (the maximum configuration used has 512 million words of memory available). This application cannot presently be converted to run on parallel systems. The only system currently capable of adequately executing this application is a 16-node Cray.

Surveillance and target detection and recognition

Research and design programs in these fields involve the production of more capable sensor systems enhanced by improvements in basic detection technology and especially target signal processing. All these programs are computationally intensive in their design phases. Many of the resultant systems also require the use of HPC for their operation. Programs examined in this section include the design and evaluation applications listed in Table 11.

Table 11. [Omitted] Surveillance Design Functions

Threat

The threat to U.S. national security from the proliferation of significant capabilities in this field or the diminution of this country's technological lead arises principally from the loss of tactical surprise, but also from a routine loss of operational flexibility on the part of units on patrol, conducting reconnaissance, or pre-positioning in anticipation of hostilities. Stealth has long been the hallmark of submarine operations, and is becoming increasingly important in air and land operations as well. Any foreign developments that increase the protectability of these platforms would negate both a significant investment and tactical advantage, and place U.S. military forces at greater risk. Improved surveillance capabilities, especially wide-area and long-range surveillance, on the part of a country of national security concern would reduce the potential for achieving tactical surprise through force maneuvering (such as in Desert Storm). Development of remote and/or fine-grained sensors on the part of countries of national security concern

would jeopardize the security of RDT&E programs and friendly infrastructure by enabling better enemy observation and targeting.

Countermeasures

Were one or more countries of national security concern to acquire a sufficiently high-powered computational capability to pursue the development of advanced sensors, these developments could not be readily negated by U.S. counter-developments. Significant improvements in the ability to detect and identify U.S. weapons platforms on the part of a country of national security concern might be mitigated by modifications to operational profiles, but would more likely necessitate costly design changes. For example, the threat from an improved long-range search radar might be offset by flying at a different altitude or maintaining a longer stand-off range, but each of these affects other aspects of military operations. It would be desirable, although more difficult and expensive, to maintain the optimal operational profile and mitigate the threat by reducing the aircraft's radar cross section.

Typical applications

Automatic target recognition and radar signature prediction applications seek to reduce on-board processing requirements through the development of target recognition templates and to provide accurate predictions of sensor performance under various operating conditions. Algorithms are being developed that will provide accurate performance predictions of the capabilities of radar systems to identify targets in the presence of ground clutter or jamming. The development of templates, including those for potential "stealth" and other low radar cross-section targets, and prediction algorithm processing are very computationally intensive, with processing times of up to several thousand hours on computer systems rated at 24,000 Mtops and higher. Very high-powered computers are required to enable the simultaneous solution of numerous sample sets. Performance increments permit more simultaneous solutions, yielding more accurate templates. These applications currently employ parallel processing and could be converted to execute on very large clusters of powerful workstations.

Research and development projects to advance the state of the art of acoustic sensors require extensive modeling and simulation of various environments as well as algorithm development for sensor signal processing. These applications typically employ the most powerful vector computers available (e.g., a Cray C916 at 21,125 Mtops), although some of the code is being converted for execution on MPP machines (e.g., a Cray T3D at 10,056 Mtops). Typical applications, such as large finite-element analysis or 2-D ocean acoustic models, require large amounts of closely coupled, high-speed, 64-bit word memory, making them unsuitable for conversion to run on clusters or networks. Bottom contour modeling for shallow water acoustic problems requires an absolute minimum of 8,000-9,600 Mtops of processing power to execute. U.S. Navy acoustic processing applications are classified, to prevent the use of the code by countries of national security concern, even they should possess the requisite hardware.

Non-acoustic anti-submarine warfare (NAASW) sensor development protects are also computationally intensive. The most difficult phase is establishing the basic physics of advanced, non-traditional sensors and integrating them into a system. Developmental code is being executed on a 64-128-node Intel Paragon (approximately 2,000-4,600 Mtops); a typical task executes overnight. These tasks could be converted to run on workstation clusters, but the resultant increase in execution time (up to two weeks) and decrease in accuracy are deemed unacceptable by U.S. standards. Once fully developed, the deployed sensor suite is expected to require only about 500 Mtops of computing power.

Cartography is an increasingly computationally intensive area, as more maps and charts are created in or converted to digital format. The development of computer-aided topography has also led to the development of mapping applications in embedded systems (e.g., navigation suites). However, cartography is generally not time-constrained, which results in the use of economically feasible rather than the most operationally desirable computers. The Navy is developing a capability to rapidly produce digital topographic maps based on synthetic aperture radar (SAR) imagery. The system is
being developed on the same Paragon computer (2,000-4,600 Mtops) as the NAASW applications. To process the TOPSAR data in near-real time for combat support, a minimum of 8,000 Mtops and possibly as much as 24,000 Mtops of computing power will be needed.

Survivability, protective structures, and weapons lethality

Deep underground and other very hard shelters are important and difficult targets. The development and testing of advanced algorithms for simulation of weapons effects against hard targets is required for development of better protected structures (hardened targets) and for design of weapons for use against them. The same applications code is also used for development of advanced vehicle armor and improved armor-penetrating weapons. Improvements in anti-tank weaponry have placed a premium on advanced armor design. High-performance computing is critical to the efficient processing of these applications. New technologies, such as high-power mobile-laser development, have generated new requirements for weapons effects modeling. Programs within this functional area include the design and evaluation applications listed in Table 12.

Table 12. [Omitted] Survivability and Weapons Design Functions

Threat

The threat to U.S. national security from the proliferation of significant capabilities in this field or the diminution of this country's technological lead is from the development by countries of national security concern of improved armor, enhancing their combat capabilities, or anti-armor weapons of increased lethality, placing U.S. combat forces at greater risk. Deep underground shelters, particularly relevant to military operations in the developing world, could be locally designed or improved, complicating the collection of intelligence about

these structures and possibly negating U.S. developments in deep penetration weapons.

Countermeasures

Were one or more countries of national security concern to acquire a sufficiently high-powered computational capability to pursue the design of enhanced anti-armor weapons, U.S. forces could mitigate the effects through modified tactics. However, this would entail a loss of operational flexibility that could only be restored by concomitant improvements in U.S. armor systems. Some less costly countermeasures, such as thermal decoys, might be effective, depending upon the exact nature of the threat. The fielding by a country of national security concern of more survivable armed combat vehicles might be countered by modified tactics (e.g., firing from shorter ranges), but most likely would require the development of more lethal anti-armor weapons.

Typical applications

Modeling of the interaction between warheads and defensive structures is very computationally intensive, due both to the complexity of the code and variables, and to the requirement for multiple iterations. The complexity of the models increases significantly under certain scenarios, such as high attack angles or very high impact speeds. For example, 3-D modeling of a symmetric, transonic, low angle of attack situation can be modeled on a Cray Model 2 (1,098 Mtops) in two hours, but a full (i.e., asymmetric) model requires 40 hours. A typical penetration model against advanced armor, using the same computer, requires approximately 200 hours per run, while the modeling of kinetic kill effects against hybrid armors can take up to 2,000 hours. Full optimization analysis requires up to 14,000 hours of run time (multiple iterations) for each candidate armor type. Modeling the effects on a complex structure (as opposed to an armor plate) can take up to several hundred hours per iteration when run on a Cray C916 (21,125 Mtops).

The development and evaluation of deep penetration weapons requires multiple iterations of 3-D finite-element analyses, and non-linear equations which become very complex due to the high pressures and short time scales of the events being modeled, coupled with the requirement for high spatial resolution. The problem is similar to hybrid armor modeling, since the different strata to be penetrated have unique properties, and the coupling between the layers is complex. Current research is being conducted using a Cray C916 (21,125 Mtops).

Nuclear blast simulation for the development of protective structures is fundamentally different from,-and significantly less complex than, nuclear detonation simulations used for weapons development and testing. These simulations involve 2- and 3-D finite-volume-flow, finite-difference, and fluid-motion algorithms to model the effects of nuclear blast fronts on stationary and moving objects (such as shelters or vehicles). Two-dimensional modeling of one or two seconds of nuclear blast requires about 200 hours on a Cray C916 (21,125 Mtops), and a 3-D model requires about 600 hours. The algorithms are being

adapted for execution on massively parallel HPC systems such as a Cray T3D (10,056 Mtops) and a Thinking Machines CM-5 (10,457 Mtops).

The Smart Munitions Test Suite program is developing a system that will provide a simulated combat environment on a test range and collect and analyze weapons testing data in real time. The most difficult aspects of this application are the data communications and image processing requirements. Data communications are handled by double-wide HIPPI bus interfaces that accept data input at 70 MHz. Image processing is handled by a 128-node partition of a Thinking Machines CM-5 rated at 5,194 Mtops, which is being up-graded to a 14,410 Mtops CM-5 to provide additional realism in the visual presentations.

Key judgments-advanced conventional weapons programs

*While HPC is clearly of great importance to critical research and development projects, no clear lower performance bound is apparent. This is due in part to the time dependence of computational resources. Economic considerations (i.e., cost of idle assets) aside, most programs can be executed on less capable equipment if the executor is not bound by a tight schedule.

* The great variability in performance requirements is also due to the emergence of powerful workstations and high speed networking, which have accelerated the trend toward parallel processing on smaller computers. Indeed, the U.S. national security community is making ever-greater use of consumer-grade computers for sophisticated applications.

* More than two-thirds of the applications for which data are available can be carried out using computers below the threshold of controllability defined in Chapter 3. Of those remaining, about five percent require the use of computers in the 7,000-8,000 Mtops range. A smaller but still significant number of applications require the use of computers of at least 10,000 Mtops.

* There are RDT&E applications of great national security significance, the proliferation of which should be strictly controlled. Some applications, such as acoustic sensor development and associated acoustic modeling, automated target recognition algorithm, and template development, cannot be executed on computers less powerful than 20,000 Mtops with significant high-speed memory. Thus far, these applications cannot be readily adapted to run on networks or clusters of conventional computers.

Military Operations

Computers are used in direct support of military operations for real-time control of weapons systems, sensor data processing, decision support information processing, and critical support functions, such as weather prediction. The migration of these functions to computer systems from special-design processors has enabled the integration of functions into battle management systems. These applications are of great importance in that people's lives and military missions are immediately at stake and processing must occur in real-time. The prevention of significant

developments in this area on the part of countries of national security concern is of high priority to U.S. national security. Table 13 summarizes the major mission areas examined for this study.

Table 13. [Omitted] Military Operations Functional Areas

Direct support to military operations is the fastest-growing area for national security high-performance computing. This is principally the result of rapid advances in the state of the art that have greatly increased computer performance while simultaneously reducing the size, weight, and power requirements of very high-performance computers. Today's deployed and embedded computer systems are significantly more capable than the headquarters mainframe of ten years ago.

Although many of the applications that support military operations are inherently parallel and therefore readily adaptable to networked or clustered environment, the use of HPC is required for some fielded systems because of size, weight, and power consumption limitations. Additionally, interconnect bandwidth may limit the ability of clustered or networked computers to efficiently process the large volumes of data inherent in many of these applications.

C4 target engagement, battle management, and information warfare

Command and control is the sine qua non of military operations. Military operations are characterized by broad scope, fast pace, multiple threats, and large, multi-disciplinary forces. Effective acquisition and assimilation of threat data and efficient force management require high through-put and reliable data processing and communications. High-performance computers are critical for some of these applications. Provision of sensor data, especially from national technical means, directly to tactical units requires exceptionally efficient data processing and communications.

The impetus behind current development efforts is the need to be able to operate within the adversary's "decision cycle." That is, U.S. forces must collect, understand, and react to sufficient information more rapidly than can the enemy. Data collection includes the acquisition of information about the enemy (intelligence), the environment (third parties, geography, weather), and friendly forces. Understanding the data-turning the data into useful information-includes data processing and database activities, and the presentation of the information in useful formats. Reaction involves making and communicating decisions, whereupon the cycle begins anew (with collection of feedback data from friendly forces and observation of enemy reactions). This entire process is highly reliant, upon efficient communications, which are also facilitated by computers.

The new emphasis on information warfare (IW), however defined, also places a premium on the number and performance of fielded computer systems and their interconnectivity. Whether processing friendly data or intelligently manipulating an adversary's data processing, information warfare will require significant computing power. It is likely, however, that a large number of efficiently

networked workstations will prove more useful for many IW tasks than a few HPC installations.

It is important to note that, in this field, developers tend to work with what they can get, rather than what might be truly optimal. The computers used tend to be small relative to the size of the problem. Extensive use of commercial off-the-shelf (COTS) equipment is dictated by budgets and the requirement for reduced development cycles. Thus, most of the extant applications are executed on computers below the uncontrollability curve.

Another unique feature of the communications-intensive applications in this area is that the critical performance elements are not the hardware, but software and network architecture. These features are not, however, accounted for by the conventional measures of computer performance.

Threat

The threat to U.S. national security from the proliferation of significant capabilities in battle management or the diminution of this country's technological lead is from the potential for a country of national security concern to develop the capability to prosecute military operations more rapidly than our ability to react. The development of capabilities for managing diverse forces over large areas would present a more complicated threat to U.S. military forces. Improvements in this area would be most marked in less developed countries, the adversaries most often encountered today. Because of its heavy -reliance on information systems and networks, the U.S.-both military forces and the "rear"-is vulnerable to information warfare attacks.

Countermeasures

Were one or more countries of national security concern to acquire a sufficiently high-powered computational capability to achieve significant improvements in their C41 and battle management, these developments could be countered by concomitant improvements in U.S. battle management, revised tactics, and improving the security of friendly information systems. Regardless of the efficacy of such countermeasures, however, the pace of military operations would be intensified and the time available to make critical decisions would be significantly decreased, circumscribing the available options. The efficacy of battle management systems depends as much, if not more, on doctrine as the system itself, and the hardware-computers in this case-is not necessarily the most critical element.

Typical applications

Shipboard infrared search and tracking (SIRST) systems are being developed to detect anti-ship and other cruise missiles. Currently available sensors, principally radar, cannot be relied upon to detect or track ultra-high-performance cruise missiles such as the Russian "Sunburn," which skims the water's surface at high speed while rapidly maneuvering. The algorithms for use in the sensor's processing system are being developed on a 328-node Intel Paragon (8,980 Mtops). The

deployed system is likely to require a computer capable of delivering about 6,500 Mflops of sustained computational power (about 13,000 Mtops) for real-time operations; a Mercury parallel processing computer of "only" about 7,400 Mtops might be minimally sufficient. This application is readily parallelizable; however, the ability to do so is constrained by the memory and input/output intensity of the processing, and ultimately by size, weight, and power consumption constraints for deployable configurations.

In addition to infrared optical sensors, visible light sensor systems are also under development. Development is currently being carried out on a 24,000 Mtops HPC, and it is currently estimated that the deployed sensor processing systems will require similar computing power, although in a smaller, lighter form. While the development work can be converted to run on a cluster or network of lower performance computers, the deployed systems will be constrained by physical limitations.

Integrated systems to provide communications, database functions and data fusion, and decision support processing and display are being developed to enhance a combat commander's ability to make higher quality (more well-informed) decisions in shorter time frames. Such an integrated battle management system requires exceptional computational power; however, this power can be efficiently provided by distributed computer systems. Battle management functions are also readily scalable, making them suitable for initial implementation on readily available commercial equipment with a view toward upgrading with more and/or more powerful computers as funding becomes available. The critical level of technology for systems currently in use is represented by IBM SP2 and Silicon Graphics PowerChallenge workstations with performance capabilities in the 100-1,000 Mtops range. Combat direction systems on mobile platforms, such as ships, submarines, and aircraft, are also expected to take advantage of the scalability of battle management functions to an extent feasible within size, weight, and power consumption constraints. The F-22 avionics suite will execute about 1.6 million lines of code on a pair of computers with CTPs of about 9,000 Mtops. The AN/BSY-2 Submarine Combat System developed for the Seawolf attack submarine, for instance, executes over five million lines of code on a network of more than 100 embedded Motorola processors in various sensors, control, and display systems. 65

Modeling and simulation, long the domain of RDT&E and headquarters organizations due to the cost and complexity of the resources required, are increasingly being integrated into fielded decision-support systems. Realistic, real-time simulation of the actual or expected order of battle in any given scenario can be performed using current computing and display technology. The goal is to be able to apply simulation technology to enhance the ability of tactical units to assess an immediate battlefield situation and examine alternative responses. Applications programs include interactive simulation of battlefield situations in real time, including accounting for the effects of obscurants (countermeasures such as smoke) and weather-related visibility diminution. The simulation programs are being developed on workstations, but the simulations themselves are to be executed on remote MPP computers with performance ratings in excess of 8,000 Mtops. The

performance requirements for future fielded versions have not been determined, but will likely be well above the 1,000 Mtops range of current battle management processors. With the increasing sophistication of computer simulations and cost of live training, HPC is increasingly used for military training. However, most of these applications are executed in a distributed fashion on uncontrollable computer systems.

The Attack and Launch Early Reporting to Theater (ALERT) system is being developed by the U.S. Air Force to rapidly process and route missile launch detection data from Defense Support Program (DSP) satellites and other sensors to theater commanders as an outgrowth of problems encountered in trying to find and destroy Iraqi SCUD missile launchers and defend against SCUD missiles in flight. The system comprises a central processing suite of three Silicon Graphics Onyx servers (1,700 Mtops) and 14 networked Onyx workstations (300 Mtops). 66

As demonstrated during Desert Storm, switching is the bottleneck in telecommunications networks, Modern switches, especially in data and record communications networks, are computers. Faster switches provide higher information through-put and therefore a more capable communications infrastructure. A highly capable communications network does not necessarily require high-performance computers. An appropriate architecture and efficient software are much more critical to system performance than raw computing power. During Desert Shield/Storm, most of the communications architecture was implemented on Sun SPARCstation 4/300 (20.8 Mtops) and SPARCstation 10 workstations (53.3 Mtops). Although the network proved inadequate for operational requirements in late 1990, by the time the ground attack was launched in February 1991, the network was operating efficiently. No hardware was upgraded, however, the entire performance enhancement was due to software improvements, Emerging special applications, such as adaptive routing and video conferencing, will generate requirements for higher performance computers in the communications architecture.

Meteorology

Weather forecasting is not generally considered in reviews of programs of national security concern, but the fact is that accurate weather prediction is of critical importance to a wide array of military activities, from intelligence collection to tactical operations to strategic planning. The anticipated weather was one of the most important planning factors for the D-Day assault on Normandy in World War II, and the adverse weather that was encountered complicated the operation. Nagasaki was destroyed by an atomic bomb because the initial intended target was obscured by clouds.

More recently, the timing of the launch of Desert Storm in the Persian Gulf was significantly affected by meteorological conditions. Tactical weather prediction is critical for planning and conducting military operations, especially air operations, airborne assault, amphibious operations, and high-precision weapons employment. Clearly, the side with the best understanding of the weather-the longest range,

finest grained, and most accurate forecasts-has significant advantages in initiative, planning, flexibility, and conduct of military operations.

A problem with analyzing meteorology as a national security application is that the basic techniques for military weather forecasting are the same as for civil weather prediction. The principal differences are the consumer, the level of detail required, and the potential consequences of inaccuracies. Unique features of military weather forecasting include the availability of data from in unique sensors and the requirement to be able to provide highly accurate, long-range (5-10 days), fine-grained (less than 50 km resolution) forecasts for any area of the world on short notice.

Threat

The threat to U.S. national security from the proliferation of significant meteorological capabilities or the diminution of this country's technological lead is from the potential for countries of national security concern to improve their intelligence collection operations and, in the event of hostilities, military planning.

Countermeasures

There are no feasible countermeasures against improved weather prediction on the part of a country of national security concern. The effects would have to be ameliorated by altering operations plans to take improved enemy planning into account. This could significantly constrain the available alternative courses of action.

Typical application

Numerical weather prediction is centralized at several sites that provide world-wide meteorological support to all the military services and other U.S. government organizations. Unlike civil weather forecasting, weather prediction for the U.S. military is more concerned with detailed weather prediction over relatively small areas than with global or regional analysis, which results in the requirement for the use of powerful computers. While a typical global weather model with 120 km resolution can be executed on a workstation with performance in the 200 Mtops range, typical tactical weather models with 45 km resolution require computers rated in excess of 10,000. Calculation of weather forecasts in littoral areas to resolve complex air-ocean interactions is even more demanding.

Numerical weather prediction for all armed services is performed on 8-node Cray C90s (10,625 Mtops), which are considered barely adequate to support operational requirements. (An 8-node C90 is rated at 3,000 Mflops of sustainable performance on weather-specific benchmarks.) This computer can routinely generate, on a timely basis, regional five-day forecasts and 36-hour forecasts with 45 km resolution. Special analyses can be produced on an ad hoc basis by concentrating computer power on a limited problem set. The Navy provides special forecasts with resolutions as fine as 20 km, while the Air Force generates special forecasts of greater duration, up to seven to ten days, or with as much as 5

km resolution over a shorter period. An up-grade to a 64-node Cray T90 (well over 100,000 Mtops) would permit the routine production of ten-day forecasts with resolutions of up to 5 km, and even 1 km in some circumstances. A system is also under development that will rapidly generate I km resolution, three-hour forecasts over relatively small areas in support of chemical and biological weapons defense. This system requires a Cray C916 (21,125 Mtops).

Attempts have been made to adapt the CFD code for numerical weather prediction to parallel processing environments. One study analyzed the performance of a parallelized version of a mesoscale-cloud-scale numerical weather prediction model running on a cluster of up to 16 IBM RS6000 workstations (of an estimated approx. 40 Mtops each) connected via a 10 Mbps Ethernet LAN. For a fixed problem size, this configuration ran 4 times slower than a single-processor Cray - Y-MP (500 Mtops)and did not scale well past eight processors. Even with interconnects 100 times faster than Ethernet-, performance models predicted the cluster to have only half the performance of the Cray. 67 While clusters may be used to develop and test parallel codes, additional research is needed to determine precisely how well clusters employing high-speed networks and more advanced workstations address numerical weather prediction models. The results cited here results suggest that clusters have difficulty achieving the time-to-solution exhibited by the Cray configurations mentioned above.

Surveillance and target detection and recognition

Surveillance for intelligence collection, reconnaissance, and self-protection is one of the fastest growing application areas for HPC. Improved sensors and processing systems must manipulate a large volume of data efficiently, placing a premium on computational power, memory, and communications. The related area of target detection requires very powerful computers to enable the analysis of large arrays of data in an effort to achieve greater target detection capabilities with existing sensors. Target recognition processing is similar to the RDR&E processes for target prediction and template development, although computational requirements may be lower due to the use of templates.

High-speed digital signal processing (DSP) is critical to providing timely target detection and recognition capabilities. Certain basic functions (e.g., FFT) can be implemented in special purpose processors, but efficient coherent signal processing that can be modified to meet an evolving threat can be achieved only in software. For real-time image processing in support of reconnaissance and weapons systems, extremely high-powered computers are required due to the large amount of data to be processed and the complexity of the algorithms. High-speed, high-bandwidth DSP can be effectively incorporated in radar, acoustic, and infrared sensors. Lower performance computers may be useful for these applications, but are not capable of ensuring real-time detection of targets with weak signatures or target discrimination in high target-density environments, and are thus unsuitable for operational environments.

Threat

The threat to U.S. national security from the proliferation of significant surveillance capabilities or the diminution of this country's technological lead arises largely from the potential that a country of national security concern may greatly improve its sensor performance through computerized post-processing. However, some countries of national security concern possess the knowledge required to integrate high-performance computers into existing sensor suites to achieve improved target detection and identification capabilities. Either could result in a higher detection probability for low radar cross section and other "stealth" aerodynamic platforms (e.g., aircraft and cruise missiles.); increased capability to detect and engage small, fast targets (e.g., anti-aircraft and high-performance anti-ship missiles); and improved capability to detect and localize submarines.

Countermeasures

The achievement of enhanced sensor performance by countries of national security concern would require significant effort to negate. U.S. technology in the areas of radar signature reduction and acoustic quieting are already state-of-the-art; further improvements would be costly and uncertain of success. Decoys and masking could prevent target localization, but may confirm the presence of a U.S. platform,

Typical applications

The Theater Missile Defense Ground-Based Radar (TMDGBR) system is a deployable sensor suite employing an X-band phased-array radar to perform search, fire control, and kill assessment functions. The system currently under development requires massively parallel HPC to control the radar, detect, identify, and track targets, and compute fire control solutions for multiple high-speed targets. computers used are commercial-grade MASPAR 2264 computers; each system comprises four computers (5,152 Mtops each). This problem is not amenable to solution through employment of larger clusters of smaller computers in lieu of HPC at this time because of the very high rate and volume of data through-put required and current limitations on interconnect technology.

Signal processing for synthetic aperture radar (SAR) applications is computationally intensive. A SAR produces high-resolution images by combining coherent broadband signals, involving several computational steps. SAR signal processing has historically been implemented in specialized processors, but the deployed systems are inflexible. PVP computers offer flexibility, but lack the ability to process the data in real time. MPP computers have successfully been adapted to provide the required flexibility for tactical use while maintaining throughput. With the development of special-purpose computers for use in airborne platforms, continuous real-time processing of SAR reconnaissance images is expected to require computers in the 300+ Mtops range for sensor platforms.

Creation of a SAR image is a three-step process requiring about 500-1,500 floating-point operations per pixel. The first step requires interpolation from polar to grid coordinate systems, and parallelizes well. The second step is the performance of a 2-D FFT to create the basic image. The final step, phase gradient

autofocusing, remove-, artifacts of the platform's motion from the image and is the most computationally intensive step. Use of massively parallel computers (e.g., ncube 2 with 256-1024 nodes E 413 Mtops and 16,000-64,000 node CM-2 ((3 512-2,471 Mtops) provides the ability to produce a useful image in tens of minutes, as opposed to the hours formerly required. 68 Tests of an Intel Paragon for SAR image processing indicated that the machine could only sustain about 40-50% of its peak rate. Although the processing is highly parallel, computer clusters and networks are currently less useful because of interconnect speed limitations.

Long-range unmanned aerial vehicles (UAV) operating at high altitudes are being used as SAR platforms, and development is underway to greatly increase their on-board processing capability to permit wider area searches at greater resolutions while minimizing communications requirements. On-board data processing reduces communications requirements by a factor of 8-10. 69 Current systems, which deliver up to 5,000 Mflops of sustainable performance (i.e., up to about 10,000 Mtops 70), provide the capability to survey 1-5 km with resolutions of 0.3-1 foot. To this computationally intensive task has been added the requirement to support a variety of mission packages, such as optical imagery or signals intelligence, in addition to the SAR suite. Estimates of the upper limits of these requirements range between 100,000 and 200,000 Mflops (i.e., more than the largest machines currently available). Various systems under development include one that will employ a developmental embedded Cray Research system with an estimated performance exceeding 30,000 Mtops, and the Honeywell EPHC-10 "Touchstone" embedded version of a 64-node Intel Paragon. Given the efficiency factor of the Paragon architecture for SAR processing, a single-node "Touchstone" computer is expected to be able to process in real-time a l,OOOx2,000 pixel image strip. However, the Tier 2 UAV will be required to process data from a 1,5OOxl5O,OOO (50 km wide) strip, which will require about 50,000 Mflops of sustained performance (about 80,000 Mtops). The Tier 2+ UAV will be required to simultaneously process SAR and IR imagery, requiring an on-board MPP computer in the 10,000-12,000 Mflops range (i.e., about 16,000-20,000 Mtops).

Sensor processing and data integration suites, such as on the joint Surveillance and Target Attack Radar System (JSTARS) and Airborne Warning and Control System (AWACS) aircraft, also require powerful computers, for sensor data processing, database functions and data fusion, graphic displays and other decision support functions, and communications control. JSTARS is optimized for monitoring and control of ground operations, while AWACS provides the same functions for air combat.

Computer support in the JSTARS aircraft is currently provided by three separate systems: the Programmable Signal Processor, display processors, and a general purpose computer. The Programmable Signal Processor integrates four processors, each executing about 1,500 MOPS (million fixed-point operations per second). Programming is in microcode. The display processors are workstations based on DEC Alpha processors (172 Mtops). The general purpose computer is a Raytheon 6000. None of these systems is state-of-the-art. Reprogramming is difficult and time-consuming. A multi-purpose HPC system is being developed to integrate the functions of these three computer systems into one (although workstations will still

be used for displays) to provide greater target resolution capability, faster and wider area searches, and more flexibility (i.e., through simpler procedures for reprogramming).

Computer support requirements for AWACS aircraft are similar to those for JSTARS, with more emphasis on radar signal processing, which was recently upgraded. The special purpose radar signal processor now provides about 12,000 Mflops of sustained performance; the old processor was only capable of 600 MOPS. The general purpose processor has been upgraded from 2.5 MIPS to 100 MIPS. The emphasis in radar signal processing is to achieve a very high probability of target detection, with very low false alarm rates, and accurate target tracking. Multi-node versions of the EPHC-10 "Touchstone" computer system are being developed to meet the computing needs of both the JSTARS and AWACS platforms, as well as UAVS. Each node of an EPHC-10 is the equivalent of a 64-node Intel Paragon (2,621 Mtops), and the goal is to integrate sufficient nodes to provide 40,000-80,000 Mtops of performance. Equivalent computational power could not be provided by clustered or networked workstations due to size, weight, and power consumption constraints. Configurations that rely on interconnections, such as clusters and networks, are also vulnerable to vibration and other environmental stresses that make them less than optimal. In addition to providing high-performance capability in a small form factor (@8,000 Mtops/ft3), the EPHC-10 is also very power efficient, which is of critical importance in an environment where only a few kW of electrical power are available.

Enhanced acoustic signal processing capabilities will be required in deployable systems with the development of the multi-line towed acoustic sensor array. In contrast to the relatively low data rates in current single-line arrays (e.g., TASS), multi-line arrays will generate about 2 billion bits of information per second. Real-time processing of this data will require about 60,000 Mflops of sustained performance (i.e., at least 120,000 Mtops). Moreover, MPP architecture is inappropriate for this type of processing, providing only 10% efficiency.

Key judgments-military operations

*Computational support to military operations is probably the fastest-growing area of national security HPC applications.

*Development or acquisition of comparable capabilities by countries of national security concern would present an immediate threat to U.S. military personnel and operations, and are not easily countered.

*The predominance of real-time processing requirements for computing support to military operations, coupled with the large volumes of data to be processed for most applications, necessitates the use of very high-powered computer systems. - Virtually all applications examined require computers of at least 7,000 Mtops; most require computers of more than 10,000 Mtops, and many require computers more powerful than currently available in deployable configurations (i.e., more than 20,000 Mtops).

*HPC support to military operations is predominantly required in air, land, and sea mobile configurations, placing significant size, weight, and power consumption constraints on the acceptable characteristics. There are also unique environmental stresses, such as vibration, that constrain choices of architecture. These constraints severely limit the potential for use of clusters or networks of conventional computers in many military operational applications.

Key Findings and Conclusions

There exist a significant number of applications of national security concern with extensive computational support requirements. Proliferation of some of these applications could be slowed or prevented through export controls on HPC.

While the control of HPC exports will have little or no effect on foreign programs to develop first-generation nuclear weapons, it may significantly impede the ability to develop second-generation weapons or verify the capabilities of existing weapons. Given the worldwide availability of computing resources below the level of controllability, any efforts to interdict first-generation nuclear weapons programs through computer export controls would not be effective, and would likely be so futile as to damage the credibility of export controls more generally. Significant cryptologic capabilities can be achieved through the use of widely available computer equipment, such as clustered or networked workstations or simple massively parallel processors, making cryptologic applications inappropriate as a basis for establishing an export control regime or defining a control threshold.

The control of HPC exports has the potential to degrade foreign RDT&E capabilities and preclude some applications. In a growing number of applications, this degradation will be minimal-by denying foreign programs the use of the most efficient systems available. Most RDT&E programs -of national security concern- about two-thirds of those examined in this study-can be conducted at some level of adequacy through the use of clusters or networks of uncontrollable computers. However, some applications, such as the development of advanced acoustic signal-processing capabilities, cannot be conducted without the use of the most powerful computers available, principally because of the limited efficiency of parallel architectures to execute certain classes of algorithms. Table 14 summarizes the computational requirements of the RDT&E applications discussed in this study.

Table 14. [Omitted] Summary of Representative Computational Requirements for RDT&E

Prevention of acquisition of deployable HPC systems by countries of national security concern might significantly limit or prevent the development of sophisticated military operational threats to U.S. national security. While many functions, such as C41 and battle management, can be implemented on distributed and/or low-powered systems, significant applications of HPC for military operations will continue to require very high-performance computers in small, deployable configurations. Table 15 summarizes the computational requirements of the military operational applications discussed in this study. Nearly all of those

examined require computers of at least 7,000 Mtops, and most require systems of 10,000 Mtops or more.

Table 15. [Omitted] Summary of Representative Computational Requirements for Military Operations

A growing proportion of requirements can be fulfilled by clusters or networks of computers.

The combination of the increased performance of uncontrollable computers, and the nature of ,Many applications that are readily adaptable for parallel processing has resulted in the ability to efficiently execute many applications on clusters or networks of personal computers and workstations. Having been enabled by technological developments, this trend is accelerating. Researchers prefer to have more individual control over computer processing, and this has been made possible by the migration of many applications from the large centers to local networks. Additionally, budget constraints limit the ability of many organizations to purchase or use the most powerful computer systems, whereas they may be able to achieve useful results through networking of more readily available equipment. The use of otherwise idle time (e.g., nights and weekends) on ubiquitous computers is often a more efficient allocation of computing power than the use of dedicated HPC. There continue to be applications for which clusters or networks of computers cannot be used.

Some applications simply cannot be converted for parallel processing at this time. These programs cannot be executed on distributed (clustered or networked) computer systems. Such applications are characterized by the requirement to process very large arrays of tightly coupled data. Man also require real-time or near-real-time processing of data; such applications that process large volumes of data cannot today be effectively executed on systems with extensive interconnections.

For other applications, constraints on feasible computer architectures arise from environmental considerations. Computer systems for ground or sea (surface and subsurface) deployment have significant size and weight constraints. Airborne systems are even more constrained, and are also subject to power consumption limitations and the increased need for highly reliable interconnections.

There are groups of HPC requirements above the level of uncontrollability.

Figure 10 depicts the distribution of the national security applications that have been most closely examined during this study. The applications at the lowest end of the chart have already effectively been "given up," in that the computers necessary for their execution are currently readily available on the international market in the form of expandable computers below the HPC control threshold or because they can be effectively executed on distributed architectures.

Figure 10. [Omitted] Distribution of Minimum Computational Requirements

There are groups of applications that require the use of HPC above the level of uncontrollability. Many of these applications are unsuited, in their present forms, for execution on distributed architecture systems. Given the economic and technological changes in the HPC environment, it is not clear to what extent these applications will continue to require HPC for their execution in the future.

Chapter 4 Notes

57 US Department of Defense High-Performance Computing Modernization Office (HPCMO), Requirements Analysis (S&T Report) (April 1994), pp. 5-7.
58 HPCMO, Requirements Analysis (Draft DT&E Report) (April 17,1995), p. 11.
59 HPCMO, "System Speeds (GF) -'94 Questionnaire" informal database report (May 1995).
60 HPCMO, "Export Control" informal database reports (May 1995).
61 See, for example, Harvey, et al., op. cit., pp. viii, 21-23; also, Jack Worlton, "Some Myths About High Performance Computers and Their Role in the Design of Nuclear Weapons," Worlton & Associates Technical Report No. 32 (June 22, 1990).
62 Department of Energy (Code NN-43), "The Role of Supercomputers in Modern Nuclear Weapon Design," (April 27,1995), pp. 2-3.
63 Bruce Schneier, Applied Cryptography (New York: John Wiley & Sons, 1994), p. 131.
64 A large number of people were interviewed in the process of collecting data for this chapter, from which were derived most of the data on specific configuration and processing requirements. Appendix B lists the organizations that provided data for this phase of the study.
65 Bob Brewin, "Software Battle Dogs DOD," Computer World (May 15, 1995), p. 80.
66 William B. Scott, "Air Force 'ALERT' System Speeds Missile Warnings," Aviation Week & Space Technology (May 1, 1995), pp. 56-57.
67 K.W. Johnson, J. Bauer, G. A. Riccardi, K.K. Droegemeier, and M. Xue, "Distributed Processing or a Regional Prediction Model," Monthly Weather Review (November 1994), pp. 2558-2572.
68 Steve Plimpton, Gary Mastin, and Dennis Ghiglia, "Synthetic Aperture Radar Image Processing on Parallel Supercomputers," Communications of the ACM (July 1991), pp. 447-448.
69 Michael A. Dornheim, "New Sensors Show Two Paths to Reconnaissance," Aviation Week & Space Technology (July 10, 1995), p. 46.
70 These figures are necessarily approximate. As discussed on age 47, CTP figures derived from
estimates of sustained (vice peak) performance are rather general.

CHAPTER 5. APPLYING THE BASIC PREMISES ANALYTICAL FRAMEWORK: RESULTS AND POLICY IMPLICATIONS

We now combine the results of Chapters 3 and 4 and apply the analytical framework outlined in Chapter 2. First, the framework is applied for the present time (circa mid-1995), to determine (a) whether or not a threshold satisfying the three basic premises currently exists; (b) if so, what the upper and lower bounds of the threshold are; and (c) what options there are for selecting a specific threshold within this range. The durability of the threshold in the future is then examined.

Establishing the Existence of a Valid Control Threshold

Figure 11 shows the distributions of computer systems found on the June 1995 Top5OO 71 list of the 500 most powerful computers installed throughout the world and the minimum computational requirements, in Mtops, of approximately 600 research, design, testing and evaluation applications of national security interest. In addition, a number of applications, chosen for their relevance to this study and discussed in Chapter 4, are shown individually in the upper portion of the graph. For these, the minimum requirements and performance (in Mtops) of the actual systems used are indicated.

Figure 11. [Omitted] Threshold Analysis: June 1995 Snapshot

The three vertical lines (at roughly 4,200, 4,800, and 7,600 Mtops) reflect the lower bound for a viable control threshold in 1995, 1996, and 1997. Derived in Chapter 3, see Figure 6) they represent the performance of the most powerful uncontrollable general-purpose systems for these years.

The current snapshot makes clear a number of points:

* There are high-performance computer systems that are controllable, and which deliver computing power not available from uncontrollable systems. Since the lower bound is defined by the most powerful uncontrollable systems, Top5OO supercomputer sites with CTPs above the lower bound are controllable, by definition. 72 In June 1995, there were at least 250 installations of computer systems with CTPs above the lower bound. The third basic premise is satisfied.

* There do exist applications of national security interest that require computing power greater than that which can be delivered by uncontrollable systems. This study makes no claim of identifying all such applications. The existence of some, however, means that the first basic premise is satisfied.

* The number of such computationally demanding applications, seen as a fraction of the set of applications of national security interest, is currently small. At present, the number of such applications appears to be diminishing as advances in computer technology make it possible to carry out existing applications on other than the most powerful systems available. In some categories, such as nuclear weapons design and cryptology, the remaining computationally demanding applications are not likely to be pursued by any but the most sophisticated

countries of national security concern. 73 The number of such applications in the RDT&E category is diminishing rapidly, although some applications will always require the most powerful computers available. In the future, most such computationally demanding applications are likely to be found in the military operations category. Nevertheless, the fact remains that:

* The computing power needed to carry out a very large fraction of applications of national security interest is no longer controllable. One of the most distinguishing features of Figure 11 is the large number of RDT&E applications that require fewer than 1,000 Mtops of computational power. The inability to deny potential adversaries the computing power to perform these applications must be recognized as a fact of life. This same tendency can be observed in other categories of applications. In particular, as discussed in Chapter 4, essentially all of the basic applications of national security interest in the nuclear and cryptography categories are below the control threshold. Applications in these categories can no longer provide a legitimate basis for an export control regime. Moreover, if it hasn't happened already, it is all but inevitable that some day an adversary will use an American-made computer to design or operate a system that harms American citizens or property.

The second premise-that there are countries with the military and technical wherewithal to make effective use of the computing power being controlled-is reflected in the analytical framework only in that if the premise is false, there is little point in applying the analytical framework to begin with. Determining the validity of the second premise is properly the role of foreign policymakers and the intelligence community, but a case can be made on widely available information that there continue to be countries of national security concern that can use high-performance computing capability effectively for many of the applications discussed earlier in this study.

As a practical matter, most of the cost to the government of maintaining the regime is in the overhead of establishing the basic regulatory mechanisms; the cost differential between controlling exports to, say, half a dozen countries versus a dozen is minimal. (Note, however, that the cost to industry is directly proportional to the number of sales subject to licensing and the number of sites subject to control, both of which are a function of the number of countries subject to export controls.) The most important question is whether or not the export regime should exist at all, not the number of countries to which it should apply. Table 16 provides a general indication of selected countries' ability to use high-performance computing capability in a number of important application areas. The table indicates that several countries satisfy the second basic premise, but the on-going revalidation of foreign capability should be an integral part of export control reviews in the future.

In summary, applying the analytical framework shows that at present, the three basic premises hold. The current (1995) lower bound for a viable control threshold is approximately 4,200 Mtops.

Table 16. [Omitted] Foreign Capability in Selected Applications

Selecting an Appropriate Control Threshold

Selection of a specific control threshold should take into account not only the current upper and lower bounds but also the distribution of applications and the trends which will affect the viability of the threshold until the time it is next examined. Figures 10 and 11 show that there is a significant clustering of applications in the 7,000-8,000 Mtops range that is likely to remain above the lower bound until at least late 1996. There Lire other critical applications that lie in the 10,000-15,000 Mtops range.

Policy options

Under the current assumptions, the lower bound will rise slowly from 1995 to 1996, but will increase sharply in 1997. Therefore, the following threshold ranges may be considered.

1. CTP 4,200-5,000. This option reflects the "control that which can be controlled" philosophy. This choice will remain marginally viable through 1995, but will rapidly become obsolete in 1996 and 1997. This threshold is likely to be disputed by those who would argue that small clusters of powerful workstations can perform nearly all applications a single system of this power can perform. This may or may not be true for specific applications, but there are likely to be continual calls to justify this choice. This threshold may be useful if the control threshold is meant primarily to establish a regulatory trigger that causes individual sales to be investigated more closely.

2. CTP 5,000-7,000. Based on the analysis of applications in Chapter 4, there are very few applications that would be compromised at CTP 7,000 that are not already compromised at CTP 4,000. While the controllability threshold will exceed 7,000 Mtops in 1997, a CTP 7,000 threshold would be more clearly viable through 1996 than the CTP 5,000 threshold. The short-term economic benefit of a CTP 7,000 threshold is likely to be small, since only several tens of units at this level, carrying price tags of multiple millions of dollars, have been sold throughout the world. Within 1- 3 years systems in this range of performance levels will be sold for under a million dollars, greatly increasing market demand.

3. CTP 7,000-10,000. There is little compelling reason to set the threshold within this range. A number of significant-applications clustered at these performance levels would slip under the control threshold. These include acoustic sensor development, Synthetic Aperture Radar (SAR), Topological SAR (TOPSAR) signal processing, and Shipboard Infrared Search and Track (SIRST). While the lower bound will inevitably force the threshold to be set at these levels in the future, the potential national security costs probably outweigh the potential economic gains in the near term.

4. CTP 10,000-15,000. Several key applications are clustered at or just above this level-- military weather prediction, unmanned Aerial Vehicles (UAV), and simulation of nuclear blasts (given the availability of test data). The national security cost of compromising these applications is significant and, in the near future, unnecessary.

Threshold selection with different underlying assumptions

The lower bound analysis was based on the assumption that the systems defining the lower bound are in fact controllable for two years after they are introduced. This assumption might be overly conservative. In the workstation markets, it is not uncommon for 70-90% of lifetime sales to be made in the first two years. The actual point of uncontrollability could be as early as a year for some models. If we assume that systems become uncontrollable a year after introduction, the rationales for the threshold numbers change somewhat. Thresholds of CTP 5,000-7,000 will remain viable through 1995, but are likely to become increasingly obsolete in 1996. A threshold at CTP 7,000 is likely to remain viable through the end of 1996 and early 1997.

The estimations of applications' computing requirements are based on experience in the United States. It is often claimed that scientists functioning in hardware-poor environments develop improved models and algorithms to compensate. While this is not true for every application, it may be true in selected instances. It would be extremely difficult to compare the algorithms employed for applications of national security interest in, say, Russia or China with those used for comparable applications in the United States. Nevertheless, if and when such information becomes available, it should be reflected in the analysis. The net result will be a decrease in the computational requirements of some applications, causing them to move to lower CTP levels in graphs like Figure 11.

Using the Methodology in the Future

'The discussion above has demonstrated how the analytic framework can be used to (a) establish the current range of viability of a control threshold, and (b) offer options for the selection of a specific threshold at a specific point in time. It is crucial, however, that the methodology be applied at regular intervals. The industry and the world are sufficiently dynamic that few detailed projections are likely to be accurate for much more than a year. We recommend that the analytical framework be applied no less frequently than once a year. An annual review should re-examine the following elements:

1. The lower bound of controllability. The analysis should identify those computing systems, domestic and foreign, which are considered uncontrollable. Particular attention should be paid to trends in architectures and technology usage as the technologies evolve. The emergence of symmetrical multiprocessor systems his had profound implications for export control. Periodically, there will be other technology developments of comparable impact. The earlier these can be identified, the more accurate the analysis will be.

2. The applications distributions. While applications trends are less dynamic than technology trends, they are by no means static. The computing requirements of known applications should be periodically reevaluated. Thanks to breakthroughs in algorithms, the computing requirements for certain applications may drop significantly. Such events are particularly important when an application is one used to define the upper bound for threshold selection.

3. The emergence of new applications. Advances in technology enable new, previously inconceivable applications to be carried out. One should not assume that because all current applications will eventually slide below the controllability threshold that some day there will be no applications left above it. This may be true for some categories, but not for all. 74

4. The most powerful computing systems available. The distribution of the most powerful systems available sets a ceiling on what the upper bound can be. Changes at the high end of this distribution, particularly a decline in the rate of increase, could indicate a potential narrowing of the spread between the most powerful systems and the lower bound of controllability. Such a narrowing would reduce, perhaps very significantly, the range over which a control threshold is viable.

At a minimum, a review of the lower bound of controllability (1) should be performed annually. Within the framework developed here, this could be accomplished quickly and at low cost simply by reviewing vendor literature and trade publications, and obtaining estimates of the size of the installed base from vendors or market researchers. Item (4) could similarly be easily determined from publicly available sources such as the Top5OO list of supercomputing sites. Determining the applications distributions and identifying emerging applications (2 and 3) is more difficult. Because applications and computational methods change more slowly than the hardware/software technologies, however, it may not be as necessary to review the methods as the technologies. Furthermore, as pointed out in the next chapter, a key question to be answered is which applications can, and cannot, be carried out using only uncontrollable technologies. It is not unreasonable for policymakers to expect some of the burden of proof (and cost) of demonstrating that particular applications have slipped below the controllability threshold to be borne by those most strongly favoring a relaxation of controls.

Chapter 5 Notes

71 J. Dongarra, H. Meuer, and E. Strohmaier, "Top5OO Supercomputer Sites," http://parallel.rz.uni-mannheim.de/top5OO/top5OO.html. This listing is not 100% complete and reflects some companies more faithfully than others. Some companies contribute data to the list directly, while others do not. Furthermore, there are classified installations too sensitive to represent in any fashion on the list. Nevertheless, the list is a good indicator of the general distribution of systems at various performance levels, and is by far the most accurate source of installation data readily available.

72 At the same time, there may be individual models with performance below the lower bound that are manufactured in small numbers and have particular physical and operational characteristics that make it possible to regulate their export and use successfully. Foreign practitioners would not have to obtain such systems to acquire lower-bound performance; it would be easier to acquire less controllable systems.

73 For example, extremely powerful systems are still used for nuclear weapon design, but the purpose is to design safer weapons and stockpiles, simulate detonations in the absence of live tests, etc. These issues are not likely to be of great or the same concern to potential adversaries. The computing power needed to design a nuclear weapon that will explode is no longer controllable.

74 Recall the often-repeated predictions made during the early days of computing that the total world-wide need for ENIAC-class computers was less than a dozen.

CHAPTER 6. LOOKING TO THE FUTURE: TRENDS AND ISSUES

Technology and Applications Trends

The current control environment is particularly dynamic, feeling the effects of dramatic changes in technology as well as the economic and geopolitical repercussions of the end of the Cold War. These changes have a direct impact on the three basic premises, whose continued validity remains a precondition for a successful export control regime for high-performance computing. This section describes some of the most salient trends in both technology and applications. The final section discusses the continuing viability of the control regime in light of these trends.

Continuing improvement in performance of high-end, controllable, and uncontrollable systems

The existence of powerful systems whose export can be controlled is a necessary prerequisite to a successful export control policy. Figures 12 and 13 show that there will continue to be controllable systems for at least the next two years, and that their performance levels are projected to increase at a rate at least as great as the lower bound of controllability.

Figure 12. [Omitted] Trends in Distribution of Top5OO 75 Installation

Furthermore, Figure 12 shows that growth in the number of installations in all performance categories is strong. The data reveal the following:

*Growth continues to be strong in all performance categories. In spite of reductions in government funding, many leading computational centers have been able to find the resources to acquire leading-edge systems.

*The spread between the most powerful and least powerful on the Top5OO list is getting wider. This reflects, in part, the great variability in performance that can be provided by massively parallel systems that can be scaled from tens to hundreds or thousands of processors.

*The drop-off in installations in the 0-2,000 Mtops range is of course due to the fact that the list, by design, contains only 500 systems. Growth remains strong in the workstations and PC markets that inhabit this range.

Figure 13 shows trends in CTP mean, median, and high and low percentiles of the Top5OO systems. The trend in the top 25 most powerful installations (95th percentile) is projected to jump significantly next year as this level works its way out of the bubble of 16-processor Cray C90 installations (CTP 21,125). As it does so, it is likely to increase more rapidly than the other statistical measures shown in Figure 13, reflecting the widening of the gap between the most and least powerful systems in the Top500 list.

Figure 13. [Omitted] Top5OO Trends and the Lower Bound of Controllability

Significantly, the data seem to indicate that the lower bound of controllability and the median of the distribution will track each other for at least a couple of years into the future. In other words, the performance of all but approximately 250 of the most powerful installations will be attainable using systems whose export cannot be adequately controlled. Nevertheless, the existence of these 250 systems above the lower bound of controllability indicates that there will remain, for at least the next two years, levels of computing power that can be controlled.

The growing role of networked systems

High-performance computing systems today rarely operate in isolation. It is more typical for individual computing systems to function in the context of a network of systems, that often joins together a number of different kinds of computers, from personal computers and workstations to massively parallel and vector-pipelined systems. Networked systems, which can range from loosely coupled PCs and workstations or, a local area network (LAN) to heterogeneous distributed systems spanning
continents, are usually developed to provide some combination of the following benefits: 76

*sharing of resources (e.g., peripheral devices, data, processors)
*increased through-put
*more cost effective computing
*improved performance on individual problems

Because the export control regime focuses on reducing potential adversaries' ability to carry out applications of national security concern, the most critical benefit is the last one: improved performance on individual problems. If the power of a number of systems can be harnessed in a cost effective manner to reduce the solution time of individual problems, the implications for export control are considerable. A network of systems offers greater performance than any individual component, yet is only as controllable as its most controllable component.

Distributed and parallel systems development is the subject of intensive research and development, and much progress has been made in recent years. A World Wide Web site at Carnegie Mellon University has pointers to over 75 distinct development projects. 77 As improvements are made in interconnect technology and the systems software needed to parallelize code and manage a distributed system, the distinction between a network of computers and an integrated parallel system is becoming blurred.

At the same time, the field, particularly in the area of reducing the solution time of individual applications, cannot be considered mature. While individual projects are beginning to reveal some of

the potential of networked systems, 78 it is clear that achieving the performance benefits is not always a straightforward task. 79 Furthermore, there is very little hard data on the relationship between different kinds of hardware/software/network configurations and applications of concern to the export control community that can guide policy makers in their decisions about how to incorporate networked systems into the control regime.

It is clear that the role of networked systems in applications of national security concern and the resulting implications for export control policy must be studied in much greater depth. In particular, efforts must be made to distinguish those applications that can be performed effectively on networked systems, from those that cannot. We have tried to make such a distinction on the basis of necessarily limited interviews with leading practitioners in key application areas. However, if export control policy is to be truly based on a factual, objective foundation, a great deal more time and effort must be spent on this problem. Such efforts must include in-depth analysis of the computational approaches being used, the problem size, the real time constraints, the relationship between core and user-interface computations, etcetera, and the relationship of these factors to existing architectures and systems. The applicability of networked systems to particular applications should be demonstrated through actual implementations. Studies along these lines stand not only to benefit policy makers, but also the fields of computing and computational methods in general.

The changing nature of computing applications of national security interest

This study has not discussed directly the implications of a changing geopolitical environment on U.S. national security interests. These effects are reflected indirectly, however, in the changing nature of applications being pursued by the military and civilian high-performance computing communities. The nature of the applications and technologies used are influenced by national security priorities, funding allocations, and the characteristics-both technological and economic-of the technologies that can be acquired. Collectively, these factors have resulted in the following observable trends:

Growing diversity in the application of high-performance computing. Supercomputing was once characterized by the use of sophisticated number-crunchers with extensive power and cooling support systems such as those operating in computer centers of the national laboratories. It is now possible to place computing power previously available only in stationary supercomputers on vehicles such as aircraft. A future Cray Research system will offer 20 Gflops of sustained performance in a chassis the size of a mailbox and weigh 32 pounds! 80 Supercomputing power is widely used not only for the traditional applications of design, modeling, and data analysis, but for real-time operational control as well. The latter in particular has specific requirements beyond "getting the right number."

Growing reliance on simulation. Simulation can be a highly cost-effective means of testing systems., training personnel, and exploring scenarios. Commercial technologies today have performance and graphics capabilities that permit realistic

modeling of a system's environment. Particularly significant is the trend toward "hardware in the loop," which involves, for example, interfacing the guidance system of a cruise missile directly with a simulator in real time in order to evaluate actual missile response under a variety of conditions, Simulation has also emerged as an important tool for applications in which live testing is undesirable. For example, simulations of nuclear blasts are needed when physical tests are infeasible.

Growing reliance on commercially available technologies. It has never been the case that all applications of national security interest have been run only on the most powerful systems supercomputer manufacturers could build. Applications have always been constrained by the amount of money that could be used for system acquisition. As military budgets have been squeezed since the end of the Cold War, the need to stretch acquisition dollars has become more acute. While some installations continue to be able to afford the most powerful systems (as indicated by the high-end Top5OO distribution), evidence points to a tendency among a growing number of practitioners to use commercial, off-the-shelf (COTS) technology, which has excellent price/performance characteristics and rapidly increasing capability. Additionally, initiatives to "re-invent government" through, inter alia, procurement reform emphasize the use of COTS equipment wherever possible in lieu of equipment specifically designed or modified for military use.

Growth in the role of distributed systems. The increasing power and capabilities of mass-market computers and workstations, coupled with developments in high-speed interconnects, have fostered the proliferation of distributed systems. These systems are particularly relevant to data-intensive but loosely coupled applications, such as some of the elements of C 4 1 (e.g., data fusion, database mining, communications). Where not constrained by physical limits (on size, weight, or power consumption), distributed systems offer the potential for delivering high-performance computing capabilities from relatively inexpensive hardware. Such systems place a premium on inter-computer communications, however, which suggests that the limits to adversarial exploitation of distributed architectures will arise from inadequacies in infrastructure rather than computer system availability.

The Continuing Viability of the Current Control Regime

Implications of Trends in Technology and Applications

In the near term, the current export control regime will remain viable. Strictly speaking, the three basic premises described in Chapter 2 hold.

There are still applications with very demanding computational requirements. Not only are such applications being performed today, but new applications will arise in the future. Technological advance has always resulted in the emergence of new, previously unconceived applications. There is no reason to believe this will not continue.

The computing power required for these applications can be controlled. For the near term, the data indicate that there will continue to be a gap between the lower

bound of controllability, and the maximum, defined by the most powerful systems available. Reductions in federal funding for high-end systems, the rapid development of SMP systems, and the growing role of aggregated, or clustered systems will not, in the near future, result in the evaporation of the need for high-end, integrated systems. Cray's C90 and T90 series, large configurations of Intel's Paragon, IBM's SP2, and Cray's T3D, etcetera continue to be installed. Although they may incorporate commercially available technologies, large configurations still need specialized technologies and/or extensive vendor expertise to install and run. Systems at this scale, price range (greater than $10 million), and installed base (units and tens of units) remain controllable.

There continue to be countries of national security concern with the wherewithal to use such computing power to the detriment of U.S. national interests.

Whereas the three basic premises hold, the overall efficacy of the HPC export control policy is declining. As long as there are some applications that can only be performed satisfactorily on controllable technology, a control regime can be viable. But as growing numbers of applications of national security concern are performed on technology at or near the controllability threshold, the scope of effectiveness of the policy decreases. The most computationally demanding versions of a specific problem might not be attempted because practitioners cannot afford (or choose not to buy) the necessary computing system. These practitioners will do the most they can with the technology they acquire. When the technology they use lies close to or below the controllability threshold, the export control regime can do little to deny potential adversaries the computing hardware needed to address problems of the same type and size.

The growing diversity of HPC applications, particularly in the area of operational applications, underscores a growing weakness of the current export control regime: the use of the Composite Theoretical Performance metric as the chief measure of system performance.

The role of the composite theoretical performance metric
The CTP metric was developed during the early 1990s to provide a new method of measuring performance for export purposes, to replace the increasingly problematic processing data rate (PDR). 81 Based on a system's hardware features, 82 the metric is, by design relatively easy to derive and use, applicable to broad categories of architectures of software- and applications-independent. Regarding the last criterion, CoCom members felt that hardware controls should not be based on performance measures (e.g., benchmarks) that were software dependent. 83 The formula was updated in 1993 to correct some initial flaws and make the metric more closely correspond to the observed system performance. The three criteria above remained intact.

It can be argued that the CTP metric functions as well as any performance metric could, given its goals and the intended use. It was designed to be a rough indicator of raw performance, not actual performance on a given machine for a given application. (There is no single benchmark in existence that provides accurate performance data for even a fraction of existing systems on broad categories of

applications.) It was designed to enable export control personnel to quickly differentiate systems whose export should be closely scrutinized from those that could be exported more readily, with minimum bureaucratic involvement.

In a world in which both applications and computing systems are becoming more diverse, however, the inherent weaknesses of any metric that adopts a one-size-fits-all approach become more evident. These weaknesses derive not so much from the specific formula used, but from the underlying criteria on which

it is based. Specifically,

1. The CTP metric implies that systems with the same CTP are equally suited to all applications. There are few who would argue that this is true in practice, but the CTP metric provides no basis for any other conclusion. This implication is a direct result of the application-independent nature of the metric. This feature of the CTP is the cause of much of the confusion in discussions debating the role of clustered systems in determining a control threshold. That the CTP formula does not provide a means of computing the composite theoretical performance of networked systems is a technicality that avoids, but does not resolve, the underlying problem. The reality is that a system's performance on an application is a function not solely of the system's raw computing capability (or even architecture), but also of the nature of the application itself. Two systems with comparable CTP might have significantly different overall utility in different application domains. If the goal of the export regime is to limit the ability of potential adversaries to carry out those applications, there may be good reasons for wanting to distinguish between two such systems. The current CTP-based system provides no formal means of doing this.

2. The CTP metric does not account for many system features that, in today's world, can be as important as raw performance in determining a system's utility. Traditionally, supercomputing has focused on number crunching, on raw computational performance. It was not inappropriate for the export control metric also to concentrate on computational performance. While raw processing power remains, arguably, the most important characteristic of high-performance systems, in today's diverse computing environment a number of other characteristics are growing in importance. These include:

a) Some systems simply may not have enough memory, main or otherwise, to perform certain applications. For example, according to a recent NASA study in which NAS Parallel Benchmarks were ported to SGI workstations. 84 Slightly over half of the codes in the suite were too big to run on any NAS workstation . . . ." There is no way to account for such limitations by examining the CTP metric. Furthermore, moving data back and forth between main and secondary storage (or between cache and main memory) remains a relatively time- expensive operation. Although systems with smaller memories may eventually produce a correct result, the time penalties of memory management can make a system unsuitable for particular applications.

b) Interconnect bandwidth and latency. The relationship between the computing power of the processing elements and the nature of the interconnect that joins them plays a very significant role in determining the set of applications for which the system is suitable. Systems with low bandwidth and high latency may be able to run applications that are highly parallel, or that require many independent executions of the same code, but not those which require large amounts of inter-processor communications. 85

c) 1/0 Speed. Many applications depend on particularly high data exchanges with external devices. A system that is unable to support high-volume I/O could be unsuitable for particular applications, even if the raw performance is high.

d) Size. Embedded systems have strict size requirements. A system may not be suitable for embedded applications if it is the size of a refrigerator.

e) Robustness, Systems designed to operate in the controlled environment of an office building may not operate well in the field where they may be subject to conditions of extreme temperature, moisture, and vibration.

f) Reliability. Some applications may not be performed easily if a system's mean time to failure is too low.

g) Real-time operation. Data processing systems may generate results quickly, but often cannot guarantee that a result will be obtained within a specified amount of time. Ensuring real-time operation on a massively parallel system, for example, requires non-trivial modification of the system's operating system.

While the CTP is an improvement over its predecessor and has served the export control regime rather well in the past, its inherent shortcomings and omissions may now be severe enough that the use of the metric should be re-examined.

Conclusions and Recommendations for Further Study

The export control regime has traditionally been technology centered, focusing on supercomputers and, more specifically, supercomputer performance. The regime has evolved to take into account advances in technology and the inherent limitations of export control enforcement mechanisms. Controllability has, appropriately, become a significant variable in policy formulation.

Although it has always focused on limiting the ability of countries of national security concern to carry out computationally demanding applications, the policy has continued over many years without rigorous evaluation of what those applications are, the degree to which the required computational resources can be controlled, and which kinds of computing systems may be applied effectively to them. This point is clearly illustrated in the persistent reference to nuclear and cryptographic applications as the principal motivations behind a control regime.

Without neglecting the focus on technology, we recommend a more balanced approach, with greater emphasis on applications than has been the case in the past.

Recommendation 1: Perform annual reviews of the export control regime, applying a methodology that is open, repeatable, and based on reliable data.

Each review of the export control regime necessarily takes into account not only current circumstances, but also projected trends. But the world is highly dynamic. Projections about the future are notoriously inaccurate. If reviews are held annually, reliable projections beyond a year in the future are not as necessary.

Recommendation 2: Significantly enhance the analysis of applications of national security interest.

An application-oriented approach can provide greater flexibility in export control administration, while remaining truer to the basic purpose of the export control regime. To stimulate discussion on this point, we suggest the following procedure:

1. Draw up a list of those applications of national security concern that are to provide the basic rationale for the continuance of an export control regime, that is, which satisfy the first basic premise. Recognize that @i given application area (e.g., stealth aircraft design) may contain several applications that are similar in intent but qualitatively different in scope, size, and, ultimately, computational requirements. Such versions of applications should be listed independently.

The importance of preventing the proliferation of these applications must be assessed. Not all applications of national security interest are of equal concern, nor do they all require the same degree of protection. It remains unclear at this point whether the kinds of applications examined for this study form as compelling a justification for export controls as did nuclear, cryptologic, and ASW applications during the Cold War. The likely impact of deciding not to or failing to prevent the proliferation of these applications should also be determined.

2. For each application, determine whether or not it can be performed successfully on uncontrollable technology at a cost comparable to that of doing it on controllable technology. Cost should reflect not only hardware and software acquisition, but also the costs in time and human effort needed to perform the application. A report from Cray Research comparing supercomputers and workstation clusters points to a relevant issue:

The time, money and work required to assemble a group of comparatively low performance workstations, develop the software to distribute and load balance jobs, get the users to parallelize their codes where required, administrate and charge resource usage appropriately, and provide parallel debuggers and performance analysis tools for distributed jobs will cost far more than [the cost of purchasing the Cray system with
a comparable amount of raw processing power]. 87

The absolute value of these costs might be difficult to determine, but is less important than their value relative to the costs of performing the same activities on controllable technology.

3. If an application in (2) can be performed cost effectively on uncontrollable technology, then remove it from the list. The computational capability needed for the application can no longer be controlled. If the application can be performed on uncontrollable technology, but only with great effort and uncertainty, then the controlled technology still gives the United States a competitive edge. The application should remain on the list.

The lack of shared understanding about what are the critical applications that should, and can, be protected has been an irritant between industry representatives and government policymakers. Without specific data, it is difficult to answer the claims that, on the one hand, "all applications of national security concern can now be performed on, e.g., clusters of workstations," or, on the other, "that there are 'lots' of applications that can and should be protected," Such a list should help establish a shared understanding of the validity of the first basic premise.

The list also implicitly would validate the third premise as well. By the way the list was constructed, the only applications that remain are those that can only be performed effectively on controllable systems. If there are no features that make computing systems controllable, then the number of applications on the list drops to zero. It is important to keep in mind that the absence of an application on a list constructed in this way means simply that the application cannot be denied effectively by efforts to restrict access to HPC hardware. There may still be other necessary technologies (e.g., manufacturing technologies) that may serve as control points.

it is important to note that the computational requirements of these applications, established using the methodology of this study, may not trace out a neat pattern along the CTP scale. There may be applications with relatively high CTP equivalents that are taken off the list because they are "embarrassingly parallel" or otherwise well suited to clusters of systems with high aggregate performance. There may also be applications with relatively low CTPs that remain on the list because they can be properly performed only on machines that are controllable, but of only modest performance. For this to be the case, system characteristics other than raw performance must be taken into account.

Once the core applications list has been constructed, the applications-centered approach needs a mechanism for distinguishing the systems, or classes of systems, that may be subject to controls from those that do not. The mechanism need not function with great precision, but need only attract attention to those systems whose export should come under more careful scrutiny. As was the case with the CTP formulation, the mechanism should be easy to apply and use, and should apply to a broad spectrum of systems. Based on this study's preliminary evaluation of significant applications, the mechanism should take into account at least the following: raw performance (possibly measured by the CTP), amount of main memory, interconnect bandwidth and latency, 1/0 speed, physical size, robustness, reliability, ability to perform real-time operation.

This study has focused on the performance requirements of application,,. An appropriate "trigger" mechanism in the future might take into account applications' minimum requirements in these other categories as well. For export control purposes, a trigger for the licensing process could be the application profile, which is at some sense minimally in relation to the profiles of other applications on the core list. If properly formulated, this mechanism could permit weakness in one category to offset strengths in other categories. For example, even if a system has a high theoretical performance, it might still be exportable if it suffers from inadequate 1/0 or memory.

Recommendation 3: Significantly improve the quality of data related to applications of national security interest.

The discussions surrounding the export control regime, including this study, have suffered from a lack of concrete, reliable data. While this study has tried to provide a more rigorous framework and firmer factual basis for the discussion, it is limited by the quality of data that could be obtained given its short duration (from late April through late July, 1995). Better data need to be gathered in It least the following areas: HPC usage in applications of national security interest, actual distribution of high-end systems, and foreign usage of domestic and imported HPC technology.

1. HPC usage in applications of national security interest. Until recently, data about computational requirements of applications of national security interest were not gathered systematically. While the HPCMO surveys have considerably improved matters, they do not reflect all applications of national security interest. Furthermore, it is difficult to draw from them the kinds of data needed for the current study, let alone for a procedure such as that proposed under recommendation number 1.
To establish the performance requirements for individual applications in this study, practitioners were interviewed to determine the actual computer configuration used, and minimum configuration required, for the application. The CTP rating for these systems were used to establish the minimum and preferred performance (in Mtops) requirements.

While these data represent a step toward greater quantification of the computational requirements of applications of national security concern, they are far from ideal. In particular, they reflect actual practice, rather than a more fundamental understanding of the computational nature of the applications. Determining the true minimum computational requirements should be based both on empirical and theoretical examinations of the computational structure of the applications, and on the applicability of a variety of computational methods and architectures to their solution. Furthermore, the figures are subject to the same criticisms made of the CTP metric itself, as discussed above.

2. Actual distribution of high-end systems. Discussions surrounding the export control regime have also been permeated with data about the "availability" of products from vendors, with little regard to which high-performance systems are, in fact, being purchased and installed. To say that a system of X Mtops is

"available" from a vendor when the largest configuration installed at a user site is X/4 Mtops confuses more than it clarifies. Data sources such as the Top5OO Supercomputer Sites should be used to provide data on the distribution of sales of advanced systems, but even these could be improved. Similarly, projections about applications' requirements should be noted, but treated carefully.

3. Foreign usage of domestic and imported HPC technology. The second basic premise states that a justifiable export control regime requires that there be potential adversaries with the wherewithal to use the computing systems effectively. A more focused evaluation of the capabilities of target countries would be helpful. While Chapter 3 established that the lower bound of controllability is now being determined by Western systems, it is widely assumed that non-Western countries will (a) take advantage of the growing power, modularity, and availability of the basic building blocks of HPC systems to construct their own, and (b) apply uncontrollable Western systems to their own applications of national security concern. These assumptions should be tested on an on-going basis. While it is clearly difficult to gather data about foreign military uses of HPC, data about foreign civilian usage can be helpful in establishing a baseline.

Recommendation 4: Conduct a study of the implications of networked computing systems with regard to export control.

As mentioned earlier networked computing systems is a field of rapid development with significant potential impact on applications of national security concern. These systems do not lend themselves to easy classification using a single metric like the CTP, are not easily controlled, and will continue to be a problematic element in export control policy formulation.

Recommendation 5: Cultivate comparative advantage through means other than control of hardware exports.

The primary purpose of the export control regime has been to preserve United States' advantage in national security applications with the goal of providing a relative advantage in capability, in money, and in saved human lives. While the emphasis has often been on denying potential adversaries certain computational capabilities, the policy also, in fact, succeeds when potential adversaries are forced to acquire technology at greatly increased cost, effort, delay, and uncertainty. At a time when trends in technology and funding practices are driving practitioners throughout the world to use many of the same technologies, attention must focus on distinctions besides qualitative differences in hardware that the United States can use to maintain a relative advantage in the global arena.

For example, close working relationships between national security practitioners and systems developers should continue to be encouraged, to ensure that the practitioners have access to advanced technology well before their foreign counterparts. Procurement procedures should continue to be improved. Individually and collectively, U.S. practitioners enjoy a depth and breadth of experience that is unparalleled and not easily obtained in other countries. A rich collection of conferences, publications, and other vehicles for capturing and

disseminating this expertise exist. They should be promoted, and other means of preserving and cultivating this important strategic asset should be explored.

75 J. Dongarra, op. cit.
76 For a quick overview and pointers to other resources, see Andrew S. Grimshaw, "Enterprise-Wide Computing," Science, 265 (August 12,1994), pp. 892-894.
77 http://www.cs.cmu.edu/afs/cs.cmu.edu/project/scandal/public/www/research-groups.html.
78 T.E. Anderson, et al., "A Case for NOW (Networks of Workstations)," IEEE Micro, 15 (February 1995), pp. 54-64.
79 Grimshaw, p. 894.
80 Cray Research Literature, 1995.
81 Ramsbotham, op. cit.
82 Specifically, the CTP formula is based primarily on instruction execution times and word lengths. A key feature is a specification of how to aggregate the CTPs of multiple computational elements in a single configuration. This specification takes into account, in a much less direct fashion, intra-system communications bandwidth and efficiency losses in multiprocessor systems.
83 Ramsbotham, p. 3. All benchmarks that involve running code measure not only hardware performance but also, indirectly, compiler and operating system performance.
84 Clustered Workstations and Their Potential Role as High Speed Computer Processors," RNS-94-003 (April, 1994), p. 26.
85 J. Mohr, "Clustered Workstations: The Dominant Parallel Architecture?" RCI Ltd. (1994).
86 For a treatment of some aspects of this problem, see R. Games, et al., "Real-Time Communications Scheduling for Massively Parallel Processors (Position Paper)," Proceedings of the IEEE Rcal-Tittle Technology and Applications Symposium (May 15-17, 1995), Chicago, IL. In their conclusion, these authors from Mitre Corp. state, "Commercial MPPs today have a tremendous amount of raw network capacity and the tendency is to assume that such high capacity translates into predictable communications performance. Our experiences and analysis indicate that this is not the case. . . ."
87 Karen Allen, "Will Workstation Clusters Replace Supercomputers?" Cray Research (December 21, 1994).

APPENDIX A.GLOSSARY OF ACRONYMS

ACW Advanced Conventional Weapons
ADP Automated Data Processing
AFB Air Force Base
ALERT Attack and Launch Early Reporting to Theater (System)
ANRG Advanced Numerical Research Group, Hyderabad, India
ASCM Anti-Ship Cruise Missiles
ASW Anti-Submarine Warfare
AWACS Airborne Warning and Control System
BARC Bhabha Atomic Research Center, Bombay, India
C41 Command, Control, Communications, Computers, and Intelligence, formerly C31 (less "computers"). It has recently been suggested that the acronym currently in vogue is C4I2-Command, Control, Communications, Computers, Intelligence, and Information."
ccm Computational Chemistry and Materials Science (a CTA)
CDAC Center for Development of Advanced Computing, Pune, India
CEA Computational Electromagnetics and Acoustics (a CTA)
CEN Computational Electronics and Nanoelectronics (a CTA)
CF Computational Functions (DoD DT&E projects)
CFD Computational Fluid Dynamics (a CTA)
C-MMACS Center for Mathematical Modeling and Computer Simulations, Bangalore, India
cocom Coordinating Committee-A now-defunct multi-lateral organization that cooperated in restricting strategic exports to controlled countries.
COTS Commercial, Off-The-Shelf
CPU Central Processing Unit
csm Computational Structural Mechanics (a CTA)
CSTAC Computer Systems Technical Advisory Committee
CTA Computational Technology Areas (DoD S&T projects)
CTP Composite Theoretical Performance-The measure of computer system performance used by the U.S. government to implement export controls, expressed in Mtops (q.v.).
cwo Climate, Weather, and Ocean Modeling (a CTA)
DBA Database Activities (a CF)
DES Digital Encryption Standard
DoD (U.S.) Department of Defense
DSP Defense Support Program
DSP Digital Signal Processing
DT&E Developmental Test and Evaluation
EAA Export Administration Act
EAR Export Administration Regulations
EM Environmental Quality Monitoring and Simulation (a CTA)
FDDI Fiber-Distributed Data Interconnect
FFT Fast-Fourier Transform
FLOPS Floating Point Operations Per Second-A conventional measure of CPU performance.
FMS Forces Modeling and Simulation/C41 (a CTA)
Gflops Giga-FLOPS (billions of FLOPS)

HIPPI High-Performance Parallel Interconnect
HPC High-Performance Computing (or Computers)
HPCMO (U.S. DoD) HPC Modernization Office
I/0 Input/Output
IR Infra-Red
IR&D Independent Research & Development
IRSTIR Search and Track
ITMVT Institute for Precision Mechanics and Computer Technology
IW Information Warfare
JAST Joint Advanced Strike Technology
JSTARS Joint Surveillance and Target Attack Radar System
KB Kilo-Bytes (1024 bytes)
Mflops Mega-FLOPS (millions of FLOPS)
MHz Mega-Hertz
MIMD Multiple Instruction, Multiple Data
MIPS Millions of Instructions Per Second
MKP Macro-Pipeline Processor (A Russian computer)
MOPS Millions of (fixed-point) Operations Per Second
MPP Massively Parallel Processing
Mtops Millions of Theoretical Operations Per Second-The units of measure used to express a computer's Composite Theoretical Performance (CTP), which calculates computer performance as a function of CPU processing power and system architecture.
NAASW Non-Acoustic ASW
NASA National Aeronautic and Space Administration
NDST National Defense Science and Technology University (of the PRC)
NOW Networks of Workstations
NSA National Security Agency
OEM Original Equipment Manufacturers
PC Personal Computer
PRC People's Republic of China
PVM Parallel Virtual Machine
PVP Parallel Vector Processor
RDT&E Research, Development, Test, and Evaluation-The entire process for designing and fielding new systems; encompasses both S&T and DT&E functions.
RISC Reduced Instruction Set Computer
RTDA Real-Time Data Acquisition (a CF)
RTMS Real-Time Modeling and Simulation (a CF)
S&T Science and Technology
SAR Synthetic Aperture Radar
SIMD Single Instruction, Multiple Data
SIP Signal and Image Processing (a CTA)
SIRST Ship-board IRST
SMP Symmetrical Multi-Processor
TA Test Analysis (a CF)
TASS Towed Array Surveillance System
TMDGBR Theater Missile Defense Ground-Based Radar
TPCC Trade Promotion Coordinating Committee
TOPSAR Topological SAR

UAV Unmanned Aerial Vehicles
VAR Value-Added Re-seller

APPENDIX B. FACILITIES VISITED AND PEOPLE INTERVIEWED

The authors are grateful to the following people for providing background information for this study and/or reviewing draft versions of the report. Their participation does not necessarily imply endorsement of this study or its findings.

Industry:

Alex Computer Systems

AT&T Global Information Systems
Connection Machines
Convex Computer Corporation
Cray Research, Inc.

Digital Equipment Corporation

Hewlett-Packard Company
IBM
Intel Corporation

Lockheed Missiles and Space Company
Raytheon Company
Science Applications International Corporation
Silicon Graphics, Inc.
Sun Microsystems

Andrew Talbot

Gerald Matthews

Daniel Hillis

Jim Balthazar, Steven Wallach, Lana Yeager

Michael Allen, Douglas R. Goodman, Charles W. Hayes III, James L. Novakoff

Don Ames, Keith Melchers, Robert Rarog

Constantine Anifantis, Roger Grossel

Aaron Cross

David Hoger, Timothy Mattson, David Rose, Elliot Swan, Robert Wallace, Peter Wolochow

Alan Brown, Vaughn Cable, Jeffrey Newmeyer, Thomas Boak, Samuel Earp, William Grossman, Ann Scott

Luke Alexander, Angela Steen, Donald Stevenson, Bill Van Loo

Academia: Center for Analysis and Prediction of Storms and School of Meteorology, University of Oklahoma, Kelvin K. Droegemeier Center for International Security and Arms Control, Stanford University, John R. Harvey National Center for Atmospheric Research, Robert Chervin Stanford University, John H. Barton University of Arizona, David Hixson University of Tennessee, Knoxville, Jack Dongarra Department of Commerce: Bureau of Export Administration, Ian S. Baird, Sue E. Eckert, William Reinsch, , Maureen Tucker, Joseph Young Department of Defense: Advanced Projects Research Agency, Howard Frank, Stephen L. Squires Defense Information Systems Agency, John A. Gauss Defense Technology Security Administration, Howard P. Ady III, Paul Koenig, Oksana D. Nesterczuk, David S. Tarbell High-Performance Computing Modernization Office, Larry Davis, Anthony Pressley Institute for Defense Analyses, Cathy McDonald, Roger Sullivan

MITRE Corporation: Richard Games
National Security Agency: Charles Harry, Jeffrey C. Padgett
Office of the Secretary of Defense: Kenneth Flamm, F. Barry Horton, Jeremy Kaplan, A. James Ramsbotham, Mitchel B. Wallerstein
Department of the Army: Army Research Laboratory, Dave Brown, Gordon L. Filbey, Virginia Kaste, Kenneth Kimsey, Richard Lottero, Steven Schrami, Armament Research, Development, and Engineering Command Richard Fong, Theodore Vladimiroff, Combined Arms Center Robert Banning, George Kather, David Marlin, William McCollum, Robert Ramsdell, Stephen H. Robinette, National Simulation Center Kenneth Bernard, Herbert Westmoreland, Corps of Engineers Jimmy P. Balsara, Missile Command Jonathan A. Mills, Greg Smith, National Training Center, William Wallace, Tank Automotive Command, Grant Gerhart, John Schmuhl, White Sands Missile Range, Joseph Ambrose

Department of the Navy: Fleet Numeric Weather, Center Leo Clarke, Randy Nottenkamper, David Whalen, Naval Air Warfare Center, Steven Kern, Naval Command, Control, and Ocean Surveillance Center, Adi Balsara, Keith Bromley, Robert A. Dukelow, Don Eddington, John R. Evans, Richard Freund, Aram Kevorkian, Robert C. Mathews, Lynn Parnell, Rik Pierson, Robert Wasilausky, Naval Research Laboratory, Stanley Chin-Bing, Richard Priest, Randall P. Shumaker, Naval Space and Warfare Command, Marvin Cohen, Naval Surface Warfare Center Michael S. Brown, Michael Cheamitru, Gordon C. Everstine, Richard Lorey, Charles D. Milligan, Robert Tacey, Thomas Tinley, Naval Undersea Warfare Center Ernest Correia, Michael Forbes, Anthony J. Kalinowski, Norman I,. Owsley, Stephen, Schneller, Abraham Shigematsu, Craig Wagner, Office of Naval Research Jim Fein, Gary M. Koob, Marc Y.E. Pelaez, Paul Quinn,

Department of the Air Force: Air Combat Command Ralph Kohler, Richard Linderman, Richard Metzer, Electronic Systems Command Antonio Amoroso, David Arpin, Jonathan Bernhardt, John Broglio, Alan Budreau, Steve Carlon, Alan Cohn, Larry Fisher, Audie Hittle,

Peter Hughes, Jerry McKiloop, Phillips Laboratory, Donald A. Chisholm, Rome Laboratory, Richard Linderman, Test and Development Center, Lynda Davila, Wright Laboratory Dennis Andersh, Douglas Davis, Bryan C. Foos

Department of Energy: Lawrence Livermore National Laboratory, Charles Ball, Randy Christiansen, Charles Cole, Richard Gronet, Ronald F. Lehman II, William Quirck, Paul Slokowski, Stanley Trost

Other U.S. Government:

Central Intelligence Agency, Kenneth Tasky, Mark Zimmermann
National Meteorological Service, James Howcroft
National Security Council, Michael B.G. Froman, Robert D. Kyle, Robert S. Litwak, Daniel B. Poneman