CHAPTER 9. WHAT IS TO BE DONE?

If the Soviet states are to occupy leading positions in science and the international advanced technology industries, their scientists and engineers must have access to high-performance computing far beyond what they now have. Functionality, cost, and time-to-market are critical competitive variables. Western automotive, aerospace, computing, and other industries rely heavily on high-performance computers to improve each of them. Widespread access to powerful computing facilities by the scientific community is today a prerequisite for broad-based advance in many spheres of research.

To date, the Soviet high-performance computing sector has not delivered nearly the computing resources needed by science and industry. Furthermore, the sector is at a critical stage, fighting for survival. Production of high-performance systems has all but stopped. There are many potential users for HPC, but few who are willing and able to purchase HPC systems, particularly indigenous ones. Competition from Western computers is increasing. Funding continues for major industrial projects nearing completion, but at levels which are inadequate to do much more than keep development teams together. Funding for new large-scale systems has not materialized. The leading developers of HPC, ITMVT and NIIUVM, are struggling to maintain an organizational structure and expertise which can continue HPC development in the future. Under these conditions, Can Soviet HPC survive? If so, in what form? What are some options for giving the scientific and industrial computing communities access to needed high-performance computational resources?

9.1 Implications for HPC developers

Without funding, there will be no Soviet HPC sector. Without a market, it will only struggle, and not become a world-class player. There are three basic sources of revenue: direct government funding, revenues generated from internal activities, i.e., the sale of finished products, product re-sale, contract work, etc., and foreign or domestic private investment. Until the economy as a whole improves to the point where the government and industry have resources which could be devoted to high-performance computers, the opportunities for funding to support the development of complete large-scale systems are very limited.

The government is currently not willing to fund new projects on a large scale; industry does not have the resources (or is not willing to spend them) to purchase Soviet high-performance computers, let alone fund expensive R&D programs. The R&D facilities themselves are ill-equipped to earn revenue from the sale of finished products manufactured by factories. We have discussed how contract work is not likely to generate large income streams. Foreign investment such as that provided by Sun Microsystems will serve a useful purpose in keeping Soviet engineers engaged in work which will hone and advance their skills. However, foreign firms investing in Soviet R&D will be interested in research results or the development of portions of systems, and not the development of complete systems. They may provide tools (e.g. workstations and CAD systems) which are useful for other development, but funding for other projects will have to come from elsewhere. Income may be generated through strictly commercial activities if R&D facilities depart from their traditional domains and serve as re-sellers of Western equipment. Even if this is highly profitable, how willing are organizations going to be to invest the proceeds in high-performance computing development if there is no market?

A consequence of decreasing revenue stream and more highly focused small scale funding is that the HPC sector will not be able to carry out large-scale development projects which support R&D across the entire infrastructure of upstream industries. The types of projects undertaken in the future are likely to smaller scale, relying less on new developments in other industries or industrial sectors.

To be healthy, the Soviet HPC sector must have a market. Obviously, a market provides sources of revenue and should provide a demanding customer base which helps shape the direction of technological advance. If the economy does improve, or national priorities shift, it is possible that the government will fund Soviet HPC at higher levels than it does now. Such funding by no means guarantees that the end results will be competitive on either the foreign or domestic markets, however. As a principal customer, the government has helped individual companies survive; but direct government funding has not had a good track record of producing commercially viable products, either in the West or in the Soviet Union. If Soviet high-performance computers are not competitive, the market will be limited to those organizations which, for foreign and domestic national security reasons, are forbidden to use non-domestic advanced technologies. Such a market is smaller now than it once was, and will not lead to a vibrant HPC sector.

Regardless of the level of government funding, the Soviet HPC sector will, for its own health, have to try to develop a commercial presence. To do this, R&D and production must be united so that the proceeds from the sale of finished products can be reinvested in research and development. This will not be accomplished easily, however. As decentralization and privatization progress, providing R&D facilities with large-scale production facilities will have to be accomplished by acquisition, construction, or strategic alliance. The first two require funds currently unavailable; the latter is unlikely in light of the poor market for HPC products. It would be easier for factories to hire individuals from existing R&D institutes, but this will only happen if the factories determine that HPC is a better investment than other, less advanced, technologies.

Assuming the economy improves and funding from some quarter is available for R&D, how can Soviet HPC become competitive? By their nature, high-performance computers of all types and sizes depend on a large number of highly sophisticated upstream industries. Although some of the component products may be considered ubiquitous, such as certain types of microprocessors in the West, they still require advanced R&D and manufacturing capabilities.

Our study has shown how strongly the nature of the Soviet economic system and the technologies available from upstream industries have influenced Soviet HPC. Although in some respects it is currently easier for HPC developers to interact with other organizations and suppliers, the traditional infrastructure is not yet well suited to support HPC development.

It is critical that the Soviet HPC sector reduce its dependence on indigenous industries. Although cultivating a complete infrastructure for HPC was necessary in the past, the Soviet HPC sector cannot afford to serve as the principal driving force for upstream industries. It also must take advantage of advances in computer technologies as they occur elsewhere in the world. This is not to say that current efforts to keep leading teams in all facets of computer technology employed should be terminated or that the HPC sector should not use indigenous products when they are of acceptable functionality, quality, and quantity. Such efforts have a role in maintaining expertise and should be continued. It is the case, however, that if the HPC sector is forced to rely on inadequate or immature domestic technologies, the resulting HPC system will have long development times and will not be internationally competitive. Developers, in industry and academia, will not be able to enjoy the benefits of rapid iterations in computer design and development. The HPC sector cannot afford to wait for the domestic industries to de-monopolize and rise to world-class levels.

Incorporating Western technology is no panacea either. Most advanced systems require at least some customized components. Establishing the necessary business relations with Western manufactures will be problematic for financial, geographic, and political reasons, although less so now than during the Cold War era. Even acquiring off-the-shelf components requires much more effort and expense for a Moscow-based organization than one located in Silicon Valley. We have discussed the additional complications for organizations such as NIIUVM in Ukraine who have only indirect access through Russia to the Western currency needed to purchase Western goods.

The basic point, however, is that the greater the use of advanced, off-the-shelf technology–either Soviet or Western–the easier it will be to build advanced systems in a timely, competitive manner. This will be the easiest way for the Soviet HPC sector to ride the wave of technological advance in the international community.

Such a strategy has implications for the types of systems developed. One cannot build a system directly comparable to a Cray using off-the-shelf components. Soviet industry will probably have to relinquish the objective of competing with such machines for the time being. Recent Western experience shows that high performance can be obtained through high levels of parallelism using non-exotic technologies. For example, Intel's Supercomputer Systems Division has built multiple generations of systems based on off-the-shelf Intel processors such as the 286/287, 386/386, and i860 microprocessors. Furthermore, Intel has tried to incorporate commercial technologies to the highest degree possible, using commodity CMOS memory components, the 5.25" hard-disks used in most workstations, as well as widely used I/O, networking, operating systems, and computer language standards and technologies. Transputer-based systems can also be constructed from readily available technology. Transputers, communicating via built-in serial links and running widely available software, can be configured into multiprocessor configurations relatively easily, as demonstrated by the many value-added resellers of transputers throughout the world.

The reasoning driving the development of massively parallel HPC in the West is essentially the same as that underlying a key element of the Soviet HPC paradigm: parallel processing is a viable route to high performance. For this reason, developing massively parallel systems based on Western technology would not be a severe philosophical departure for many Soviet groups involved in HPC. Unlike those of the Soviet industry, however, the components used in Western parallel systems, even though commercial, are very close to the state-of-the-art.

Using existing technology will have a significant impact on system design. No longer will developers have the freedom to create exotic processor architectures or memory modules. Many issues of interconnect structure and construction will be determined, or at least bounded, by the nature of the technology available. Should this strategy be pursued, it is likely that we will see a confluence of the approaches taken to hardware design.

The major challenge in designing systems along these lines is not the hardware, however, but the systems and applications software needed to use it effectively. Companies like Intel have spent millions of dollars and hundreds of man-years researching the most appropriate way 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 sophisticated (and often proprietary) systems software, the hardware is all but useless.

Fortunately, software development plays to Soviet strengths (or, plays less to their weaknesses) than hardware development. Although software development does require a basic parallel hardware platform to be tested realistically, it does not require nearly the network of supporting industries and products that hardware development does. Adequate software development platforms are readily available for commercial personal computers and workstation. A great deal of development and simulation can be done by engineers with these tools alone.

The effort required to develop software to make the hardware discussed above functional should not be underestimated. Given the tools and intellectual capacity currently available in the Soviet states, however, it is possible to make the effort.

Soviets often claim that they have been able to coax great utility out of weak hardware through the use of more sophisticated algorithms and models. It has not been a goal of this study to confirm or disprove this claim. If it is true, however, parallel systems like those just described should offer an excellent opportunity to capitalize on this strength. Advances in massively parallel systems are now less a hardware issue than a software and algorithms issue. This fact offers Soviet engineers one of the best opportunities to make their systems internationally competitive. It remains to be seen whether they can capitalize on it.

9.2 Implications for HPC Users

The issue of accessibility depends on at least two factors: the availability of systems, and the number and type of individuals who are granted access to them. In the past, most high performance computers like the El'brus and PS- series have been installed either at closed military sites, or at non-military sites with specific, intensive data processing tasks, such as seismic data processing. In either case, the machines have not been generally available for the scientific community at large. The latter has had to work on outdated BESM-6 style machines or ES mainframes. The goal of using the El'brus systems at the center of extensive collective-use computing centers did not materialize.

Two recent trends could help reverse this state of affairs. First, influential individuals like G. G. Ryabov, the director of ITMVT, have expressed a strong design to make machines available to the broader research community through the creation of computer centers based on machines like the MKP. In this manner, several organizations could pool resources to acquire a powerful system, reducing the cost to each. Although there are few indigenous machines which could serve this purpose and the financial state of most organizations (including the government) is so precarious that establishing a center would be difficult, the intent is a good one, and is likely to serve the scientific community well in the long-term.

Second, the central planning organizations no longer dictate which organizations will acquire which machines. One of the reasons so few civilian organizations had El'brus computers was that those that had been built were taken by privileged customers. It is now possible in principle for any organization with the necessary resources to purchase Western technology, provided it does not fall under CoCom export control restrictions, and sometimes even if it does. The issue is no longer one of allocation priorities.

Of course, such a center requires an extensive telecommunications infrastructure to be widely accessible, but significant efforts are underway in both the private and public sector to improve telephone lines, establish satellite links, lay fiber optic cable, etc. This access will probably not be a major barrier to access for long.

The issue of hardware availability remains problematic. Soviet HPC users have not been well served in the past. The short-term prospects for acquiring systems from the indigenous HPC sector are very poor. What are the prospects for acquiring Western systems?

Although individual Soviet users have been able to accomplish much in weapons design, real-time mission control, etc., the CoCom export control regime has been very effective in denying Soviets access to high-performance computing, or at least making acquisition and use of foreign systems very difficult. The Soviet HPC user community has been hurt much more than the development sector by export control regulations.

Because of recent advances in technologies, some opportunities will be available to Soviets regardless of the export control regulations. For example, powerful workstations are now being manufactured in quantities of half a million or more. Thanks to public domain software, these can be clustered together to execute a variety of jobs or portions of jobs in parallel. Although the peak performance for a single task may not be at supercomputer levels, such arrangements greatly increase system throughput.

Although the Cold War and relationships between the Western countries and those of the former Soviet Union have improved dramatically, significant national security concerns remain. The West has a legitimate concern that its high-performance systems will be diverted to military use, or, through lack of effective proliferation control efforts, finds their way into the hands of restricted third countries.

At the same time, there is a widespread desire in the West to modify the export control regulations to take into account a) the realities of global advances in computer technologies and b) the possibilities that the Soviets themselves can become part of the export control solution rather than the problem. Computer technologies have advanced so quickly and are manufactured in such quantities that efforts to control the diffusion of many types of mass-produced technologies are futile. If the Soviets can demonstrate their ability and to work together with Western governments, vendors, and users to establish a control regime which will keep sophisticated technologies from being diverted to military use 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 the Soviet transition to a market economy. The issues of export control of dual-use technologies, including HPC, has been taken up in a number of forums, including [Good88; Schm91; Good93; Nrc93].

The question is how to build sufficient trust that formerly restricted technologies can be exported to the Soviet states. This issue is addressed in [Good93]:

> 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. [Soviets] must be given the opportunity to demonstrate understanding of and respect for the national security concerns of the United States and cooperate in preventing diversion and proliferation of sophisticated technology.

This document proposed a three-track framework for confidence-building in which systems could be selectively installed and used in the Soviet states. 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 tracks are loosely coupled in the sense that movement from a more secure to riskier positions on each track can be made at different rates. This flexibility makes a wide range of potential confidence-building sequences possible [Good93].

Successful confidence-building will not be easy. While there is reason to believe that the willingness of many Soviets to cooperate has increased, the government's ability to control has decreased. This means that it is not sufficient to rely on government assurances alone; individual users and organizations must have a vested interest in cooperation as well.

There are numerous possible confidence-building sequences. Two are suggested in [Good93]. Russian scientists frequently claim that they have excellent algorithms. A joint project could be established to enable them to implement these algorithms on Western hardware platforms. At a first stage, a team of Russians could undergo training at a Western university in software development for a particular Western massively parallel system. At stage two, the Russians could implement their algorithms, developing programs to run on the parallel machine. This could take place in Russia on workstations with the appropriate software development environment. At stage three, the Russian team could work on debugging and tuning their algorithms together with Western colleagues on the Western machine. At stage four, a small configuration could be installed in Russia under the joint supervision of Russian and American researchers and Russian and Western export control administrations. As long as the Russians comply with all associated agreements, the installation could be upgraded annually, increasing the number of processors, the amount of main memory, etc.

A second example could be the creation of a 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 would run Western applications, or specifically approved Russian applications. At stage two, 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. Such a commission would have to include 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. 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 stage three, 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 Russians 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 to a broader circle of users and/or applications, including deserving university students.

There are many more possible examples. In all cases, however, if they would join the international computing community, Soviet users will have to work together with the West to demonstrate an understanding of legitimate Western national security concerns and their willingness and ability to address those concerns.

9.3 Implications for Policy Makers

It is beyond the scope of this study to make the argument that it is in the West's best interests to draw the Soviets into the international community. If this argument is accepted then in light of the above discussion a number of changes to policy should be made [Good93]. Specifically,

- Controls on technologies of which 100,000 units or more have been sold should be significantly reduced, unless there are compelling reasons to the contrary.

Regardless of one's view of the geo-political aspirations of the Soviet states, the reality of export controls is that technologies produced in this volume are very difficult to control. From an economic perspective, it is in the interests of American industry to permit the sale of such products to the Soviet states. The market for workstations and chips which meet this criterion is much larger in dollars than that for high-end machines. In both the East and the West it is easier to sell one hundred $10,000. systems than one $1 million unit.

- Western export control administrations should 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.
- Western export control administrations should give favorable consideration to a number of test-case sequences of confidence-building measures.
- Soviet administrations should establish export control regimes which satisfy the legitimate security concerns of Western nations and demonstrate their continued effectiveness.
- A variety of ‘‘soft’’ controls, or means of verification of the end-use of high-performance computer technology should be evaluated as a part of a sequence of confidence-building measures.

A variety of measures can be used to monitor the use of individual systems. These include, but are not limited to, using the computers to store detailed logs about certain aspects of computers usage such as which programs are being used by whom for how long, patterns of system resource usage by individual programs, etc. Although such information is not sufficient to identify the higher-level problems being solved by a particular program, it is very useful in giving a general idea of how a system is being used.

9.4 Directions of Future Research

This study has concentrated on high-performance computing technologies and their associated R&D facilities. Although we have provided some information about the broader context, we have not presented a complete picture of all elements in the HPC sector. In particular, our discussion of the policy issues surrounding HPC has been sparse. Over the years, extensive discussions about HPC have been held at the highest levels of the Academy of Sciences, the State Committee on Science and Technology (the Council on Supercomputing in particular), the industrial ministries, and the Military-Industrial Commission, whose function was to coordinate activities in areas of strategic importance of the various ministries. The latter controlled most of the financing and production facilities needed to bring HPC research to fruition. A study of these discussions could reveal important information about how decisions were made to fund one line of development and not another, which external events (e.g. advances in Western technology, the U.S. Strategic Defense Initiative, etc.) had the greatest impact on HPC policy and why, and how the views of policy-makers served as a selection environment for individual research teams.

Similarly, a study of Soviet HPC would not be complete without studies of the production facilities, upstream industries, and users. Studies of these groups of players could contribute important information about how Soviet HPC developed in the past and, more importantly, how it might develop in the future.

The Soviet states have not completed their transition from the social, economic, and political systems of the Soviet Union. Neither have the R&D facilities discussed in this study. While we have observed certain trends, the complete process of transition will only become apparent in some years when conditions have stabilized. A number of Western researchers have examined issues of organizational transformation, concentrating specifically on the path from one organizational form to the next. For example, Hinings and Greenwood discuss organizational archetypes and categorize paths by which organizations make a transition from one archetype to another [Hini88; Gree88]. Gemmill and Smith and others discuss the application of dissipative structures theory to the area of organizational transformation [Gemm85; Leif89]. Our study offers initial data on the transformation of selected organizations. Continued study will provide valuable data with which to evaluate or improve the theories of organizational transformation.

We have discussed the ability of the Soviets to develop advanced technology within one industrial sector. Further study is needed to determine how generalizable our conclusions are to other advanced technology sectors, and to prosaic technology sectors as well.

We have postulated some possible future directions for the Soviet HPC sector. Will it survive these turbulent years? In what form will it emerge? What role will this sector play in the international community? In some respects, the HPC sector is at the edge of a new era. The old HPC projects are dying, but new ones have only begun to emerge. Although there have been some encouraging changes, dawn has not yet arrived; we will have to wait to see what the new day will bring.