

## **CHAPTER 8. CONCLUSIONS**

### **8.1 Introduction**

In this chapter we review some of the overarching features of Soviet high-performance computing and cross-cutting themes of our study. Soviet high-performance computing has a rich and complex history, and the dynamic of technological and organizational development has been shaped by inter-related factors: elements of an organization's environment, levels of technological availability and organizational slack, belief systems and research strategies, and, not least, by the technologies and organizational structures themselves. Following a discussion of the contribution of the Soviet HPC sector to the scientific computing community, we examine the factors which have shaped Soviet HPC development in the past, and the impact the reform process has had on the HPC systems and the ability to development them. We will also discuss the impact of the reforms on organizational structures and the implications for HPC R&D capability.

In the next chapter, we will discuss the prospects for developers and users of Soviet high-performance computing, policy issues for Russian and Western policy-makers, and avenues for further research.

### **8.2 The Provision of HPC Capability To the Scientific Community**

What has been the contribution of the Soviet high-performance computing sector to the Soviet scientific community? For all the years of research and resources invested in this sector, the amount of computing power provided has been disappointing. The most advanced systems have been characterized by extremely long (over 10 years) development cycles, plagued by reliability problems, and manufactured in only moderate numbers.

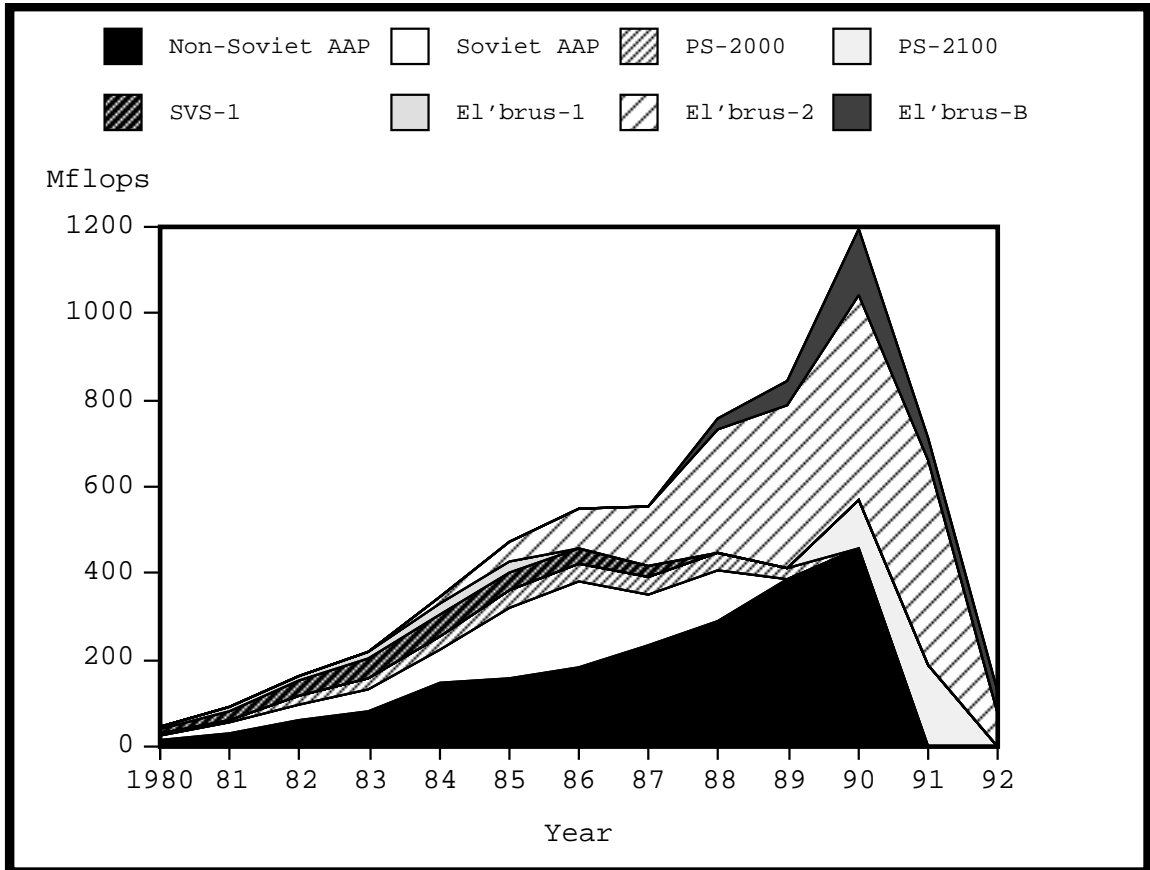


Figure 8-1 Annual Output of Series Produced Soviet HPC (in Mflops)

Figure 8-1 represents the annual contribution from 1980 through 1992 of the Soviet high-performance sector in terms of Mflops delivered by machines in series production. The specific annual figures are estimates, based on data of the total number of machines manufactured and the years of production. The graph is intended only to show in a rough way the amount of computing power delivered to the military and civilian scientific communities by the HPC sector. It does not reflect the utility of individual types of systems. In particular, the graph reflects performance on single precision or double precision (in the case of 24- and 32-bit machines) floating-point operations. The utility of machines such as the PS-2000 is under-represented, since while this machine had poor performance

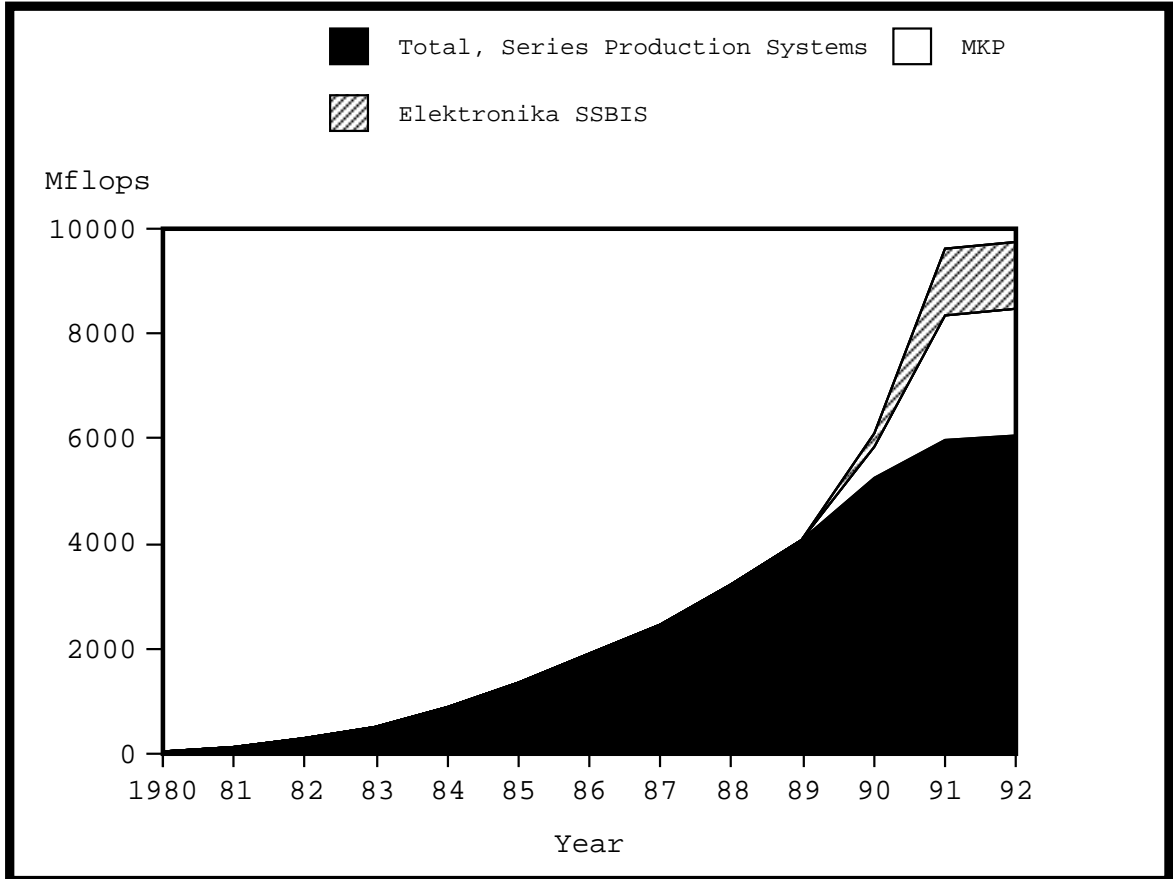


Figure 8-2 Cumulative Output of Soviet HPC (in Mflops)

on 48-bit floating-point operations, it had quite good performance on 24-bit fixed-point operations. Many applications required only the latter. The graph does not reflect the ease or difficulty of programming the system. The attached array processors cannot be considered general-purpose, since they executed only a limited set of library functions. In contrast, the El'brus computers had good performance on a much broader spectrum of applications.

Figure 8-2 shows the cumulative computing power, also measured in Mflops, delivered by the Soviet HPC sector. Two systems which could, in principle, dramatically increase the amount of computing power reached the prototype stage during these years.



Figure 8-3 Comparison of Annual Output, in Gigaflops, of Cray Research, Inc. and Soviet HPC sector

Source:[Estimates, based on CRI annual reports]

They were not included in figure 8-1 because they have not yet entered series production. It is unlikely that they will. Figure 8-3 compares the aggregate computing power delivered by the Soviet HPC sector with that delivered by Cray Research, Inc. during the same period. Those systems which have been manufactured to date lack customers. In the absence of a market, more will not be manufactured.

Several items in the preceding graphs are noteworthy. First, the “computer gap” between East and West has not been understated. The argument is frequently made that the Soviets had sufficient computing power for certain high priority applications such as nuclear weapons design and space mission control. It may well be the case that in selected

applications, and given a strong emphasis on the development of powerful algorithms and models, the computing power available was sufficient to meet certain objectives. However, in an age when computational methods have taken a place alongside experimental and analytical methods as a major tool of scientific research, scientific advance in a broad spectrum of research domains depends on the ability of large numbers of scientists to access advanced computing facilities. The Soviet scientific community has suffered considerably because of this lack. In chapter 9 we will discuss the options available to this community for gaining access to advanced computers.

Second, although the Soviet Union was the only Eastern Bloc country to have a serious high-performance computing sector, the attached array processors manufactured in Bulgaria and East Germany made a major contribution to scientific computing in the Soviet Union, in terms of raw Mflops delivered. Only during the latter half of the 1980s, as the El'brus-2 finally entered volume production, did the Soviet HPC sector begin to deliver large amounts of computing power.<sup>1</sup>

Third, during the 1990s the HPC sector has experienced a catastrophic decline in the amount of computing power delivered. In 1991, the Council for Mutual Economic Assistance (CMEA) was disbanded, and trade between its member countries began to be conducted on the basis of international prices [Wsj910107; Iht910629]. Transactions were no longer part of government-to-government agreements, but arranged by individual firms and industries. The volume of trade dropped precipitously. The sale of attached array processors by Bulgaria to the Soviet Union came to a standstill. As East Germany was re-united with West Germany, the sale of computer hardware also declined dramatically.

---

<sup>1</sup>We are deliberately limiting our discussion to high-performance computing. During these years significant numbers of mainframes and minicomputers also were being manufactured. For lack of other options, the scientific community often had to rely on these.

As the Soviet Union entered its final year, huge budget deficits forced dramatic reductions in state orders, which had been used to acquire most of the high-end systems like the El'brus. Individual customers could not afford, or chose not to purchase, indigenous high-performance systems. Orders for these systems declined to next to nothing by the end of 1992.

The decline in production of 1990-1992 is reflected in the flat curve during these years in figure 8-2. In reality, however, the amount of computing power from indigenous systems declined during these years as many El'brus systems were either turned off to conserve energy, or sold and scrapped for their precious metals content. We do not know how many systems have been affected.

### **8.3 High Performance Computing in the Soviet Context**

As we look across the landscape of Soviet high-performance computing, we see that nearly every machine experienced great difficulties in reaching the prototype and series manufacture stages. Carried out within the Soviet economic and political system, the projects had many difficulties in common. Nevertheless, the specific reasons for these, and the relative impact of each varied from project to project.

Before the late 1980s, the nature of the Soviet economic structures and management created an environment which harmed the development of HPC systems more than it helped. High performance computing depended on an extensive infrastructure of supporting industries and enterprises under the management of an equally vast network, generally hierarchical, of ministerial, departmental, and Party administrations. In theory, this centralized arrangement should have streamlined development through greater coordination within the infrastructure, and promoted important projects through priority allocation of resources. In practice, it often gave none of these benefits, even within the military sector.

HPC illustrates the limits of prioritization under such a system. HPC systems required the priority development of an enormous number of technologies. The reality was that resources were not infinite, and HPC had to compete for them with many other advanced technologies, both civilian and military. Stern admonishments from the ministers and members of the Military-Industrial Commission could hasten development of one project for a time, but often at the cost of delays in others. Delays in any link of the development chain delayed completion of the end product.

The Soviet system of economic management provided a nearly impenetrable web which snared projects and reduced the speed and efficiency with which they could be carried out. Arrangements between organizations in the HPC infrastructure had to be made through the a bureaucracy characterized by administrative barriers difficult to circumvent and long chains of approval which needed to be negotiated. In order to take advantage of the centralized features of the economy, one first had to penetrate to the center. Projects with a high-level champion—an Academician, a minister, or an influential member of an important government committee—often penetrated more easily than those that did not.

Negotiating the management labyrinths was a formidable task, but enlisting the participation of the the enterprises needed to provide the material inputs to a high-performance computer was also difficult. We have discussed the disinclination of factories to upset taut production schedules with the introduction of new technologies. When the volume of state orders exceeded a factory's capacity, the factory had an incentive to concentrate on continuing established production rather than suffer a drop in production to upgrade to a new technology.

The monopolistic nature of the Soviet economic system also hindered HPC development. Many critical components of a typical Soviet high-performance computer were manufactured at one, or at best a small number, of locations. When the development of a

given component at a certain plant was delayed or otherwise disrupted, HPC developers had little choice but wait, try to use political means to speed development, or develop the component in-house. Either option led to delays. Acquiring the same component at a different plant was usually not an option. Given captive customers and production plans cast in terms of indicators such as numbers of units or volume produced, plants compromised on quality, causing difficulty for customers and downstream industries that depended on their outputs. The El'brus computers in particular experienced frustrating delays in construction and testing due directly to the unreliability of components.

These features of the environment affected all Soviet HPC projects, but not all equally. Features of the HPC technology itself played a significant role in determining the degree to which the hindrances could be overcome. Table 8-1 outlines some of the elements which have differentiated projects and shaped their R&D cycles. Of particular importance was the degree to which the HPC project stretched the technological capabilities of the supporting infrastructure. The projects with some of the longest development times—the El'brus and Elektronika SSBIS computers—were driving forces for the computer industry, forcing the development of a broad spectrum of new constituent and supporting technologies. In the terms of our conceptual framework, the strategy was to develop systems which drove domestic industry, which required building systems for which the technological availability was low.

It was one of the great dilemmas of the Soviet high-performance computing sector that to raise the technological level of the HPC infrastructure as a whole, the end product had to push the boundaries of technological capability of the entire infrastructure. This increased the complexity of R&D enormously—both technically and administratively—and ultimately slowed development. The paradox was that by pushing so many technological fronts simultaneously, the development cycle of a system was significantly stretched out

<u>Environment</u> Relationship with factories (eestablished, long-term vs. weak, short-term) Existence of influential “champions” in industrial, policy making bodies	<u>Technology</u> Level of complexity Degree to which technology stretches capabilities of supporting industries Ease of manufacturing Complete system vs. add-on computational element
<u>Technological availability</u> Nature of components, subsystems (series production vs. prototype) Availability of components (plentiful vs. difficult to acquire) Availability of necessary tools Availability of know-how (depth of experience in building computers)	<u>Organizational structure</u> Integrated, rigid organizational structures vs. flexible, autonomous units
<u>Organizational slack</u> Large-scale vs. small-scale funding Integrated vs. fragmented funding	<u>Beliefs (design principles)</u> Goals (design system for production vs. to demonstrate a concept)
	<u>Strategy</u> Drive supporting industries vs. use existing technologies

Table 8-1 Elements Differentiating Soviet HPC Projects

because of the time needed to develop and acquire the technologies, and the delays involved in integrating a large number of prototype technologies simultaneously. In order to keep pace with the rate of advance in world-wide state-of-the-art, the next generation systems also required large, rather than incremental, advances in the supporting technologies. This was the paradox: the greater the reach forward, the greater the probability of significant delays.

Within the Soviet system as it existed through the mid-1980s, there probably was no good solution to this dilemma. The national security policies of both the East and the West forced HPC developers to rely on inadequate indigenous industries rather than take advantage of developments world-wide. Had the rate of development of supercomputers in the West not been as rapid, the pressure for advance in the Soviet Union might have

been less, and placed less stress on the infrastructure. While this might have made it possible for the supporting technologies to advance in a more continuous fashion, it is likely that given their incentive structures the supporting industries would have reduced their rate of advance to something less than what actually took place, yet still not improved reliability.

Most of the HPC projects discussed in this work did not push the technological capabilities of the infrastructure. With the partial exception of the multiprocessor computing systems with programmable architecture developed in Taganrog, systems developed with the Academy of Sciences and the Ministry of Higher Education used equipment already in series production. The MARS-M used existing El'brus technology, and most of the others were oriented around the technology used to build the ES mainframes.

Although these projects did not suffer delays directly resulting from the need to develop and test new components and supporting technologies, they did suffer from the lack of close ties with industry. The projects were greatly assisted by the efforts of NITsEVT and ITMVT which provided the technology necessary to build prototypes and lent supporting voices within the Ministry of the Radio Industry and policy-making circles. In spite of this, the academic projects experienced delays (of varying severity) through their relationships with the factories assigned to support their work. Although they used components in series production, they also suffered from low technological availability and waged on-going struggles to acquire the necessary components with the necessary quality. As the examples of the MARS-M and ES-270x demonstrate, from the factories' perspective, the construction of a prototype for an academic research institute was of secondary importance behind their main production, especially when the former required resources and facilities which could be applied to other products.

Although the Soviet economic system has been widely criticized for a lack of responsiveness to customer needs, the plans for each economic unit were based on the expressed needs of other economic units, which served as inputs into the planning process. If there was no expressed need for a given product, particularly those serving the military sector, there was little reason to incorporate it into the planning process. Supporting academic projects through to the prototype stage could be justified on the grounds that this served to advance the research of the field. Putting the systems into series production could be justified only if there were customers willing to take the machines, either because of a perceived need or because of higher-level pressure to do so.

The reality was that there were few customers for academic high-performance computers. Although offering high performance on paper, the systems did not deliver the performance advertised, were unbalanced, were too difficult to program, or otherwise did not suit users with real, demanding applications. As we will discuss below, during the reform period, interest in such unproven machines only declined. Furthermore, some of the systems, while built with standard components, were too complex in their construction to be easily incorporated into existing factory production lines. For these reasons, it was not in a factory's interests to assimilate production of the academic machines.

The technology and the guiding principles shaping their development were to a large extent responsible. A primary goal of the academic machines was to demonstrate the viability of novel architectural concepts. For many reasons, including professional integrity and the bias of funding organizations, the development of unique, sometimes radical, systems was imperative. While the industrial ministries sometimes sought to duplicate the efforts of Western computer manufacturers as in the case of the ES and SM families, such "uninteresting" efforts would have run counter to the mission of the academic research community and would have a difficult time finding high-level support. A machine was

considered successful if it demonstrated a concept, even if it did not meet the needs of users.

The lack of an industrial orientation was not a failing of academic circles per se. Although attitudes towards the appropriate mix of fundamental and applied research carried out in the Academy of Sciences and Ministry of Higher Education fluctuated over time, there has seldom been doubt that the purpose of these institutions is to advance new ideas and expand the base of knowledge. The academic orientation, however, did little to narrow the gaps between the academic researchers, the industrial facilities necessary to build machines, and the users needed to justify the production of such systems. Not only was the bureaucratic distance between an Academy of Sciences research institute and a Minradioprom factory great, the philosophical distance also discouraged active cooperation. The end result was that while they have made some contributions to the body of HPC research, purely academic projects have contributed little to the Soviet computer base.

High-performance computing research advances most quickly through experimentation. A design's strengths and weaknesses are never fully understood until they have been implemented and the effects of memory volumes and access times, cable lengths, signal propagation times, bus synchronization clocks, interconnect latency, and the myriad of other factors related to a physical implementation have their full impact. Advances are made through constructing prototypes, observing their behavior, and building new prototypes taking into account the lessons learned. Theory and conceptual design are necessary, but cannot fully compensate for a lack of experimentation. Similarly, the effectiveness of systems software and algorithms can only be fully evaluated when executed on a physical machine.

Considerable time and effort were expended in all Soviet HPC projects to acquire the resources and components necessary to construct physical machines. Thanks to the resulting delays, the research cycles of individual projects were extended, and the rate at which the lessons learned from building one machine could be applied to its successor was reduced. We can only speculate about the progress of Soviet HPC had the average development cycle been 2-4 years rather than 5-10. It is likely that new ideas would have been generated and tested more quickly, and unpromising lines of development rejected sooner. The state-of-the-art of the sector would probably have progressed much beyond its current state, both conceptually and in performance.

There are few examples of HPC projects which had the technological characteristics, development philosophy, and environmental conditions to be developed quickly and successfully. Thanks to a pragmatic development strategy oriented towards industry, a construction which lent itself to mass production, the use of existing, available technologies, close ties between the research and series production facilities, and considerable high-level support, the PS-2000 was one of the most successful Soviet HPC projects in terms of the length of the development cycle and the levels of series production. The El'brus-B, while not necessarily a high-end system at the time it was built, similarly profited from the use of proven technologies, an industrial orientation, a construction of moderate complexity, and the close ties between ITMVT and the Moscow SAM Plant.

On the other hand, the PS-2000 successors demonstrate more of the difficulties characteristic of the majority of Soviet HPC projects. Although developed with the same industrial-orientation and manufacturability as its predecessor, the PS-2100 experienced delays because of the need to develop a new generation of gate-arrays for the processing elements.

#### 8.4 Technological Paradigms and Trajectories

Is there a “technological paradigm” shaping machine architectures throughout the Soviet HPC sector as a whole? As we pointed out in chapter 2, the definition and scope of usage of this term is unclear in the work of Dosi and others. The term is most often applied to a scientific community; is it also applicable at the level of individual projects? If so, what is the relationship between the two? Dosi and others have not addressed the latter question satisfactorily.

At the community level, a paradigm should be founded on an “‘outlook,’ set of procedures, definition of the ‘relevant’ problems and specific knowledge related to their solution” [Dosi82, 148; Dosi84, 14] which predominate within the community. As we examine the set of Soviet HPC projects discussed in this study, we can identify some general characteristics which they share. All of the projects have a design objective of achieving maximal performance within the constraints of other design objectives and the technological base available (or projected to be available). A key direction of advance from one generation to the next is towards higher performance. To attain high performance rates, the overwhelming majority of systems rely on some form of parallelism.

Two features of the developmental environment are shared by all Soviet HPC projects. High performance has been a consistently important parameter to users and sponsors, policy makers who make decisions about project funding, and the broader scientific and technical community in both the Soviet Union and the West. Although peak performance figures are only marginally useful in describing a system’s utility or applicability to particular classes of problems, they are easy to compute and one of the few metrics which can be applied unambiguously to a broad spectrum of machines. In the case of the El’brus-1 and El’brus-2, a slightly more useful Gibson-3 benchmark performance figure is common. In either case, performance is unquestionably the most commonly mentioned

parameter in HPC. Appropriately or no, it has come to symbolize the level of technological advance within the sector.

A second feature of the environment shared by all Soviet projects has been the weakness of the supporting infrastructure and upstream industries relative to their counterparts in the West. For decades, developers have complained about the low quality of components and materials, the low functionality of the available tools, etc. Soviet HPC developers have not been able to build individual processing elements with a performance comparable to the Western state-of-the-art. They have been forced to take alternative approaches to machine architectures. The predominant characteristic of these approaches is that they incorporate parallelism. In theory, a given level of performance can be reached using a single fast processing element, or a greater number of slower elements. The reality is much more complex, but basic deficiencies in the technology available to designers all but forced them to pursue parallelism. Furthermore, parallelism consistently has been viewed as a means of overcoming reliability problems in the underlying hardware through redundancy.<sup>2</sup>

“Achieving high performance and reliability through parallelism and modularity” sums up the technological paradigm impacting the Soviet HPC sector as a whole. There are few other goals or elements of machine architecture which are shared by Soviet designers. A dominant characteristic of Soviet HPC is the great diversity of approaches toward achieving parallelism, and the very small number of distinct projects which pursue any given one. Table 8-2 reviews the broad spectrum of architectures built within the So-

---

<sup>2</sup>The reality is that the goals of high performance and high reliability are not likely to be achieved simultaneously when the underlying component base is unreliable. Larry Snyder makes this point well: “It is often erroneously thought that connecting multiple copies of unreliable components together can achieve both reliability and performance. Though reliability *may* be achieved, that is, some parts *may* be functional at a give time, high performance is achieved only when all parts are functional all of the time. Thus, improved performance is achieved only when all parts are functional all of the time” [Dong92b, III-3].

Machine	Type of architecture
ES-2700	Attached array processor
ES-2701	Macro-pipeline/coarse-grain compositional language
ES-2703	Programmable architecture multiprocessor
ES-2704	Reduction/dataflow
ES-2705	Analog multiprocessor
El'brus-1,-2	Shared-memory multiprocessor / Stack-based
El'brus-3	Shared-memory multiprocessor / VLIW
MKP	Shared-memory multiprocessor / pipeline
Elektronika-SSBIS	Vector-pipelined
PS-2000	SIMD
Sibir'	Loosely-coupled array processor system
ES-1191	Mainframe host with vector processors

Table 8-2 Spectrum of Architectural Approaches in Soviet HPC

viet HPC community. While policy makers in the Soviet centralized, command economy made decisions about which projects to support, the technical features of Soviet HPC were not generally determined centrally. Although customers had requirements that had to be met by the HPC developers, the specific design was almost always determined, or at least suggested, by the researchers themselves. The systems reflect not only user requirements, but also the design philosophies of individual researchers.

Although the community as a whole has taken widely divergent approaches to achieving the goal of high performance through parallelism, we can find design elements which are shared by subsets of projects. For example, the El'brus and Elektronika SSBIS rely on shared memory; most of the academic projects including the Homogeneous Computing

Systems (OVS), dynamic architecture machines (MDA), and macro-pipeline processor systems use distributed memory. The Multiprocessor Computing Systems with Programmable Architecture (MCS PA) use a hybrid shared-distributed memory system. The PS-2x00 line also is a distributed memory system. The PS- series, OVS, and MCS PA in particular view reconfigurability as a promising means of improving the match between architectures and algorithms.

Other issues create different groupings. The OVS, MCS PA, El'brus, and PS- series use homogeneous processing elements. The MDA and macro-pipeline systems have homogeneous computational elements, but incorporate other types of processing elements for control functions. During Burtsev's tenure at ITMVT, the philosophy of incorporating a variety of special-purpose processors into an El'brus-2 configuration was pursued actively. The MARS-M and Sibir' projects are two extremely different systems which incorporate heterogeneous processing elements.

Soviet HPC can also be categorized by whether they operate in a single-instruction, multiple-data (SIMD) or multiple-instruction, multiple-data (MIMD) mode, the nature of the interconnect system, whether or not the system is designed to be attached to a general-purpose mainframe host, the use of horizontal architectures, the reliance on high-level language constructs, etc. Machines which are grouped together by one criterion are frequently not grouped together by another. In short, with regard to machine architecture, it is difficult to find design principles, or groups of principles, which are adhered to by the HPC sector as a whole. In the case of Soviet HPC, the number of principles shared by significant subsets is not large. In short, beyond achieving high performance and reliability through the use of parallelism and modularity, there does not appear to be a clearly identifiable paradigm for the Soviet HPC sector.

Although Soviet HPC projects differ greatly, they exhibit a great deal of internal design consistency from one generation to the next. In our study we have seen a striking continuity of many architectural approaches in the machines for which multiple generations have been built: the El'brus, the PS-2x00, the dynamic architecture systems, the multiprocessor computing systems with programmable architecture, etc. Each of these families have followed a technological trajectory characterized by continuity, rather than discontinuity. In fact, it is perhaps surprising how stable the technological trajectory has been during the reform period. The architecture of the El'brus-3,<sup>3</sup> the PS-2100, the successors to the ES-2703 and ES-2704 share the dominant characteristics of their predecessors and have few features which reflect the growing turmoil of the surrounding social, economic, and political systems.

We have discussed factors contributing to the stability of the technological trajectory. Each of these families was built under the influence of a stable set of users and user requirements and/or a set of design principles held very strongly by the main engineers. In the case of machines developed for specific customers, the selection environment—strongly shaped by the requirements of those users—remained stable, at least until the breakup of the Soviet Union and in some cases (the El'brus-3 in particular) later.

The technology itself played an important stabilizing role in several different ways. First, requirements for compatibility forced a new machine to share many basic features with its predecessor. Second, the longer a machine had been under development, the more costly in time and money it would be to radically alter the design. The El'brus-3, for example, was initiated around 1984. By 1990 the machine had been rather completely designed and arrangements had been made with the supporting industries to provide the necessary components and subsystems. Altering the design at this point would have been

---

<sup>3</sup>We discuss the shift from stack-based to VLIW in the El'brus line below.

costly. Third, if an architectural approach was basically satisfactory yet allowed room for improvement, designers had little incentive to make the effort to master a dramatically new architectural approach.

From our study we can see that in the case of Soviet HPC, it is much easier to identify a meaningful technological trajectory and associated paradigm and selection environment for individual projects than for a community as a whole. What is viewed as an element of consistency within one line of development may not be shared by any other projects, and thus not an element of a more broadly held paradigm. In short, there are important elements of intra-project consistency which are not explainable by identifying a paradigm in effect for an entire community.

The case of the El'brus series gives us further insight into the nature of technological trajectories and paradigms. In chapter 4 we saw how the El'brus-1, -2, and -3 lie along a technological trajectory which has been quite consistent for over 20 years. In each generation, designers sought to increase performance through some combination of faster and improved components, reduced clock periods, greater volumes of primary and secondary storage, greater numbers of processors and functional units within processors, and improved processor architecture. Basic systems characteristics—coarse-grain parallelism through a moderate number of powerful processors with shared main memory, modularity, multiprocessing, independent I/O and data transmission processors, hardware support for high-level language constructions, software compatibility with previous generations—remained very similar throughout the generations.

One of the few points of sharp discontinuity in the technological trajectory was the design of the individual processors, as the stack-based architecture of the El'brus-1, -2 processors was replaced by a VLIW approach in the El'brus-3.

Did a paradigm shift take place? How much change is needed in a paradigm before one can say that a shift has taken place? These are perhaps the wrong questions. They assume that a paradigm, even at the level of a single product line, is a single, indivisible entity. The concept of a 'paradigm' must be modified to account for the fact that within a given project, certain elements can remain quite constant as the technology develops, while others can change dramatically.

A more useful way to view a paradigm is as a series of layers of finer-grained "'models' and 'patterns' of solution..." which cover the spectrum of technological problems to be addressed during the development of a complete system. We can refer to such layers as micro-paradigms. The micro-paradigms guide developers' decisions about specific subsystems or parts of the complete system. The history of the El'brus series suggests that some micro-paradigms can shift dramatically without necessarily causing a shift in others. The shift to a VLIW approach with static scheduling did not cause major changes in the principles of modularity, coarse-grain parallelism, shared-memory, etc.

Because the micro-paradigms affect the development of portions of a complete system, they are in practice not independent. Unless subsystems function well together (or, in instances where inter-generational compatibility is a requirement), the computer will be unbalanced and not deliver the performance and functionality desired. There are hundreds of examples. A memory system which is slow, or has certain bottlenecks will not be able to provide a fast processing element with data rapidly enough to avoid excessive idle time. A lack of software compatibility between generations will require extensive re-coding by software developers and users. A cooling system must be capable of dissipating the heat generated by the components. Similarly, decisions to alter one subsystem dramatically may force dramatic changes in other subsystems. For example, had El'brus designers determined that a fine-grained, massively parallel, distributed memory system

held the best promise for the future, most of the earlier ideas about processor architecture, memory structure, interconnect systems, etc. would have had to change significantly.

An important issue for technological innovation and advance is the conditions under which paradigms and micro-paradigms can change. Our study of Soviet high-performance computing indicates that several factors are necessary. First, there must be a mis-match, or growing incompatibility between the technological approach being taken and the selection environment which indicates which kinds of technology are acceptable. This can occur through a qualitative change in the requirements for a system, such as those specified by principal customers, or because a given technological approach is not able to meet the goals of quantitatively changing requirements. In the case of Soviet high-performance computing through the period covered by this study, there was little qualitative change in requirements of the principal sponsors, leading to a high degree of consistency in the technological trajectories of most HPC projects. The nature of the target applications changed little. The principal changes were quantitative--increased performance, improved functionality and reliability, and so forth.

Some systems, like the PS-2x00 series, could meet the requirements adequately through quantitative means. The essential architecture of the individual base modules differed little from that of the PS-2000. Advances were mostly extensional: uniting multiple base modules and providing for their interaction, increasing the amount of main and peripheral storage, increasing the word length, using an improved component base, etc.

Other systems, like the El'brus-3, could not meet the requirements solely through incremental extensions of existing approaches. The nature of the basic requirements had not changed, but existing technological approaches to the design of the processors could not meet the high performance demands.

Two other factors strongly affecting whether or not a shift in a paradigm or micro-paradigm occurs are the scope of impact of changes to the technology and the scope of decision-making needed to effect the change. The scope of impact and scope of decision-making are intimately related to the nature of the developmental environment and organizational structure, and the technology and associated paradigms. The scope of impact refers to the degree of coupling between elements of a technology. The broader the scope of impact, the greater the changes that need to be made in the systems which interact with the specific technology. The greater the scope of decision-making, i.e. the numbers and types of individuals and organizations involved in the decision-making process, the more difficult it is likely to be to make a decision for change.

The lower the scope of impact and scope of decision-making, the easier it will be to implement a change. In the case of the El'brus-3, the change from a stack-based to VLIW architecture had a relatively limited scope of impact and scope of decision-making. The scope of impact was limited by a series of well defined interfaces which insulated the processor from the surrounding systems, both technological and social. For example, the high-level programming languages defined the interface between applications and systems software and the underlying hardware. The compilers were the only pieces of software which directly reflected the underlying processor architecture. Hence, changes in the processor architecture required changes to the compilers, but not to existing systems or applications software. Implementing this change also required the decisions and cooperation of a relatively limited group of individuals, primarily those within the El'brus development team and selected individuals in the ITMVT and ministry hierarchy.

In contrast, similarly drastic changes to the PS-2x00 processor architectures would almost certainly have required changes to the assembler level language in which much ap-

plications software was written. Large bodies of existing software (and hence users) would have been impacted. The social cost of such a change would have been high.

In principle, Academic systems without a user community have greater freedom to make drastic changes to the architecture. They may, however, lack to financial and human resources to make changes to a many aspects of a design simultaneously.

The strategies of minimizing the scope of impact and decision-making have proved to be extremely powerful facilitators of technological advance in Western development as well. Network designers have employed layered approaches to protocol stacks as a means of reducing complexity and allowing individual vendors considerable freedom to alter implementation details of any given protocol layer. A similar strategy lies at the core of recent trends towards open systems. Open systems can be viewed as a collection of “black boxes” with well defined interfaces between them. As long as a “black box” adheres to the necessary interface standards, the internal implementation can be varied easily.

A fourth factor influencing paradigm shifts is the availability of ideas about alternative directions of advance and examples of successful implementation. In particular, we have discussed how the ILLIAC-IV, Burroughs 700 Series, FPS attached array processors with horizontal architectures, and Cray supercomputers served as sources of inspiration for the PS-2000, El’brus-1 and -2, El’brus-3, MKP, MARS-M, and Elektronika SS-BIS respectively. In each case, the Western models shaped Soviet developments in two ways. First, they demonstrated particular architectural approaches. Second, and perhaps more importantly, they showed that these architectural approaches were basically viable. The latter factor gave development teams much of the confidence they needed to pursue implementations, and, in some cases, proved valuable in obtaining higher-level approval. It is certainly not the case that all ideas found in Soviet HPC computers were inspired by

Western developments. Many ideas and their implementations are indigenous. Nevertheless, our study indicates that the availability of Western ideas has been a powerful catalyst for advance in Soviet high-performance computing.

A fifth factor is the feasibility of implementation. Feasibility is closely related to the scope of impact discussed above, but is also a function of cost, time to development, and other factors. Projects like the PS-2000 and academy projects with modest resources which stayed within the bounds of what the domestic industry could manufacture experienced greater feasibility constraints than did those, like the El'brus, which were designed to push technological boundaries.

Finally, a more human element is the tenacity with which principal designers hold on to their beliefs about the direction of advance. This is particularly evident in the work of NIIMVS in which there is only slight variation from one project to another in basic design philosophies of processor architecture, interconnect systems, etc. V. A. Kalyayev has exerted strong control over the research agenda and permitted little experimentation outside the framework of the established paradigm.

The factors we have just discussed play an important role in facilitating or hindering a paradigm shift. Our list is not exhaustive, particularly if we wish to generalize the discussion to other types of technologies. Studies of computing technologies more broadly, such as those by Kling [Klin82; Kling84] make it clear that for technologies which are socially complex, systems are more likely to evolve as a by-product of technological and social factors rather than through strictly rational decision-making. Although not socially simple, Soviet high-performance computers are not as socially complex as other systems, such as the management information systems discussed by McHenry [Mche85]. Characteristics of the technology and the immediate infrastructure necessary for their develop-

ment have considerable explanatory power in helping us understand the nature of their technological advance.

While it is beyond the scope of this dissertation to predict conclusively under what combinations of these factors a paradigm shift will occur, we are able to point to a number of changes in these factors which could very well lead to significant paradigm shifts in the future. We discuss these in our final chapter.

### **8.5 The Impact of the Reform Process on Organizational Structure**

We have discussed how changes in legislation regarding state enterprises and associations, small enterprises, joint ventures, and cooperatives have dramatically altered the opportunities and mechanisms for organizational change. An important outcome of the reform process has been a decentralization of authority and responsibility for an organization's structure and domain(s) of activity. How have the organizational structures changed, and why? What has been their impact on the development of high-performance computers?

Throughout the former Soviet Union, there has been a pronounced trend towards the fragmentation of organizational structures at all levels of society. The role of the ministries and organizations such as the Academy of Sciences in the lives of their subordinate institutes has decreased dramatically. The State Committee on Science and Technology (later absorbed into the Russian Ministry of Science, Higher Education, and Technology Policy) has retained some influence through its funding practices, but lost much administrative ability to control S&T throughout the economy. The government is still the owner of land, buildings, and much capital equipment of most organizations, but has minimal direct influence in day-to-day decisions about the activities and structure of enterprises and institutes.

Within the institutes we have discussed, we have witnessed a transformation away from a unified hierarchy toward a loosely-coupled collection of smaller-scale organizations with greater autonomy. This trend has taken place more completely in some organizations than in others. ITMVT and NIIMVS maintain a hybrid structure of an integrated hierarchical core with a number of associated but autonomous organizations existing within the shell of the institute as a whole. NIIUVM consists almost entirely of financially independent “rental collectives.” ISI is currently a collection of laboratories, small enterprises, etc. pursuing research or contract opportunities independently of each other with little coordination from central administrators. Similar patterns are, we believe, taking place at the other R&D facilities mentioned in this study.

The transformation of organizational structure reflects the balance between forces for decentralization (driven by efforts increase the organizational slack and achieve greater autonomy) and centralization (driven by the desire to preserve the ability to conduct HPC R&D). In the cases we have studied, the former has dominated the latter.

The changes in legislature allowing alternative organizational forms and local decision-making have been a powerful enabling factor. Information about successful incorporation of new forms at other institutes often encouraged leaders at the institutes we have studied to make similar changes. The principal factor driving the changes, however, was a desire to increase “organizational slack,” or the level of resources at the disposal of the organization. The latter could be accomplished by working more efficiently, by generating additional revenues, and by converting existing resources into a more flexible, useful form, i.e. the conversion of accounting rubles (*beznalichnyye*) into cash (*nalichnyye*). A basic objective has been to find a way to retain the engineers who constitute the core of the institutes’ technical capability and keep them from seeking employment else-

where. Money for wages had to be generated, and restrictions on wage levels had to be skirted.

The creation of cooperatives, small enterprises, rental collectives, and temporary collectives served each of these purposes. They provided a way to get around legal restrictions on wage levels (or, more precisely, the amount of money available to pay wages), to enter into contracts with negotiated (i.e. higher) prices, to convert accounting rubles into cash, and bring together individuals best suited to carry out a particular task in an efficient, timely manner.

Our study indicates that the nature of the revenue stream and the opportunities for alternative organizational forms have a significant influence on organizational structure. In each of the organizations in our core cases a reduction and fragmentation of income has led to a fragmentation of organizational structure. We will discuss the changing nature of income streams and its implications for technology in the next section.

## **8.6 The Impact of Reform on the Development of HPC**

The reform process initiated in 1985 by Mikhail Gorbachev has followed a complex, uncertain, largely uncontrolled path which has fundamentally changed most facets of Soviet economic, political, and social life. The goal of “democratic centralism,” of greater autonomy and use of economic mechanisms at the local level coupled with more comprehensive coordination by central government organizations has largely unravelled; the forces of decentralization and fragmentation have overwhelmed the forces seeking to improve centralization in both the political and economic spheres.

The reforms have brought about some changes which are likely to have a positive impact on innovation in Soviet HPC over the longer term. The administrative-directive form of economic management has to a significant degree been replaced by one based on eco-

conomic considerations. Enterprises have achieved the autonomy over transactions with other enterprises and organizations which Berliner called for [Berl76, 522]. The quality of feedback between suppliers and customers has increased as suppliers, facing declining markets and excess capacity, have been forced to court customers. The customers, thanks to their own weakening financial condition and control over finances, have become more demanding. The sensitivity to customers' needs has increased at factories and at research institutes whose "products" are pieces of research or technology development carried out under contract. In addition, in the case of ITMVT and NIIUVM, the research institutes have assumed much of the burden of marketing the high-performance systems. The creation of the Supercomputer Association, its composition, and the grass-roots nature of its operation are further indications of the strengthening of the customer feedback loop. An important feature of current inter-organizational transactions is that they are based on satisfying the requirements of customers rather than bureaucratic watchdogs who monitor planning indicators and procedures.

The reforms have brought a new flexibility to the organization of R&D. We have witnessed the creation of development teams and organizations which draw members from a variety of existing organizations to address specific tasks.

The reforms have created opportunities for expanded contacts with the international community, opening the doors for better professional interaction, foreign investment, and/or contracts for work or products. Joint efforts such as that between Sun Microsystems and ITMVT (Moscow SPARC Center) would have been unthinkable a decade ago. Although many aspects of technology transfer between Soviet and Western counterparts remain subject to export control restrictions, arrangements like these offer Soviet scientists the hope of accessing Western capital, technology, and know-how and the opportunity to observe Western practice more closely.

These changes have the potential for improving some critical parts of the innovation process. Idea generation can be improved through increased interaction between suppliers and demanding customers and idea “cross-pollination” across organizational and international boundaries. Gathering support for an idea can profit from the potential access to foreign funding sources and more open communication between individuals in different organizations (although the latter is partly offset by growing possessiveness over ideas of possible commercial value). The idea implementation process can be improved through the use of the more flexible organizational forms, the growing willingness of factories to manufacture products which they can sell, and the reduction of bureaucratic overhead in the development process.

These positive elements currently exist more as opportunities than as realities, however. While they will undoubtedly have a significant positive impact on innovation in the future, they are overshadowed by the negative consequences of the reform which have seriously undermined Soviet ability to carry out large-scaled R&D in advanced technologies. First, the landscape is dominated by the desperate state of the economy and a non-existent market for Soviet HPC. Second, fundamental weaknesses remain in the Soviet infrastructure. Unless the economy improves dramatically and the structural weaknesses are corrected, there will be only limited, localized benefit from the improvements listed above; there will be no Soviet high-performance computing sector of any consequence.

#### 8.6.1 Economic Considerations

The real demand for expensive, powerful systems at present is very low. The financial state of current and potential HPC users is so poor that few are able to acquire them. The ballooning federal budget deficits coupled with a policy of reducing the portion of production carried out directly for the state has caused the volume of state orders to de-

crease dramatically. Series production of Soviet HPC systems ground to a halt by the end of 1992.

The depressed economy has prevented a market for Western high-performance systems from growing as well. Western vendors with permission to sell to the former Soviet Union have found the market virtually non-existent. A Convex executive responsible for sales in Russia has stated that in spite of a number of inquiries from Russian organizations, “we haven’t seen any money yet, so they haven’t seen any computers” [Huds92].<sup>4</sup> A healthy HPC sector cannot exist without a market.

In cases where users are able to afford large-scale systems, available Soviet machines face increasingly fierce competition from Western models. Current CoCom regulations still shelter the market for high-end Soviet HPC systems such as large configuration El’brus-2s, the Elektronika SSBIS and the MKP, at least in terms of performance levels. Mid-range workstations and mainframes can be imported with few restrictions, but larger systems are either categorically prohibited, or are sold with cumbersome restrictions. In practice, users who have hundreds of thousands or millions of dollars available to spend on computing equipment are attracted by Western mid-range systems and workstations which are highly functional, reliable, and can sit beside an engineers desk. Given the rapid advances in technology,<sup>5</sup> the number (hundreds of thousands) of units manufactured, and the pressure to acknowledge global trends in computer technology evolution and relax export control restrictions, Western workstations will provide a viable alternative to indigenous systems for most Soviet HPC users. This will be a truism if, as appears to be case, production of the older generation of Soviet HPC systems (the El’brus-2 and

---

<sup>4</sup>Convex recently installed its first (legal) unit in Russia at the Joint Institute of Nuclear Research in Dubna, Russia [Hpcw930628].

<sup>5</sup>Witness the recent introduction of machines based on the Alpha microprocessor which compete favorably with Western mainframes or even supercomputers. In a recent test, an Alpha AXP system performed a sort benchmark six times faster than the Cray YMP which set the record in 1992 [Hpcw930405].

PS-2100) has ended before successor systems reach volume production. Series production of high-end systems has all but ended, yet demand for the new generation of systems is too weak to support series production of systems like the MKP and Elektronika SSBIS which have reached the prototype stage. By the time the economy recovers to the stage where potential HPC users have the funds necessary to acquire large-scale systems, the current models will be obsolete, especially in comparison with Western models which will continue their rapid rate of technological advance for the foreseeable future.

#### 8.6.2 Structural Considerations

Even if the economy improves, Soviet HPC developers face a number of structural weaknesses which fundamentally compromise their ability to develop new machines over the long term. We have mentioned how the reforms have brought increased flexibility to organizational structure and inter-organizational ties. They have also brought fragmentation, or the breakdown of ties which we discuss in this section.

A major structural challenge facing R&D facilities is the nature of the indigenous infrastructure supporting HPC development. Large-scale development still requires the participation of hundreds of upstream organizations. As a consequence of the reforms, these links are now, for the most part, not administrative (vertical), but economic (horizontal). Their existence depends on whether or not the upstream organizations feel that it is in their best interests to participate in the development or production of a particular good, and their ability to deliver. Thanks to declining production levels throughout the economy and the creation of a fair amount of idle production capacity, organizations are more willing now than previously to consider new orders. At the same time, they are not likely to take on orders which will not be profitable for them, as might well be the case for highly specialized pieces of advanced technology for which the market is limited. Even if they

wish to engage in such work, under current conditions acquiring the necessary inputs of the necessary quality is quite difficult.

More fundamentally, the Soviet economy still has a highly monopolistic nature for many products. While the number of factories making consumer electronics has mushroomed, the number of facilities able to carry out advanced microelectronics R&D, for example, has remained constant or declined. The indigenous infrastructure for Soviet HPC still lacks redundancy. As a result, the problems of delay and quality which result from non-competitive, monopolistic industrial organization will continue to plague Soviet HPC developers and overshadow some of the benefits of more flexible inter-organizational ties. We will discuss possible alternatives to this situation in the next chapter.

A second structural weakness is the administrative and financial gap between research and production facilities. Soviet policy-makers have tried for many years to bridge the gap between research and production, chiefly through scientific production associations (NPO) and research clusters such as in *Akademgorodok* in Novosibirsk which encourage R&D to be carried through the combined efforts of Academy and industry organizations. At no time have the component entities of the HPC sector been linked so that a portion of the profit realized through series production was poured directly back into R&D for the development of the next generation of products.

In the cases of NPO Impul's and ITMVT, funding for development of high-performance systems came through state funding (from the state budget or principal sponsors) ear-marked for the development of a specific system. These funds were used to support R&D and production in upstream industries as well. When a prototype was completed, factory documentation was turned over to the series production facility which built units in response to a production Plan, or individual orders from customers. Pro-

ceeds of series production were not channeled back into the R&D facility. The development of new generations of systems had to be supported by specific funding from government or wealthy customers.

The forces of decentralization unleashed by the reform efforts penetrated below the boundaries of the NPO. Factories became administratively and financially independent of the research facilities. The former, themselves struggling for solvency, had little inclination and few resources to support the R&D of advanced technology in a different organization. To the extent that research funds were available, they were redirected towards goods which would cost little to develop and which would enjoy an immediate market. Such goods, like telephones and washing machines, were developed inside the factory.

V. V. Rezanov, deputy-director of NPO Impul's calls this situation "a hole in *perestroika*." One of the goals of *perestroika* was to incorporate decentralized mechanisms in the management of the economy. However, decentralization penetrated below the boundaries of the NPO to the level of individual institutes and factories and even to the level of divisions and laboratories within institutes. At a time when prospects for continued government support for R&D projects were becoming very uncertain, the R&D facilities had less opportunity to profit from the manufacture of systems which they had developed.

It is absolutely essential for R&D facilities in the Soviet HPC sector to find an arrangement in which the research and production facilities are tightly integrated, both financially and administratively—in effect combined into one organization—so the organization can generate revenue through the sales of series production goods, which can be used to fund in-house development of new products. The alternative is for R&D always to be funded directly from the government budget, or by customers willing to fund a large portion of the entire R&D bill themselves. In the first case, the link between customer prefer-

ences and R&D efforts will be weak and/or not lead to commercially viable products, and administrative barriers between research and production will remain problematic. Under the new economic conditions, it is unlikely that customers or investors will be willing to fund massive R&D efforts unless they can be assured of recovering their investment. The number of such customers is already extremely limited, and the prospects for long-term, adequate, stable funding through them are not good.

A third structural problem is that the fragmentation of organizational structures will impact Soviet ability to conduct R&D on large-scale projects. We have discussed the fact that development of large-scale projects depends on an appropriately extensive and integrated organizational structure to provide the direction and coordination necessary to build a functional system. Such a structure need not be rigid, but mechanisms must be in place to enable a variety of organizational units which complement each other to work together towards a common goal. While a moderate reduction in the income stream helped move organizations towards more flexible structures, fragmentation resulted from a drastic decrease in and fracturing of the income stream. As funding levels for HPC R&D decreased in relative terms, institutes were left with few alternatives but to find other sources of funding, mostly through contract work or, in rare instances, through joint efforts with Western companies. In the desperate economic climate, such contracts were, as a rule, small scale and short term. Whether smaller contracts caused smaller organization units to form, or smaller organizational units sought out contracts which matched their capabilities can be debated. The influence probably ran in both directions. Either way, a rough equivalence developed. Given the difficulty in securing large-scale contracts, a strategy adopted by the leadership in nearly all the institutes we have examined was to distribute the burden of finding alternative funding sources, placing much of the responsibility for survival on the shoulders of the small autonomous units themselves.

Contract work tends to lead to a fragmented income stream. First, payment is very closely tied to the execution of specific tasks. Unless the profit margins are quite high, such work generates a minimal amount of slack resources which can be applied to other development projects. Although in principle the cost of a contract is negotiated between customer and provider, the poor economy and past practice under the Soviet system tend to keep profit margins low. Customers with little money to begin with and who are used to contract values' equaling the cost of doing the work plus a small margin are reluctant to grant providers large profits. Second, contract work tends to be customized for the consumer and does not result in a product which can be broadly marketed. Furthermore, as a function of the poor financial state of most organizations in the depressed economy, year-to-year or even month-to-month budgeting, and the multiplicity of customers, individual contracts currently are often relatively small, supporting the work of a team of engineers for a relatively short period of time. Such projects do not generate sufficient income to support the development of large-scale high-performance projects. In other words, contract-based work has limited potential for generating the volume and type of income needed to support large-scale R&D in the future.

A necessary, although not sufficient, condition for conducting large-scale R&D is a unified flow of income sufficient to meet the costs. We have mentioned possible sources: government funding, support from individual, principal customers, production and sales of goods which generate revenue.

It is an unfortunate paradox that the measures needed to preserve capability and keep development teams employed—the use of more autonomous organizational structures—threatened to fragment the structure necessary for large-scale development. It can be argued that if funding is restored the resources needed to support such a structure will again be available. This is probably so, but the ease with which Humpty-Dumpty can be put

back together is also a function of the length of time the smaller organizational units evolve autonomously and of the diversity of their directions. Over time, they will probably drift apart in terms of shared technology, shared research goals, and possibly shared culture. The ability to conduct large-scale HPC R&D will be compromised.

The preservation of some semblance of an integrated structure has depended on the persistence of the directorate, and the level of funding available to support a basic level of research on advanced projects. ITMVT, for example, continues to receive funding to finish existing projects and support the core R&D teams. Funding for the PS-2300 had, at the time of this writing, dried up completely, and the development team was returning to the institute's traditional emphasis on control systems. Itenberg's rental collective retained some ties to other organizations within NIIUVM thanks to a small, but existing market for the institute's control systems. The ISI divisions and laboratories are carrying out independent research and HPC development has ended.