

**APPENDIX A. HIGH PERFORMANCE COMPUTING:
CONTROLLABILITY AND COOPERATION**

Joint Statement of the U.S. National Academy of Sciences and
Russian Academy of Sciences Working Groups on
High-Performance Computing¹

by

Seymour E. Goodman
Vladimir K. Levin
Ivan D. Sofronov
Peter Wolcott
Aleksy V. Zabrodin

April, 1993

¹Reprinted with permission from *High-Performance Computing: Controllability and Cooperation. Joint Statement of the U.S. National Academy of Sciences and Russian Academy of Sciences Working Groups on High-Performance Computing*. Courtesy of the National Academy Press, Washington, D.C.

A.1 Introduction

With the end of the Cold War, the relationship between the United States and the Russian Federation has become less adversarial on many fronts, from foreign policy to commerce to science and technology. The hope is that the two countries' basic principles and goals about the nature of relationships between nations have drawn closer. In this context, some of the key features of the Cold War, such as the regimes governing the export of sophisticated dual use technology from West to East and from North to South, should be reviewed and possibly modified.

At the heart of the issue are questions of which technologies or uses should be controlled and which technologies can practically be controlled in terms of their manufacture or proliferation. Controllability depends strongly on the nature of the technology, its availability on global markets, and the organizational arrangements governing its use and distribution. The control of high performance computing (HPC) is particularly problematic because controllability characteristics of constituent technologies (software, micro-circuits, networks, integrated systems, etc.) vary widely, and thanks to the extraordinarily rapid rate of technological advance and diffusion throughout the world, leading-edge technologies move into the mainstream only a few years after their introduction.

In the past, control efforts have consisted of measures taken by the U.S. and its allies in the Coordinating Committee for Multilateral Export Controls (COCOM) which assumed active attempts by agencies of the Soviet Union to divert sophisticated technology. Reliance on Soviet cooperation was minimal. In recent years, remarkable progress has been made diplomatically in recasting the relationship between the U.S. and the Soviet Union/Newly Independent States from one that has been fundamentally confrontational to one that is more mutually beneficial. The cooperation of Russians can be a powerful factor in the export control equation. If the Russians can demonstrate their ability and

willingness to work with Western governments, vendors, and users in keeping sophisticated technologies from being diverted to military uses or restricted destination countries, it is possible that the iron-clad controls of the past can be eased to the benefit of commerce, scientific progress, and Russian transition to a viable market economy.

In any relationship, including that between countries, the reduction of confrontation does not lead immediately to an establishment of trust. The latter can be accomplished only through a) the multilateral establishment of procedures and mechanisms to achieve the goals of non-diversion and non-proliferation, and b) a series of small and incremental steps taken over time in which both parties demonstrate trust, trustworthiness, and a willingness to work together in mutually beneficial ways. These will necessarily involve an element of risk, since measures which give one party complete control over the actions of the other (e.g. iron-clad control over high-performance computer installations) give the latter no opportunity to demonstrate independent good faith and cooperation. Russians must be given the opportunity to demonstrate understanding of and respect for the national security concerns of the United States and cooperate in preventing the diversion and proliferation of sophisticated technology.

In the past, the Soviets' willingness to control diversion and proliferation was questioned, but their ability to do so was not. Strong, centralized political and military institutions effectively regulated sensitive technologies. Today, there are reasons to suspect that the willingness to cooperate has increased. However, it should be noted that the Russians' ability to control has decreased. Partly as a result, concerns about North-South proliferation of technologies to such countries as Iraq and Iran have grown. The Western community should acquire assurance that under the current conditions of fragmentation and decentralization of lines of authority the Russians have the ability to establish an effective, civilian, control regime.

This paper examines the present nature and inherent controllability of high-performance computing technologies. It discusses means of control in the context of broader efforts to create an environment in which the need for controls is reduced. Specifically, it sketches a three-track approach—focusing on application domains, institutional arrangements, and technologies and controls—to building confidence without unduly compromising national security or economic interests.

A.2 Controllability of High-Performance Computing Systems

The High-Performance Computing Act of 1991 defines high-performance computing as "advanced computing, communications, and information technologies, including scientific workstations, supercomputer systems (including vector supercomputers and large scale parallel systems), high-capacity and high-speed networks, special-purpose and experimental systems, and applications and systems software." The ability to store and manipulate large volumes of data and make the results accessible to local and remote users has made such systems powerful enabling factors in a wide variety of civilian and military applications. HPC systems have contributed to leading-edge developments in such diverse applications as the weapons design, integrated-circuit simulation, weather analysis, automobile crash simulation, seismic prospecting, and drug design. Computational research has become a third pillar of scientific advance, together with theoretical and experimental research, and is increasing in importance over time.

In this section we examine some of the trends in high-performance computing development in the United States and Russia. We concentrate on those technologies which could be, or are, most easily obtained in Russia.

A.2.1 Trends in HPC

Over the last half decade in particular the high-performance computing sector has been highly dynamic, witnessing remarkable advances in performance throughout all classes of machines, component bases, storage devices, architectures, software and software environments, data transmission technologies, etc. While systems with traditional vector-pipeline architectures continue to evolve steadily, the emergence of a variety of commercial massively parallel machines with performances in some instances rivaling or exceeding those of traditional vector-pipelined supercomputers has signaled an important era of transition in the field in which massively parallel processors (MPP) have a significant role in complex and increasingly heterogeneous computing environments. Building on dramatic advances in microprocessor technology, workstations, servers, and accelerator boards have blossomed into a \$15 billion market in recent years. On the order of 500,000 workstations are sold worldwide annually.

A.2.1.1 HPC Trends in the United States

Some of the high-performance computer systems which have been introduced in just the last two years by U.S. companies are listed in Table A-1.

The rapid evolution of microprocessor technology is one of the main factors fueling the construction of hardware with impressive theoretical performance. Thanks to manufacturing technologies which can place well over one million transistors on a chip and advances in microprocessor architectures, single-chip microprocessors in volume manufacture today offer 50-200 MIPS. They are small enough that thousands can be combined in a reasonably sized, air-cooled cabinet. Individual microprocessors are the engines for powerful, user-friendly engineering workstations. Several microprocessors can be placed on a single printed-circuit board which can be slipped into standard slots in workstations or even personal computers.

<u>Year</u>	<u>Company</u>	<u>Machine</u>	<u># Processors</u>	<u>Peak Performance(64-bit)</u>
(64-bit) VECTOR-PIPELINE SUPERCOMPUTERES				
1991	Convex	C-3 series	1-8	34.4-960 Mflops
	Cray Research	Y-MP C90	16	16 Gflops
		Y-MP/EL	1-4	133-532 Mflops
1992	Cray Research	Y-MP M90	2-8	666-2664 Mflops
MASSIVELY PARALLEL COMPUTERS				
1991	Intel	Paragon XP/S	66-4096	5-300 Gflops
	Thinking Machines	CM-5	32-1024	4-128 Gflops
	Kendall Square Research	KSR1	32-1088	40 Mflops/node
1992	nCUBE	nCUBE 2E,2S	8-8192	27-34,000 Mflops
OTHER				
1991	IBM	Power-visualization System	8-32	1280 Mflops
	Convex, HP	Meta series		
	IBM	RS/6000 cluster		
WORKSTATIONS				
1992	Sun Microsystems	SPARCstation10	4	400 MIPS (32-bit)
		SPARCcenter 2000	2-20	100 MIPS - 2.19 GIPS (32-bit)
ADD-IN BOARDS				
1992	Transtech	TTM110	1	60 Mflops
	Transtech	PARAstation	4	240 Mflops
	Sky Computers	SKYbolt	16	960 Mflops
	CSPI	SuperCard		
		Quad-860	4	320 Mflops (32-bit)

Table A-1 Recent U.S. High-Performance Computers

Many MPP manufacturers have chosen to use commercial, off-the-shelf microprocessors to save design and development resources. These include the Intel family of parallel systems (based on 286, 386, and i860 microprocessors), Thinking Machines CM-5 (SPARC), and the Parsytec GmbH Parsytec GC (transputers). Other manufactures, including nCUBE, Kendall Square Research, and MasPar, have chosen to use highly integrated, customized processors on the argument that by tailoring the design to include only needed functionality more processors can fit in a given space. Although such processors cost more to design and develop, these companies feel the improved performance and

functionality outweigh the drawbacks. In general, the industry has not come to a consensus over whether customized or off-the-shelf ICs are preferable.

During the past decade the performance, functionality, and availability of storage systems, high-speed networks, software engineering environments, graphics workstations and visualization software, etc. have increased tremendously (although not necessarily keeping pace with advances in microprocessor technology). High-speed networks which provide access to HPC configurations and support the transfer of large volumes of data have transformed the way in which individuals in the HPC community conduct their research and collaborate with one another.

A.2.1.2 HPC Trends in the Soviet Union/Russia

The Soviet Union has conducted research and development of digital computers since the late 1940s. Although not without achievements, the HPC industry has not been able to keep up with the scope and pace of Western development, for a variety of systemic and technological reasons. Two of the most significant hinderances have been the complex and cumbersome political and economic structures needed to support the development of complex technology, and, correspondingly, a technology base unable to support the development and manufacture of machines with world-class performance. In particular, the Soviet/Russian microelectronics industry has been unable to achieve large-scale, reliable production of chips with less than 1.5 micron technology (approximately 30,000 transistors per chip). The manufacture of single-chip microprocessors with the level of integration of even the 386 has not been achieved. Table A-2 lists some of the principal Soviet/Russian HPC computers prototyped or put into series production over the last decade.

In part to try to achieve high performance using relatively slow components, Soviet designers concentrated their efforts on parallel systems. Collectively, the Soviet/Russian

<u>Year</u>	<u>Machine</u>	<u># Processors</u>	<u>Peak Performancs</u> <u>(64-bit)</u>	<u>Status</u>
1992	El'brus-3	1-16	8.96 Gflops	near prototype
	Elektronika-SSBIS	1-2	500 Mflops	prototype
1990	MKP	1-2	1 Gflops	prototype
1988	PS-2100	64-640	1.5 GIPS (32-bit)	series prod.
1985	El'brus-2	1-10	125 MIPS	series prod.

Table A-2 Soviet/Russian High-Performance Computers

projects cover a spectrum of architectural approaches nearly as broad as that in the West, although not as deep. Two recent exceptions, initiated after the success of vector-pipeline computers in the late 1970s, are the vector-pipelined MKP and Elektronika-SSBIS.

Besides those listed, many projects, carried out chiefly within institutes of the Academy of Sciences or the ministry of higher education focused on homogeneous distributed systems. During the 1980s hardware prototypes were built using available indigenous technology, but in recent years developers have been calling such systems "transputer-like" and have focused their efforts on software, often using real transputers as a development base. Most of these researchers have become members of a newly-formed Russian Transputer Association. Many of these projects have been oriented towards developing parallel co-processors or accelerators for general-purpose host computers.

In the past, national security policies in both the Soviet Union and the Western countries forced the Soviet high-performance computing industry to develop to a large extent independently. Basic information about development trends in the West was available, but developers were forbidden to use Western components and users were forced to rely almost exclusively on indigenous computers. Some architectural innovations served partially to compensate for the weak component base, and users invested much effort in de-

veloping models, algorithms, and systems software which would compensate for computer deficiencies. The Russian party feels that such efforts have been very successful for high-priority applications. Little has been published in the West about Soviet/Russian innovations in models and algorithms in particular.

As a whole, the HPC community in Russia continues to suffer from deficiencies in storage devices, although individual high-priority configurations may be reasonably adequately supplied with moderate-capacity storage systems. Facilities for remote, interactive access to HPC installations are at best extremely limited. Only in the 1990s has electronic mail become generally available in Russia.

In recent years, orders for HPC technology and financing for development has been reduced significantly as a result of the conversion of defense industries to civilian production and the deterioration of the Russian economy in general. At the same time, many Russian policy barriers to international contact and cooperation have been lifted, making it likely that the Russian HPC community will, in the future, be more integrated into the world HPC community.

A.2.2 Controllability of HPC

High-performance computing systems have become a particularly problematic element of the export control regime. The extraordinarily rapid rate of technological advance means that products move quickly from leading-edge to the mainstream, threatening to make specific features of export control regulations obsolete before they are published. High-performance computing as a whole encompasses a wide variety of technologies from components to networks to large-scale systems to sophisticated software which have very different controllability properties. The field, dominated by American and Japanese companies is growing more international (e.g., C-DAC in Pune, India has recently started production of the transputer-based PARAM computer), as the

user base expands, production technologies are licensed outside the principal countries, and international networks provide access and rapid communications across national boundaries.

Off-the-shelf ICs are not easily controlled. The microprocessors mentioned in the previous section have been manufactured in volumes of 100,000 or more, are small enough that tens or hundreds can be packed in a suitcase, are sufficiently self-contained that units are replaced rather than repaired, and are widely available outside the COCOM member countries. While not all of them, strictly speaking, can be considered commodities according to the criteria spelled out in [Schm91], they will be within very few years.

All of the massively parallel and conventional vector-pipeline use some customized components. Customized components are easier to control than off-the-shelf or industry standard ones. They are manufactured in smaller quantities, have limited distribution, and not easily substitutable. The ease with which one could acquire the hardware necessary to build a given machine varies with the degree to which customized components and subsystems are used. We examine three types of high-performance systems which are among the least controllable from the perspective of denying the capability to construct them. We consider other systems not discussed explicitly to be more difficult to construct than these.

A.2.2.1 Intel Parallel Systems

The Intel systems grew out of work on the Cosmic Cube at the California Institute of Technology in the 1980s, a system with a hypercube structure using the off-the-shelf 8086 and 8087 microprocessors as nodes. Intel's Supercomputer Systems Division was formed in 1984 to commercialize large-scale parallel computer systems based on an implementation of standard Intel microprocessors. Between 1985 and 1992, Intel intro-

duced three more generations of machines, based on the 286/287, 386/387, and i860 microprocessors, and currently has an installed base of over 300.

In order to keep manufacturing costs low and leverage the enormous amount of research done in the workstation and personal computer sectors of the computer industry, Intel has sought to use commercial and non-exotic technologies to the greatest extent possible. For example, the iPSC/860, introduced in 1990, is constructed using the i860 commercial microprocessors, commodity CMOS memory components, the 5.25" disks used in most workstations, as well as widely used I/O, networking, operating system, and computer language standards.

An exception to this principle is the direct-connect inter-node communications system based on the proprietary VLSI communications chips which route messages from one node to another. These chips were developed and manufactured by Intel based on designs by researchers at the California Institute of Technology. Of all the hardware used in the iPSC/860, these chips would be the most difficult to acquire.

It is, however, a mistake to assume that simply acquiring and assembling the hardware is sufficient to build a high-performance system. Actual performance depends directly on the efficiency with which system resources are managed and data are moved from one location to another. Intel has invested hundreds of man-years and millions of dollars researching the most appropriate ways of managing system resources, taking advantage of the computing power the hardware offers, decreasing software development time, and providing computational results in a useful form. Without the proprietary Concurrent File System (CFS), System Resource Manager (SRM), the NX/2 operating system, and other important pieces of systems software and firmware, the hardware is all but useless. The effort and know-how required to develop the necessary systems software should not be underestimated.

A.2.2.2 Transputer-based Systems

A second type of system worth examining is transputer-based systems. Transputers are microprocessors developed by INMOS Ltd. in England. Two important qualities are the ease with which they can be configured into systems of various sizes, and the clear and stable interface between hardware and software which enables software to run on a single transputer or a network of transputers without change or even re-compilation. Because communications between processes on the same transputer or on different transputers uses identical instructions (with the inter-transputer communications taken care of by the hardware), the precise configuration of the hardware is largely transparent to the software.

The core of the transputer market currently is fault tolerant systems and embedded controllers. Significant numbers are also used in large multiprocessors and accelerator boards which can be plugged into commercial personal computers and workstations. Specialized boards are being marketed for basic computation, video frame-grabbing, graphics, A/D conversion, and more. Some boards even incorporate an i860 processor.

Construction of such boards does not require highly sophisticated, proprietary technology. One leading vendor does board design using commercially available CAD systems, purchases commercially available components and subcontracts out the PCB manufacturing and system assembly. The technology to manufacture eight-layer boards used here does not have to be state-of-the-art, and is within the capabilities of Russian manufacturers. The assembly is a combination of surface-mount (automated) and through-hole (manual) processes, also within the capabilities of Russian manufacturers.

Multiple transputers can be placed on one board (up to 32 for a board to be used in a workstation). Boards are placed into racks which are attached to the workstations via an Sbus-VME converter card which also is commercially available. No customization of

software is required for such an installation. Standard transputer software—the operating system, compilers, a toolkit, etc.—is supplied by INMOS and the vendor serves solely as a distributor. Systems with several hundred transputers and theoretical peak performances of several hundred Mflops can be configured in a straightforward manner.

Because transputers, communicating via built-in serial links and running widely available systems software, can so easily be configured into multiprocessor configurations it is difficult to prevent configurations such as those mentioned above from being assembled. To disable the hardware, the serial links connecting transputers must be disabled. Since the communications facilities are built into the transputers themselves, this is impractical to do, short of physically isolating individual transputers. In multi-board configuration, it would be possible to manufacture specialized boards in which the communications links on the board could be terminated before they reach the edge connector. This would prevent transputers on different boards from communicating with each other, while permitting a modest amount of parallel processing on each board individually.

For large systems consisting of several hundreds or thousands of nodes, the most difficult challenge is not the construction of the hardware, but the development of the systems and applications software necessary to use it effectively. Network operating systems, debuggers, and performance monitoring systems are crucial, and hard to develop. Much time must be spent porting or developing applications. Not yet widely available, such sophisticated software is still reasonably controllable.

The transputers and their basic systems software are not easily controlled, however. Over one million transputers have been manufactured, and annual world-wide sales are over 250,000. SGS Thomson, the principal transputer distributor, has offices in India and the Pacific Rim countries as well as Western Europe and North America. Nearly two

dozen companies are value added resellers, building complete systems based on the transputer.

A.2.2.3 RS/6000 Clusters

The RS/6000 workstation clusters recently announced by IBM represent an alternative path to high performance computing based on commercially available technology. IBM has widely advertised the fact that a cluster of RS/6000 workstations supplanted a Cray X-MP supercomputer at Lawrence Livermore National Laboratory (LLNL). The hardware consists of standard workstations equipped with boards manufactured by IBM for Ethernet, Token Ring, or FDDI fiber optic networks, and the associated cabling.

The heart of the cluster is in the software. Currently, the cluster can be organized as a serial-batch, or a parallel system. In the first case each program runs on only one machine, but programs can be submitted to any available computer in the cluster. Supporting this mode are the Network Queuing System (NQS) or the DQS system, a public-domain system developed at Florida State University. Supporting the parallel execution of a single program across multiple machines are the Network Linda, Parasoft Express, and PVM (Parallel Virtual Machine) environments. The latter is public domain, developed at Oak Ridge National Laboratory.

The most controllable parts of an RS/6000 cluster are the workstations and the individual network boards which are IBM proprietary. Although not yet commodities, with a worldwide installed base approaching 100,000 the RS/6000 is not easily controlled.

As the technology advances, clusters of computers increasingly will be applied to a single job. Gordon Bell, the well-known former Vice President for R&D at DEC and founder of the Gordon Bell Prize for achievement in parallel computation, states that "Important gains in parallelism have come from software environments that allow networks of computers to be applied to a single job. Thus every laboratory containing fast

workstations has or will have its own supercomputer for highly parallel applications. The rapid increase in microprocessor power ensures that the workstation will perform at near super speed for sequential applications. LAN environments can provide significant supercomputing for highly parallel applications by 1995" [Bell92].

A.2.3 Controlling the Acquisition of HPC

The number of high-end HPC installations is still quite small. For example, Intel has shipped only 300-400 parallel systems; Convex has over 1100 systems in the field; and Cray Research Inc. has an installed base of just over 300 mainline computers plus nearly 100 Cray YMP-ELs. The small numbers of units, single sources, and considerable effort required to install them are among the factors which make it relatively easy to keep track of where each system is located.

From the discussion above we can see that it is possible to construct high-performance systems with a high percentage of generally available, off-the-shelf hardware which is difficult to control. However, the construction of almost all massively parallel and vector pipeline high-performance computers requires some customized hardware and/or software. These components and the technologies necessary to produce them are the best candidates for control efforts. Software design tools for application-specific integrated circuits are commercially available from over a dozen firms (see Smith, 1992), but the technology to manufacture chips based on the designs is still controllable and should be a high control priority.

In addition, the sophisticated systems software needed to make systems run effectively is still a reasonable control target. In spite of the ease with which software can be copied, it is a very difficult task to port it from one type of hardware platform to another. It is nearly impossible without access to source code. This is particularly true of most par-

allel systems. Proprietary and closely held, the source code is perhaps the most controllable part of such a machine.

Workstations, with global production rates approaching half a million annually from a dozen or more vendors, are not easy to control; neither will computing clusters based on them. While leading vendors are cooperating with government policies to limit direct sales of restricted technology effectively, the installed base is so large that controls are "leaky", at best. Large numbers of workstations can be obtained easily in the Far East. U.S.-based workstation manufacturers have set up factories abroad. Particularly notable are Hewlett-Packard and Silicon Graphics Inc. SGI's Iris Indigo workstation, introduced in the U.S. in 1991, will soon be manufactured in China; HP built a factory in Shanghai in 1990 to manufacture the Apollo 9000 series 400 workstations.

Workstations constitute perhaps the most rapidly evolving sector of the computer industry. With development cycles under a year in many cases, prices on given models drop rapidly following introduction. Users replace models after only a few years of use, creating a large secondary market which is uncontrollable. Developing countries like India appear to have little problem acquiring workstations.

Currently the principal barriers to workstations in Russia have little to do with export control regimes. First, few organizations can afford to purchase machines costing tens of thousands of dollars each. Second, the support infrastructure is poorly developed. Western workstations will run for months without maintenance, but they do fail, and failures are more difficult to diagnose and repair than is the case for PCs. As the technology becomes more available and the economic climate improves, these hindrances will ease.

A.3 A Framework for Confidence-Building Measures

We have examined the controlability of HPC technology to restricted countries. We now consider a three-track framework for confidence-building by which systems could be selectively installed and used in Russia. The three tracks are: application domains, institutional arrangements, and the means for controlling or monitoring HPC technologies. For each track, one can envision an evolution, conditional on continued cooperation and trustworthiness, from safer, more secure positions to those which involve greater risk of diversion. The possibilities discussed below are necessarily riskier than what has been permitted for HPC in the past. The tracks are loosely coupled in the sense that movement from more secure to riskier positions on each track can be made at different rates. This flexibility makes possible a wide range of possible confidence building sequences.

The framework does not assume that cooperation at any one level of society or government, or within any particular sector is sufficient for establishing confidence. Russia today is characterized by the decentralization and fragmentation of lines of authority. This creates both difficulties and opportunities, since governments are not as able to regulate the activities—for good or for ill—of individuals and organizations as they once were. The framework requires the cooperation of all individuals and organizations involved, from users up through national governments.

The success of confidence-building measures will depend to no small measure on the creation of incentives to cooperate for all parties involved. The main idea behind the sequence of confidence-building steps is that observing the agreements at the previous step will open new opportunities or capabilities for the next step. Inherent in the framework with its emphasis on a sequence of confidence-building steps is the notion that appropriate behavior today will be rewarded by increased opportunity or capability tomorrow.

A.3.1 Application Domains

The applicability of high-performance computing applications to military concerns varies considerably. Confidence-building measures should initially focus on applications which have little importance to the military and gradually move towards those which are marginally important.

According to [Gart91], the following are examples of applications with little direct military applicability:

- Design of pharmaceuticals through the simulation of proteins and molecules.
- Structural biology. The use of simulation and molecular dynamics methods to study the time-dependent behavior of biologically important macro molecules.
- Human genome project. Computer-assisted comparison of normal and pathological molecular sequences for understanding genomes and the basis for disease.
- Computational Ocean Sciences. The development of a global ocean prediction model.
- Astronomy. The processing of the large volumes of data generated by Very Large Array or Very Long Baseline Array radio telescopes.
- Quantum Chromodynamics (QCD). Simulation of QCD yield insight into the properties of strongly interacting elementary particles.
- Computational Chemistry. Simulation of molecules and chemical reactions are critical to the development of new materials.
- Financial applications. In the West, sophisticated econometric models and vast databases consume enormous amounts of computing power. The Russian financial infrastructure is still immature, but will, hopefully, strengthen.

- Commercial applications. Reservation systems, point-of-sales systems, etc. require fast access to large databases.

Other application domains having a greater, but indirect, relevance to military capability are crucial to economies in general, and the Russian economy in particular. These include:

- Transportation. Modeling of fluid and gas dynamics in three dimensions, such as the airflow around vehicles, fluid flows within engines.
- Superconductivity. Superconductivity can be a critical factor in future power transmission technologies, instrumentation. The basic properties of superconducting materials are not well understood.
- Efficiency of combustion. Studying the interplay between flows of various substances and the quantum chemistry principles governing the reactions between them.
- Oil and gas exploitation. Utilization of improved seismic analysis techniques and modeling the flow of fluids through geological structures.
- Nuclear fusion. Understanding the behavior of ionized gasses under high-temperature conditions with very strong magnetic fields.
- Prediction of weather, climate, and global change. Development and use of models regarding the interaction between atmosphere, ocean and biosphere system enabling long-range predictions.
- Engineering applications. The structural analysis of products.

Because of the potential military application of these areas, confidence-building measures here in the areas of institutional arrangements and control regimes might proceed more slowly than in the non-military cases listed earlier.

Nevertheless, because of their importance, the greatest efforts to make progress should, perhaps, be concentrated exactly in these areas.

Some application domains have direct and critical implications for both economic and military competitiveness. Efforts should be made to explore confidence-building measures here, but with greater caution than in other application domains.

- Material sciences. The understanding the atomic nature of materials and the development of new kinds of materials.
- Semiconductor design. The modeling of how semiconductors constructed out of faster materials operate.
- Vehicle dynamics. The analysis of the aeroelastic behavior of vehicles and their stability and ride characteristics.

Finally, application domains which have great military importance, but marginal economic importance. There is little reason to seek confidence-building measures in, for example:

- Vehicle signature. The reduction of acoustic, electromagnetic, and thermal characteristics of vehicles.
- Undersea surveillance. Tracking undersea vehicles.
- Cryptography.

Movement from safer to riskier applications could take place on one machine, as the set of allowable applications grows, or in multiple installations. In the latter case, one installation might be devoted to a safe application, while in subsequent ones riskier applications might be allowed. One of the problems that has plagued this approach, and that of permitting remote access to a machine under physical U.S. control, in the past is the difficulty of carefully monitoring and distinguishing between applications as they are run.

A.3.2 Institutional Arrangements

The technical composition and technological content of an installation influence the degree to which it might be diverted. In general, installations can vary in the scope of use, i.e. the spectrum of applications, the size and composition of the user community, the degree of openness about systems use, and the physical distribution of the hardware. In addition, one can categorize installations according to whether they are managed and used by people from one country or from several. The risk of diversion increases as the scope of use increases, the hardware becomes geographically distributed, and the management of the installation becomes more closed, private, and under the control of just one country. If Russians and non-Russians are working together on the same system, it is less likely that sensitive military applications will be run.

Least subject to diversion would be government-run facilities physically located in the U.S. or another NATO country, where citizens of other countries would be permitted either on-site or remote access. The risk of reverse engineering could be kept minimal, and the risk of diversion during times of international conflict would be essentially eliminated.

A higher risk might well be suitably controlled at public, centralized, tightly managed, international computing centers sponsored and managed by individual governments or international agencies such as the United Nations for the purpose of providing advanced computing resources for non-military research in specific application domains. Over time, the center could expand the user community, possibly offering time to the international community on a competitive basis. Researchers might submit detailed proposals for systems use; individual projects, selected on the basis of appropriateness to the center's mission, could be run under the supervision of the center's staff. Fundamental to

such an arrangement would be the on-going surveillance of system activities by the international community.

International installations could also be created within private joint ventures. Companies which routinely use high-performance computing such as Boeing, Chevron, and Sun Microsystems have established joint ventures in aerospace, oil and gas exploration, and computing. These companies could work together with their Russian partners to ensure non-diversion of imported advanced technologies. To a large extent, the success of the export control regimes in the past has been due to self-policing by Western companies; reputable firms have refused to deal with suspect customers because of the possible legal, financial, and negative publicity consequences of illegal transactions. This principle should be applicable in the the case of joint ventures as well. Western partners could be given permission to import high-performance computing technologies for use in the joint venture with the understanding that they will be held responsible for any diversion of the technology.

Centers under the management of a single country could be established at state-owned institutions such as universities and government research facilities and at new or newly privatized corporations. The opportunities for diversion would be less at facilities operating with non-proprietary data and applications, where activity could be monitored by a broader circle of observers. It is not clear *a priori*, however, whether a government organization is to be preferred over a private firm. On the one hand, the Russian government could be enlisted as a partner in preventing diversion; other the other hand, a private firm involved in non-military commercial activities would likely have weaker ties to the military. Questions such as these would likely have to be answered on an organization by organization basis.

Fundamental to the success of any of these scenarios is the establishment of mutually beneficial collaborative efforts using high-performance computing between Russian and Western researchers. The degree of commitment to the relationship (and, correspondingly, the willingness to avoid actions which threaten it) will be a function of the longevity of the relationship, the promise of future benefit, and the importance of the efforts to individual researchers as well as industrial or scientific sectors and the country as a whole.

Russia has rich and extensive pools of data in many branches of science listed in the previous section. In many cases these have suffered from inadequate computing facilities to process and analyze the data properly. A fruitful area of cooperation would be the application of Western computing technology to this data. Cooperation would be more closely knit and longer-term as researchers work together to conduct studies and experiments which generate new data as well.

It is important to note that collaborative work can begin prior to hardware installation in Russia. Russian researchers could literally bring data tapes to the West for processing. Alternatively, such data could be sent electronically to Western machines monitored by both countries.

Institutional arrangements also include government-level monitoring and control mechanisms. In the past, the Soviet government was able to exert effective control over the use and distribution of sensitive technologies through strong military, Party, and State Security structures. Each of them had strong military components and worked to a large extent, at least in the area of high-technology, on behalf of military interests. If Russia is to work together with Western countries to control diversion and proliferation, an effective civilian mechanism must be established which will not only exert control in the cases

of individual installations, but also provided continuity and consistency of control from one installation to another and over time.

No such mechanism currently exists with regard to high performance computing. There is doubt in the West that such a mechanism could perform an effective job under Russia's current economic and political conditions. First, the traditional pillars of Soviet society have partially lost their ability to control as a result of decentralization measures and the unregulated activities of powerful groups such as organized crime. Second, the dire economic straits are forcing individuals and organizations at all levels of society to skirt regulations merely to survive. The sale of advanced technology for hard currency to unauthorized customers is by no means inconceivable. Third, the Russian government has stated that it intends to maintain significant military capability and the ability to re-convert enterprises to military production if necessary.

Under these circumstances it is difficult to envision a civilian authority which could effectively control the diversion and proliferation of high performance computing technologies, even though in principle COCOM-like structures and procedures could be established within Russia as they have been within Hungary. Yet such structures must exist and be effective if Russia is to be a partner with the West in this area in the longer term. It is incumbent upon the Russians to design such structures and convince the West that they are effective.

A.3.3 Technologies and Measures for Control and Monitoring

Given any combination of application domain and institutional arrangement, a variety of control measures can be implemented to regulate and monitor system activities.

"Hard" controls are those which seek to prevent diversion by making it difficult to carry them out. Measures which physically or logically control access to a given computer or otherwise restrict performance are of this nature. "Soft" controls, on the other hand, are

designed to detect violations, rather than prevent their occurrence. Confidence building measures involve a series of steps which increasingly reduce first hard controls and then soft controls. Soft controls in most cases will have to be in place longer than hard controls to provide objective verification that violations have not occurred.

A.3.3.1 Hard Controls

Hard controls seek actively to limit what can be accomplished on a computer and by whom. The Supercomputer Safeguards Plan (SSP) (Export Administration Regulations 15 CFR 776.11(f)(4)) places very stringent hard controls on systems with Composite Theoretical Performance (CTP) equal to or exceeding 195 million theoretical operations per second (MTOPS). The measures are designed to prevent unauthorized use through denial of physical access to systems by restricted nationals, strict control over the issuing of passwords, precise selection of which applications can be run and under what conditions, complete lack of connection to networks or remote terminals, etc.

While these restrictions do and will continue to accomplish the goal of controlling access, at a time when we are examining alternatives to such restrictions it is important to keep in mind that access is a necessary, but not sufficient condition for performing useful work. The performance and usefulness of a computer are dependent on many things, only some of which are taken into account in the computation of the CTP. Performance depends not only on the raw processing rate of individual processors, but also on the amount of memory available, the throughput of the interconnect system, the amount and speed of external memory, and the throughput of the I/O system. Unless processors are supplied with data at a high enough rate, they sit idle and accomplish no useful work. Software also plays a critical role. The overhead of the operating system, the efficiency with which it manages systems resources, and the effectiveness of the compiler can have a significant impact on performance. Instances in which the performance of the same

program on the same hardware is increased 100% or more, simply through the use of an improved compiler, are not uncommon.

In a real-world setting, human factors—the ease with which a user can accomplish a desired task—play a crucial role in determining a system’s usefulness. The amount of time spent programming and debugging, and the time needed to analyze and interpret results strongly influence the utility of the machine to the user.

Each of these factors provides a means of regulating the effective performance and usefulness of a system. If a system is installed with insufficient external storage, a non-mature software development environment, a lack of sophisticated applications, inadequate tools to support the visualization and interpretation of results, or is used in an environment in which the ratio of software development to execution is high, the true performance indicated by the CTP will not be realized.

As confidence is built, given installations can be enhanced by selectively relaxing the constraints just mentioned. Processing elements can be added, more external or main memory can be installed, upgraded software packages can be provided, the number of applications authorized for execution can be increased, etc. At each installation, the prospect of future upgrades provides an incentive to cooperate.

A.3.3.2 Soft Controls

Soft controls make it possible to monitor the use of a system without necessarily preventing unauthorized use. The Supercomputer Safeguards Plan requires extensive soft controls to be used in conjunction with the hard controls. These include maintaining in a secure fashion usage logs and inspecting them daily, detecting attempts to gain unauthorized access, recording execution characteristics of each program run, and the monitoring of CPU and I/O usage.

Soft controls also serve as guards against proliferation, since the physical location of a system can be easily determined.

Some sort of soft control should be used until a high level of confidence in adherence to non-diversion agreements has been reached.

A.4 Recommendations

The recommendations in this section augment those presented in the "Joint Statement of the Delegations of the RAS and NAS on Dual Use technologies and Export Administration" and should be considered within the context of the latter document.

High-performance computing technologies are evolving very rapidly, particularly in the workstation arena where new generations are introduced every 2-3 years and equipment five years old is often considered obsolete. Recommendations for the control or de-control of specific technologies are similarly quickly outdated.

- **Recommendation #1: Significantly reduce controls on technologies of which 100,000 units or more have been sold, unless there are compelling reasons to the contrary.**

Currently, microprocessors such as the i860 and T800 and many workstations fall into this category. While not necessarily commodities in the strict sense, such technologies are so widely available that control measures are very "leaky" at best.

From an economic perspective, the greatest benefit to American industry will come through the sale of large-volume products. In the West, the total size of the workstation market is an order of magnitude larger than the supercomputer market; the personal computer market is many times larger than the workstation market. In general and in Russian in particular it is much easier to sell one hundred \$10,000. units than one \$1 million unit.

With a threshold of 100,000 units, great economic gains can be made without severely compromising national security.

A basic premise of this paper is that Russians should be able to participate with Western countries in regulating the diffusion and use of high-performance computing systems, and that they should be given the opportunity to demonstrate their willingness and ability to do so. One means of accomplishing this is through the use of carefully selected sequences of confidence-building measures. Ideally, such sequences would serve as a testing ground for a variety of Russian, Western, and combined control measures, and serve as a model which could in the future be replicated. An additional benefit would be the placement of technology in Russia which could help stem the drain of computational scientists from Russia. But it is critical that the object of export control review be an entire sequence of steps, rather than an isolated installation.

- **Recommendation #2: Consider plans for the installation of individual pieces of technology within the context of a series of measures, possibly leading up to the approval of otherwise restricted technology, conditional on compliance with prior agreements.**
- **Recommendation #3: Give favorable consideration to a number of test-case sequences of confidence-building measures.**

We offer the following sequence as an example. Russian scientists frequently claim that they have developed methods of solving a variety of computational problems which are better in some sense than those developed in the West. As their contribution, in the interests of mutual cooperation in advanced high-performance computing technologies, the Russian scientists can adapt these methods to Western machines. At the first stage of a joint project, a team of Russians would undergo training at a Western university in software development for a particular Western massively parallel system. At the second

stage, the Russian team would implement their algorithms, developing programs to run on the parallel machine. This could be carried out in Russia on workstations with the appropriate software development tools. At the third stage, the Russian team would work on debugging and tuning their algorithms in concert with Western colleagues on the Western machine. At the fourth stage, a small configuration would be installed in Russia under the joint supervision of the Russian and American researchers, and Russian and Western export control administrations. Each subsequent year, as long as non-diversion agreements are not violated, the installation would be upgraded through adding more processing elements, memory, external storage, software, etc.

A second example could be oriented towards the creation of a prominent computer center which would provide computer time to individuals conducting civilian research in a variety of application domains. At the first stage, a low-end, general-purpose machine from a leading Western supercomputer manufacturer could be installed at a prominent Russian university or Academy of Sciences computer center under the exclusive control of representatives of Western export control organizations and the computer's manufacturer. At this stage the system could be used to run Western applications, or specifically approved Russian applications.

At a second stage, a set of research projects, conducted jointly by collaborating Western and Russian colleagues, would be selected and granted access to the machine. An international commission could be established with the task of guaranteeing its appropriate use. Crucial to the composition of this commission would be the full participation of the principal researchers using the system. Additional members would include a representative of the computer vendor, a representative of a Russian monitoring agency, and a representative from the Western export control establishment. Having both Russian and Western researchers involved would ensure that the commission contained the expertise

necessary to understand the applications being run. The arrangement would rely for its success on the personal relationships and interests of the researchers, and the personal stake each has in ensuring an enduring, successful collaboration.

At a third stage, the set of users and applications could be selectively widened. The international commission would retain a permanent core, with pairs of Western and Russian researchers participating for the duration of their projects.

At subsequent stages, the center could evolve in a number of different directions. The installation itself could be upgraded; the Russian could be given greater and greater monitoring responsibilities; the requirement that all projects be collaborations between Russian and Western colleagues could be removed; the center could be made available for a broader circle of users and/or applications, including deserving university students.

This second example assumes that successful use of an installation must be based on participants from the individual researchers up through the national government having a strong interest in guarding the system against inappropriate use. Although the existence of a Russian governmental structure with oversight over export control and the use of imported high-performance technology is not a sufficient condition, it is necessary.

- **Recommendation #4: Evaluate a variety of "soft" controls, or means of verification of the end-use of high-performance computer technology as a part of a sequence of confidence-building measures.**

The confidence-building measures will lead to fewer iron-clad controls over the use of particular systems, but means of verification of use should be kept in place until sufficiently high levels of trust have been established, or technological developments make them unnecessary or impractical. Computer systems can store detailed logs about certain aspects of computer usage, such as which programs are being used by whom for how long, patterns of system resource usage by individual programs etc. Although such infor-

mation is not sufficient to identify the higher-level problem being solved by a particular program, it is very useful in giving a general idea of how a system is being used. Initially, such information would be gathered by Western systems managers on location. At later stages, such information could be gathered and transmitted automatically through satellite or other communications links to individuals monitoring the system. This would provide a relatively unobtrusive form of soft control.