

# Table of Contents

## September - October 1994 -- Volume 2, Number 5

- [NAS to Join Ames Information Systems Directorate](#)
- [Improving I/O Performance on the NAS CRAY C90](#)
- [From NX to MPI: A Primer](#)
- [RAID Technology Offers Cost-effective Data Storage for CRAY C90s](#)
- [Virtual Reality, Interactive Media Highlighted at SIGGRAPH](#)
- [Linearizing Nested Loops](#)
- [Users, Developers Converge at PTOOLS Consortium](#)
- [Pat Elson, User Liaison `Extraordinaire'](#)
- [Credits for this Issue](#)
- [This issue's front page](#)

[\[Next Article\]](#)

[\[Table of Contents\]](#)

[\[NAS News Home Page\]](#)

[\[NAS Home Page\]](#)

# NAS to Join Ames Information Systems Directorate

*by Elisabeth Wechsler*

A reorganization in progress at Ames Research Center is expected to have little impact on NAS users, except to make additional technologies more accessible, according to Dave Cooper, Chief of the NAS Systems Division since 1991 and a 32 year NASA veteran.

Effective October 1, Cooper will head Information Systems, one of four consolidated Ames directorates. The other directorates are Aeronautics, Space, and Center Operations. Previously, Ames was divided into eight major entities.

## Focus on `Best' Technologies

The purpose of the reorganization has been to better focus the Center's activities on its customers, on the technologies it does best, and to make the organization more efficient and cost effective, Cooper explained.

NAS will become part of Information Systems, yet remain a distinct entity. "NAS is a very important organization--not only to Ames--but also to the NASA Office of Aeronautics," Cooper said. "Because of its importance and visibility, we've seen fit to keep the NAS organization as a separate division."

Information Systems will include all computers and communications at Ames, most of the Office of Aeronautics' supercomputing activities, artificial intelligence and expert systems, plus other elements of computational sciences such as neuroengineering, photonics, and visualization technology. This organization will manage the Ames part of the High Performance Computing and Communications Program, including the Information Infrastructure Technology and Applications and National Research and Education Network, as well as the Health Open System Trials and Surface Movement Advisor projects.

## Information Systems To Serve As Model

"We have a lot of activities and firm commitments with solid milestones on these programs," Cooper said. "What I envision for Information Systems is that we press the leading edge of these technologies and that we take and infuse those technologies into the rest of the aeronautics program at Ames, and into

Space and Center Operations. Once we've done that, I hope that we'll be able to serve as a model center for NASA in how information systems or technologies can change the way we do business. I believe these changes can make us considerably more cost effective and efficient in what we do," he said.

Cooper noted that while a few functions within the NAS organization may be downsized as a result of mandated budget cuts "we'll try to minimize the impact of the reorganization on NAS users. The result will be transparent to our customers, except that there will continue to be more emphasis on technology transfers," he said.

## **Chancellor Named Acting Chief**

Marisa Chancellor, currently Deputy Chief, was named by Cooper to be Acting Chief of the NAS Division, also effective October 1, while the position of Chief is advertised nationally. It could take up to a year before this and other senior executive positions at Ames are permanently staffed, he said.

Though he will be busy organizing and managing the new 700-person Information Systems entity, Cooper does not want to lose touch with NAS external customers. "I have created excellent working relationships with our customers and, despite my new responsibilities, I want to continue to work with these customers to help them meet their requirements."

## **New Technologies Offered To NAS Customers**

Cooper thinks that Information Systems is well positioned to offer aerospace customers better opportunities to use expert systems and artificial intelligence. These technologies, though only recently gaining mainstream commercial acceptance, can be used in a variety of ways by the aeronautics community and are "ripe for development," Cooper believes.

Cooper said he will continue to "sell and defend the NAS Program in Washington" and, working closely with Chancellor, will do everything possible to maintain the NAS budget at current levels.

## **Reorganization Team Reached Broadly Within Ames**

The reorganization team at Ames was appointed by former Center Director Dale Compton, who retired in January. When Ken Munechika came on board as Center Director, he asked the team to continue its work and advanced its schedule for submitting a reorganization plan, Cooper said.

The team, composed of 11 individuals representing all current Ames organizations, worked virtually full time for five months on the reorganization, Cooper added. Many other individuals at Ames participated by providing input to the reorganization team or by serving on one of three teams formed to identify Ames' core competencies.



*Dave Cooper*



*Marisa Chancellor*

[\[Next Article\]](#)

[\[Table of Contents\]](#)

[\[NAS News Home Page\]](#)

[\[NAS Home Page\]](#)

[Next Article](#)[Contents](#)[Main Menu](#)[NAS Home](#)

# Improving I/O Performance on the NAS CRAY C90

by Clayton J. Guest

Experienced NAS High Speed Processor (HSP) users are familiar with techniques for improving Fortran I/O speeds on the CRAY C90 system (**vonneumann**). This article is geared toward new or inexperienced NAS users and may serve as a review for experienced users.

All too often, Fortran programs seem to run forever and consume large amounts of CPU time. Users sometimes wonder why computers such as the CRAY C90, which is rated at one gigaflop per CPU, shows poor performance; frequently, I/O is the answer. Poor I/O performance often manifests itself when system CPU time approaches or exceeds user CPU time.

Users can obtain the values for system and user CPU times by including job accounting (**ja**) commands in a job's NQS script. The job accounting output contains system CPU and user CPU time in seconds, and includes the time the job started, elapsed time, and other information that may be of interest to users. Examining the **ja** output to see if the system CPU approaches or exceeds the user CPU time can point out potential I/O concerns.

If a legitimate I/O concern exists, there is no cause for alarm or for rewriting the Fortran program to improve I/O speed. Several approaches can improve performance by reducing the ratio between system and user CPU times. The following methods have been successful in improving the performance of Fortran I/O.

On **vonneumann**, do not attempt to execute Fortran programs with the I/O being sent to or retrieved from the /m filesystem directory. This directory is meant for bulk data and should not be used during program execution. Using /m for execution can produce greater system CPU time than user CPU time--so, the overall time required to execute a job becomes much greater than it should be. This does not imply that /m should not be used--simply copy data from the /m directory to a faster directory prior to executing the program and, conversely, copy the data back to /m after the program executes.

## SRFS Helps Improve Performance

The Session Reservable File System (SRFS), available to users on both **vonneumann** and the Aeronautics Consolidated Supercomputing Facility system (**eagle**) also helps improve I/O speed. Through SRFS, users can guarantee that file space will be available prior to executing a program. When file space is guaranteed, the system performs I/O better and the user CPU to system CPU ratio is better than if executing in your home directory.

If a program is I/O intense, the /fast filesystem directory is very helpful in improving I/O speeds. /fast uses the CRAY C90's solid-state storage device (SSD) hardware. The SSD has a bandwidth of 100 gigabits (gb) per second, which gives it high transfer rates and allows the system to reduce system CPU time.

To illustrate the differences among the various directories in system CPU time, a simple Fortran program was prepared, and then executed during normal work hours and conditions. The program does very little computation -- only enough to fill an array of 2,048 words. The array is written 90,000 times, then the file is repositioned to the beginning and the array is read into memory 90,000 times. All I/O is done with standard Fortran 77 statements, such as WRITE, REWIND, and READ. All I/O is binary and "well formed"; that is, I/O records are in multiples of 512 words. The associated UNICOS **assign** commands used defaults and had only directory information. The differences among directories are shown in Figure 1. Note the constant time of user CPU and wide differences of system CPU. The information was obtained before /big included the RAID disk system (*see [RAID Technology Offers Cost-effective Data Storage](#)*). Additionally, if the program had performed more computation, the system CPU might have been different.

**Figure 1:**  
**Differences in system CPU time when using different directories for I/O.**

| Directory | User CPU | Sys CPU |
|-----------|----------|---------|
| /m        | 6.37     | 21.20   |
| home      | 6.37     | 8.54    |
| BIG       | 6.36     | 6.35    |
| FAST      | 6.36     | 3.83    |

The SRFS directories /tmp, /big, and /fast have limits. These limits are shown in the man page for SRFS on **vonneumann**. Users should request only the amount of space needed when using these directories, and must copy information to and from these directories.

## Using the UNICOS Assign Command

The UNICOS **assign** command can also be instrumental in achieving better I/O performance on the CRAY C90. The man pages contain details on this command. Two of the parameters for **assign** are discussed here: the structure type (**-s**) and buffer size (**-b**) options.

The structure type declares how the data is buffered and how it is to be stored in the CRAY C90's filesystem. The optional types are: **cos**, **blocked**, **bin**, **sbin**, **unblocked**, and **u**. Each of these has a unique

method of manipulating I/O records. For example, **cos** and **blocked** are identical--only the name is different. Control words are appended to these records when written. The control words contain blocking information, such as record length. Control words are a carry over from COS, the original Cray Research Inc. (CRI) operating system. With **cos** or **blocked** structures, library buffering is performed at the system level or the data is moved to and from the user's program via data buffers maintained by UNICOS. The **cos** structure type is the default for unformatted and BUFFERIN/BUFFEROUT. Direct access and GETPOS and SETPOS are not supported for **cos** structures.

The **bin** structure should be avoided except when word addressable I/O is required.

The **sbin** structure is used with unformatted access. This option is compatible with standard C file I/O as defined in `stdio.h`. This file structure is obsolescent, and CRI will discontinue support sometime in the future.

In the **unblocked** structure, adjacent records are not delimited from one another and no control words are added. The **unblocked** structure is the default for direct-access unformatted files.

With the **u** structure, each read or write request results in an immediate system call and no system buffering is performed. Requests should be made in multiples of the sector size for the target filesystem. I/O requests should be well formed. Requests that are not well formed will go through the main Unix system buffer, where well-formed requests will use raw I/O and cause the data to be transferred to the device directly out of the user's data buffer, bypassing the Unix buffer. If data is being moved to a "slow" device the **u** structure will cause the I/O to consume large amounts of system CPU time.

Filesystem sector sizes are a function of the physical devices used and will vary according to system hardware configuration. The filesystem sector sizes for **vonneumann** and **eagle** are:

- /big -- 4\*512 words
- /fast -- 1\*512 words

If transfers cannot be well formed, it is recommended that **unblocked** be used instead of **u** -- the former will use library-buffered raw I/O.

The buffer size (**-b sz**) for the library I/O buffer is declared in 512 word sectors. The default buffer sizes are in 512-word blocks. Sequential unformatted has 48 blocks; direct unformatted has 8 blocks.

Frequently, increasing the buffer size--provided that there is sufficient memory available -- will improve the user to system CPU ratio.



*[Clayton Guest](#) has been a member of the NAS HSP consulting staff since 1993. Previously, he was a user consultant at the Ames Central Computer Facility. Guest began his career at Ames in 1975 as a programmer. He has 30 years experience in large-scale computing.*

[Next Article](#)

[Contents](#)

[Main Menu](#)

[NAS Home](#)

[Next Article](#)[Contents](#)[Main Menu](#)[NAS Home](#)

# From NX to MPI: A Primer

*by Ed Hook*

The Intel iPSC/860 **lagrange** will leave NAS on October 1. By then, the approximately 275 users of that system will have moved to **babbage**, the IBM SP2 system that was recently installed at NAS.

This transition presents users with some work to do, since message-passing programs running on the iPSC/860 are written to use calls that Intel provides in its NX library. This library is specific to the iPSC/860, so these codes will need to be rewritten for the SP2.

The first decision that must be made is which of the SP2's message-passing libraries should be the target of the port. This article assumes that Message Passing Interface (MPI) has already been chosen by the user and presents a possible path through the porting maze. MPI is a good choice and, once you've gone through this exercise, you should never have to do it again.

## MPI Allows Portable Message Passing

Devised by the Message Passing Interface Forum as a proposed standard implementation of the "message-passing paradigm," MPI will allow users to write *portable* message-passing programs. It incorporates the important functionality of virtually every message-passing library that exists and adds a number of important concepts of its own, which are needed, for example, to allow the development of "safe" parallel libraries (whose routines can't fail because they inadvertently receive messages not intended for them). Before undertaking a port to MPI, take a look at the [MPI standard](#), which is available from the [NAS Documentation Center](#).

One MPI invention is the communicator, an object that incorporates a group of communicating processes and a context within which they communicate. It is the existence of this context that allows MPI to enforce the "safety" measures mentioned above. MPI provides many facilities for creating and managing communicators. The most important thing to know is that every MPI program begins life with a useful communicator already defined; named MPI\_COMM\_WORLD, its group is made up of all of the processes that, together, constitute the application, so it allows any two of these processes to trade messages.

## Initial Steps

MPI\_COMM\_WORLD does not exist when execution begins. It is created when you call MPI\_INIT, in

order to initialize MPI's internal data structures--this call should be one of the first actions in your program. When execution is complete, all processes must call `MPI_FINALIZE` in order to "clean up" after MPI. As a final bit of preliminary advice, always include the MPI header file `mpif.h` in your source code, since the definitions there allow you to make use of MPI's facilities while remaining shielded from the nitty-gritty of their implementation. A skeletal Fortran code that uses MPI on the iPSC/860 might look like this:

```
program MPIstuff
implicit none
include `mpif.h'
integer ierror
call MPI_INIT(ierror)
! do whatever here ...
call MPI_FINALIZE(ierror)
stop
end
```

where the recommended error-checking is not shown and the interesting "whatever"--what users are really interested in converting -- is left to the imagination.

## Getting From NX To MPI

Some of the issues involved in the NX-to-MPI conversion for a "hostless" iPSC/860 Fortran program are addressed here.

Think of the nodes in an iPSC/860 application as members of `MPI_COMM_WORLD`'s group. Each member of a group has a particular rank in that group, which can serve as a replacement for the node number. So, for the usual "environment" functions, there are translations, such as:

```
me=mynode(): call MPI_COMM_RANK(MPI_COMM_WORLD,me,ierror)
nn=numnodes(): call MPI_COMM_SIZE(MPI_COMM_WORLD,nn,ierror)
```

which also illustrate the characteristic presence of the variable **ierror**, in which MPI returns a success/failure indication. Careful programmers will actually test for the successful completion of each MPI request.

## Translation is Fairly Straightforward

What about the actual message-passing calls? The translation is fairly straightforward once you understand the slight difference in the ways that NX and MPI regard messages. NX views a message as consisting of its source, its destination, its tag, and the actual bytes of data being traded; MPI adds the communicator to this description. Ignoring some niceties, most NX message passing calls have almost-

unique translations. For example, the "synchronous"

```
call csend(tag,buffer,length,dest,0)
```

becomes, in MPI,

```
call
MPI_SEND(buffer,length,MPI_BYTE,dest,tag,MPI_COMM_WORLD,ierror)
```

where `MPI_COMM_WORLD` is used as the required communicator, the data in the buffer is treated as an array of bytes, and the omnipresent **ierror** has been added. If this same approach is adopted in every case, then most of the NX point-to-point message-passing calls can be translated mechanically. (In the long run, however, this is probably not the best approach. The MPI calls allow you to specify the actual type of data being communicated. Take advantage of this because it will allow transparent data conversion if you ever move the code to a heterogeneous environment.)

Various MPI calls, for example, `MPI_RECV`, return information in a status object that the user provides. In a Fortran program, this object is an INTEGER vector of rank `MPI_STATUS_SIZE`. Once it has been filled in for a pending or received message, it can be used to give the following translations:

```
len=infocount(): call MPI_GET_COUNT(status,MPI_BYTE,len,ierror)
src=infonode(): src = status(MPI_SOURCE)
tag=infotype(): tag = status(MPI_TAG)
```

This handles all of the **info** functions except **infopid**, which always returns 0 on an i860 node, so its absence is unlikely to be a problem.

What *might* be a problem is the fact that some of the NX calls do not have any translation into MPI. In particular, **hrecv**, **hsend**, and **hsendrecv**, which treat messages as interrupts, do not fit nicely into any portable framework, so codes that use them will need significant rewriting. On a slightly less sour note, there are two other NX message-passing calls, **isendrecv** and **flushmsg**, which do not have immediate replacements but can be accommodated. For example, **isendrecv** could be replaced by **csend**, immediately followed by **irecv**, then handled as above. **flushmsg** is a more difficult case, since its successful working requires familiarity with the way in which messages are handled by the operating system on the receiving node. Again, this is not portable and MPI can't give an exact equivalent. However, if it's a problem, the receiver can fetch the status of any pending message and then call `MPI_TEST_CANCELLED` to see whether its author issued an `MPI_CANCEL` in an attempt to expunge it.

## Calls Providing Global Operations

The other large family of NX calls that appear in many codes are those providing global operations, in

which each node contributes its local data. All of this data is subjected to some interesting transformation and the final result appears on all nodes at the end of the operation, overwriting the original data. NX has separate calls for each (transformation, datatype) pair, but MPI subsumes them all under `MPI_ALLREDUCE`. An example that gives the flavor of the thing is shown in the following NX call:

```
call gdsun(x,n,work)
```

which computes the elementwise sum of the double-precision vector **x(1:n)** across all nodes and leaves the result in **x(1:n)** on each node. Here, **work** is an array that must be big enough to hold a copy of the **x** vector, since it is used as a receive-buffer for the flurry of messages that must be passed around. The MPI replacement looks roughly the same:

```
call MPI_ALLREDUCE(x,work,n,MPI_DOUBLE,
MPI_SUM,MPI_COMM_WORLD,ierror)
```

where the type of element involved (`MPI_DOUBLE`) and the operation to be performed (`MPI_SUM`) appear as arguments, rather than being encoded in the name of the subroutine. It is important to note that this is *not* an exact translation, because the final result will appear in the **work** array on each node and the **x** array will not be disturbed. This is true in general, and it means that users will need to modify their code's logic, if it employs global operations. This incompatibility aside, all of the global reductions have immediate translations into various versions of `MPI_ALLREDUCE`. One further note: frequently, the result of a global operation is really only wanted and/or used on a single node. In that case, use `MPI_REDUCE` in place of `MPI_ALLREDUCE` and you'll see a boost in performance.

There are, of course, some global operations that are not reductions. The vitally important **gsync** call maps directly to MPI's `MPI_BARRIER`, but others are more problematic. For instance:

```
call gcol(x,xlen,y,ylen,ncnt)
```

where each node contributes its local **x** vector, these are copied into **y** in node-number order on each node and the number of bytes written into **y** is returned in **ncnt**. The closest that MPI can come to this is:

```
call MPI_ALLGATHERV(x,nx,xtype,y,ny,disp,
ytype,MPI_COMM_WORLD,ierror)
```

where **nx** is the number of elements in **x** (whereas **xlen** is the number of bytes), **ny** is an `INTEGER` array, with **ny(i)** the number of elements of **x** to copy from node **i**, and **disp** is another such array where **disp(i)** gives the displacement at which to store the data copied from node **i**. The significant point here is that the **ny** and **disp** arrays are inputs to the routine; this means that you must assemble these arrays on each node (unlike the situation with **gcol** where each node need only know the amount of data that it contributes initially). This also applies to the **gcolx** routine, which corresponds in the same way to MPI's `MPI_ALLGATHER`.

## Calls With Two or More MPI Calls

There are also calls whose translations involve two or more MPI routines. The first of these, **gsendx**, allows a message to be sent to a given subset of the nodes specified as a list. To do this in MPI, it is first necessary to call `MPI_COMM_CREATE` to create a special communicator whose group is the subset of nodes involved, together with the node wishing to send the data. This communicator can be used in a call to `MPI_BCAST` to send the message, after which it should be deallocated by a call to `MPI_COMM_FREE`. Note that this approach causes the originating node to be added to the recipients of the message; this is different from the NX semantics and will require logic changes.

Finally, there's **gopf**, which allows users to create their own reduction operation by specifying the operation as a particular type of subroutine. MPI provides the same functionality, but the low-level details differ significantly. First, the workhorse subroutine must be rewritten, since NX requires that it take two arguments (the usual **x** and **work** arrays--the routine must magically *know* their ranks) but MPI says that it must take four arguments (the arrays **x** and **work**, their common rank and their datatype). Next, call `MPI_OP_CREATE` with this rewritten function specifying the operation, and an operation "handle" is returned that can be used in a call to `MPI_ALLREDUCE` or `MPI_REDUCE`. When there is no further need for the hand-crafted reduction, clean up by calling `MPI_OP_FREE` to deallocate the handle.

[More information](#) for moving from NX to MPI is available on the World Wide Web.

For another perspective, request the videotape "Porting Parallel Applications from NX to MPI," a NAS Parallel Techniques Seminar by Bill Saphir, from the [NAS Documentation Center](#). For assistance, contact the NAS Parallel Systems Science Support group through NAS User Services at (415) 604-4444 or (800) 331-USER or send email to [nashelp@nas.nasa.gov](mailto:nashelp@nas.nasa.gov).

[Next Article](#)[Contents](#)[Main Menu](#)[NAS Home](#)

[Next Article](#)[Contents](#)[Main Menu](#)[NAS Home](#)

# RAID Technology Offers Cost-effective Data Storage for CRAY C90s

*by Elisabeth Wechsler*

Cost-effective data storage for supercomputers is closer to being realized, due to successful testing at NAS and subsequent adaptations to Redundant Array of Independent Disks (RAID) technology on the NAS CRAY C90. The disks are manufactured by Maximum Strategy Inc. (MSI) of Milpitas, CA, which has actively participated in the NAS project, along with C90 manufacturer Cray Research Inc. (CRI).

## Direct Impact on Users

RAID technology is expected to have a "very direct impact on users," said Bob Ciotti, of the High Speed Processor (HSP) Systems group, NAS Systems Development Branch, who has coordinated the project and testing at NAS up to this point.

In the months ahead, further refinements will be made and tested before the technology is fully implemented at NAS, but significant results have already been achieved, he said, including increased disk capacity and better data mobility across filesystems.

In addition, RAID technology has produced significant savings for NAS and has created competition among vendors. For the first time, supercomputer sites have an alternative source for disk storage, which Ciotti believes will continue to lower prices in the future.

## Disk Capacity Increased

Disk capacity on the NAS CRAY C90 (**vonneumann**) has been increased by a net gain of 125 gigabytes (GB), which will allow users quicker access to their working sets. The increase in disk capacity was achieved by adding 200 GB of RAID storage to **vonneumann** and simultaneously moving two DS40 systems to **eagle** (the Aeronautics Consolidated Supercomputer Facility CRAY C90). "Instead of pulling files off tape robots, the RAID technology will be accessed on **vonneumann**," Ciotti said.

With the new RAID configuration, data can be transferred more efficiently. For example, between home filesystems and /big or /fast, data will move at a rate of up to 80 megabytes (MB) per second, he said.

"Currently, if a user had a file on /m that was on disk and wanted to copy it to /big, you might see a rate

of 10 MB per second," Ciotti said. RAID technology has improved this rate by a factor of eight. "If a user is using the system interactively, this improvement is quite noticeable."

The four RAID systems needed for **vonneumann** are each composed of 40 Seagate Technology drives with a total capacity of about 50 GB. The initial development system cost about \$10 per MB but, with fierce competition in the personal computer disk drive market, MSI cut the cost of their system in half over the past year to less than \$5 per MB, Ciotti said. At the time of the procurement, the RAID disks were one-third the cost of comparable alternatives, he added.

## Vendor Competition

"The existence of competition in the vendor environment--where there wasn't any before--will continue to pressure the market to lower prices even further and will provide a very useful alternative for supercomputer disk storage," Ciotti believes. He commented that the development of other sources of disk storage amounts to an indirect technology transfer to sites supporting C90 computers.

"The availability of government funds and a willingness to explore the possibilities allowed us to take a big risk," Ciotti said. "NAS played a major role in getting these systems up and running."

While acknowledging pioneering efforts at the Center for Communications Research, Princeton, NJ, and the Minnesota Supercomputer Center Inc., Minneapolis, to advance cost-effective disk storage for supercomputing, Ciotti pointed out that both of these organizations used custom software, whereas the NAS project did not. "NAS is the first site to use a CRI-provided driver with MSI RAIDs. You can [now] get this environment off-the-shelf."

## HSP3 Procurement Provided Impetus

The impetus for the NAS project was an HSP3 procurement requirement to provide an Intelligent Peripheral Interface (IPI -3) software driver running over a High Performance Parallel Interface (HiPPI) channel. According to Ciotti, NAS was the first site to obtain this driver from CRI. A separate contract, awarded under full and open competition, was given to MSI to provide the disk arrays.

Ciotti pointed out that the CRI software was "extremely useful in enhancing the performance of the MSI RAID. It allowed us to direct work at the operating system level to the device best suited for the task at hand."

This technology combines the advantages of both the CRI and MSI devices by reconfiguring the filesystem according to what the components do best. The CRI manufactured DD60 disk drive is good at doing lots of small tasks quickly, while the RAID is good at transferring large blocks of data, Ciotti said.

## Other Improvements

However, there are other improvements that can be made but which CRI is reluctant to do because the modifications would not be generally applicable to CRI's proprietary drives, Ciotti added. "We're looking at ways to optimize performance beyond where we are today to deliver the best performance to NAS users."

Results from random I/O testing indicate that CRI proprietary disks worked better primarily because UNICOS 8.0, Cray's operating system for the C90, optimized their access. Applying the same organization techniques with other modifications to the RAID could bring all performance to within 10-20 percent of CRI's best performing disk, Ciotti said.

"CRI still has the edge for the highest performance disk systems," he said. "You just have to pay more for it and it's not clear that the additional performance would be effectively utilized in the NAS environment because we can work around the greater latency of the RAID systems."

Another area in which RAID technology offered improvement was fault tolerance, which MSI achieved in several ways. Due to the fact that RAID technology combines 40 physical drives into one logical unit, it is "imperative" that a failure of any one drive not destroy any of the user's data, Ciotti said, noting that "on the MSI system, up to four drives can fail without a user losing any data."

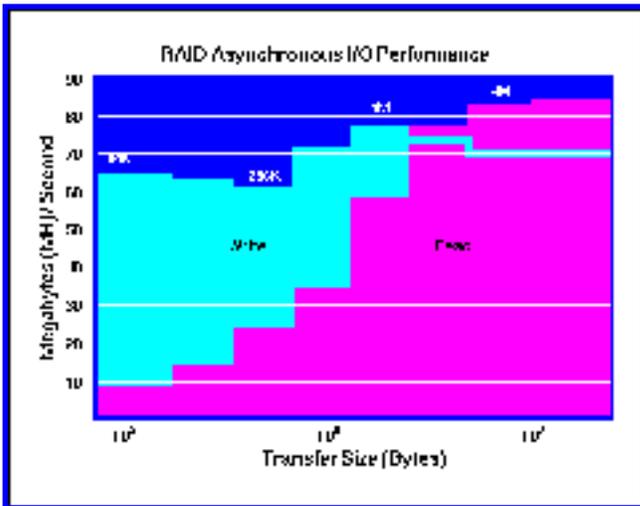
Hard failures of an entire disk drive are automatically replaced while the system is online--in a matter of minutes--and are transparent to the user, he continued.

## **Using RAIDs as /big**

"Currently, we're testing the RAIDs by using them as /big and are working on a proposed implementation plan that will be presented to a future meeting of the NAS User Group. The plan would create four large home filesystems and eventually combine the functionality of home and /m," Ciotti explained.

A team headed by Jim Crow, NAS Computational Services Branch HSP Manager, will take the results of Ciotti's work and investigate potential applications of the RAID technology at NAS.

For more information, refer to Ciotti's paper, "RAID Integration on Model E IOS," published in the Cray User Group (San Diego) Conference Proceedings, 3/94. To obtain a copy, send email to [ciotti@nas.nasa.gov](mailto:ciotti@nas.nasa.gov).



*This graph shows transfer rates on the Redundant Array of Independent Disks (RAID) using asynchronous I/O. Users should be able to achieve these rates from within their codes and with some of the system utilities that will be modified (for example, /bin/cp). Smaller transfers are slower, particularly for read operations. The filesystem structure that NAS plans to implement in its test prototype will place files of less than one megabyte (MB) on Cray Research Inc. DD60 disk drives, while files larger than one MB will be stored on the RAID devices.*

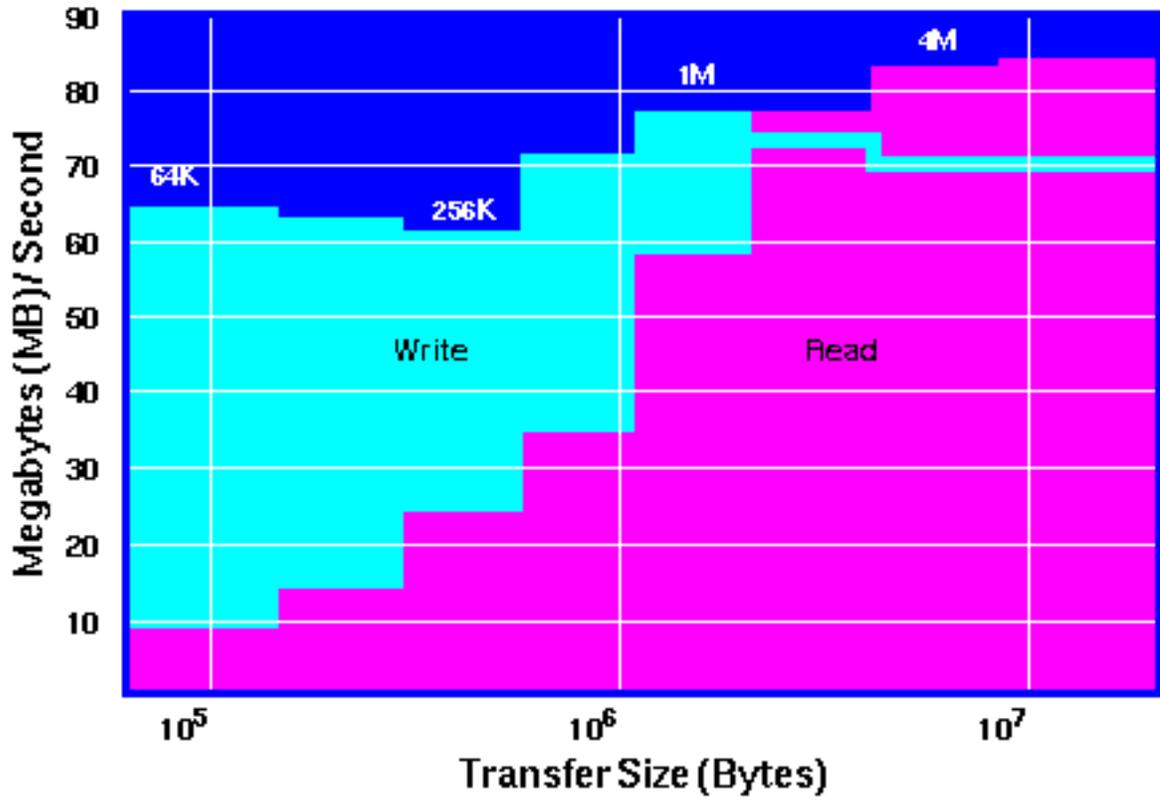
[Next Article](#)

[Contents](#)

[Main Menu](#)

[NAS Home](#)

### RAID Asynchronous I/O Performance



[Next Article](#)

[Contents](#)

[Main Menu](#)

[NAS Home](#)

# Virtual Reality, Interactive Media Highlighted at SIGGRAPH

*by Jean Clucas*

Visitors to the SIGGRAPH conference in Orlando soon discovered the "VROOM," or Virtual Reality Room, the latest major addition to this annual conference. The second floor of the conference was devoted to the VROOM, which highlighted scientific applications of virtual reality, using several different approaches.

One section, the "CAVES," consisted of dark rooms in which three Electrohome Marquis 8000 projectors cast stereo imagery onto the floor and walls, appearing to envelop users. Viewers, wearing Stereographics LCD stereo shutter glasses, took turns wearing a location sensor that controlled the display. The CAVES operated in "local" mode, showing a precomputed dataset and "interactive" mode, in which users could interactively steer a simulation running on an IBM SP2 or Silicon Graphics Inc. (SGI) Challenge symmetric multiprocessing servers. Audio was supplied through multiple speakers and the imagery for each room was created by an SGI Onyx with three Reality Engines.

A second approach to virtual reality was the "BOOM Room," which provided several displays using Binocular Omni-Oriented Monitor (BOOM) technology by Fakespace Inc. A BOOM monitor offers stereo display and is supported by a jointed arm, which allows it to be held close to the eyes and moved with the user.

The BOOM Room included the NAS Virtual Windtunnel, where users interacted with precomputed simulations of air flow around aircraft. Steve Bryson (NAS Applied Research Branch and Virtual Windtunnel team leader), Sandy Johan (NAS Systems Development Branch), and several student volunteers helped demonstrate the Virtual Windtunnel.

A third section of the VROOM, "Learn More," provided hands-on workstations running multimedia documentation of the VROOM. The presentation was viewed using the document browser Mosaic, a tool for retrieving and viewing multimedia documents from the Internet, underscoring the growth of the World Wide Web over the past year. Linking virtual reality, visualization, remote communication, and "the Web" was another hot topic at SIGGRAPH.

Steve Bryson organized and helped teach a course, "Developing Advanced Virtual Reality Applications." Other instructors were Steven Feiner, Columbia University; Randy Pausch and Dennis Proffitt, both of the University of Virginia; Henry Sowizral, Boeing Computer Services; and Andries van Dam, Brown

University.

Bryson also organized a panel, "Research Frontiers in Virtual Reality," in which Feiner, Pausch, and van Dam participated, as well as Philip Hubbard, Brown University, and Frederick Brooks, Jr., University of North Carolina at Chapel Hill.

"Virtual reality is today a technology that almost works...the resolution is so poor that the wearer has 20/200 vision [and] is legally blind. Trackers keep viewers tethered...[time] lags mean that scenes swim about," Brooks observed.

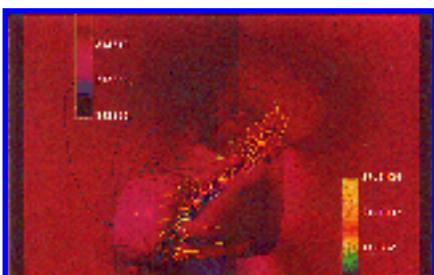
On the other hand, he commented that "people really do feel as if [the are] present, although no one is fooled." He predicted that "we will see high-resolution, low-lag systems doing serious applications within three years."

Other panelists discussed approaches to Brooks' aptly stated "getting past the `almost' stage." Hubbard proposed compromising detail or accuracy for speed, in situations where users wouldn't miss the loss of precision. Van Dam focused on the unique interface and object needs in virtual reality, and how current graphical user interfaces and object-oriented languages fail to meet those needs.

Pausch discussed the ways in which virtual reality must work with human kinesthetic sense. "By the use of grasping, we take advantage of our ability to perform three-dimensional manipulations." One of his suggestions was that truly manipulatable devices be integrated into the virtual environment. That type of kinesthetic interaction is already arriving in the CAVES, noted NAS Virtual Wind Tunnel team member Chris Gong. "The stair stepper in the `Stepping Into Reality' Cave helped make it feel more real. The more senses you can hit, the better." Developed by Edmund H. Baur of the U. S. Army Research Lab, "Stepping Into Reality" allowed users to "run" (using the stair-stepper) through a simulated terrain.

The prominence of the VROOM, with the extensive virtual reality installations as well as multimedia hands-on demonstrations, suggest that interactivity and virtual reality are being examined with a new seriousness.

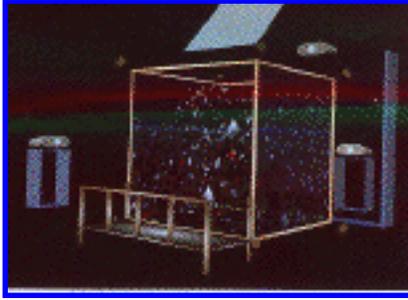
Bryson recalled a few years ago when counter-culture advocate Timothy Leary participated in a SIGGRAPH panel on virtual reality. "His presence was widely used to show what a flaky field this is. Now, we have people like Fred Brooks on the panel. This suggests the coming of age of virtual reality," he noted.



**Figure 1**

*Image representing the front of a high-speed train inside a tunnel. The surfaces represent cross-sections, iso-surfaces and boundary surfaces. The composite object is pseudo-colored with pressure and is composed with iso density lines colored with cross-flow velocity. The image was created with a custom C++ data visualization toolkit that emphasizes serial*

*composition of visualization techniques. The image (courtesy of SIGGRAPH) was created by Jean Favre, George Washington University.*



**Figure 2**

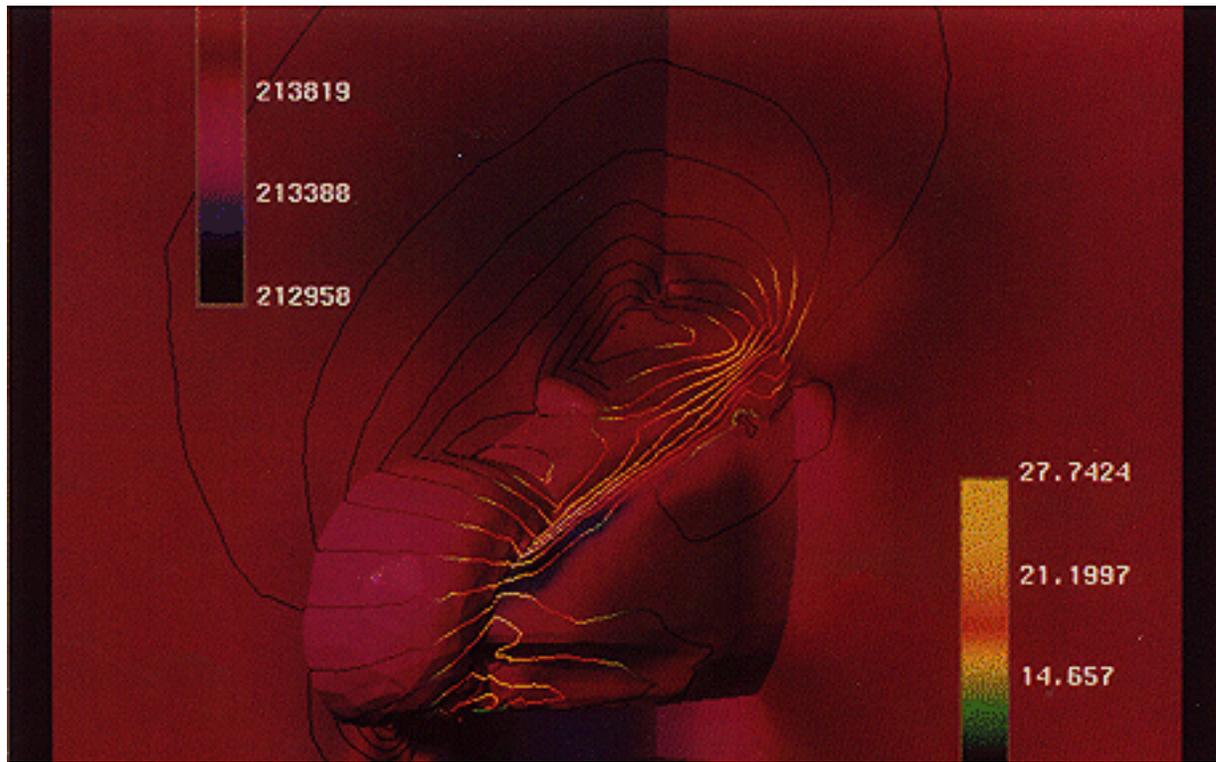
*A computer-generated rendering of the Cave, a 10 ft x 10 ft x 9 ft theatre with rear-projection screens for the walls and a down-projection screen for the floor. The illustration shows the projector layout, which, for the down projection screen includes a mirror. The Cave is a research and development project of the Electronic Visualization Lab at the University of Illinois, Chicago.*

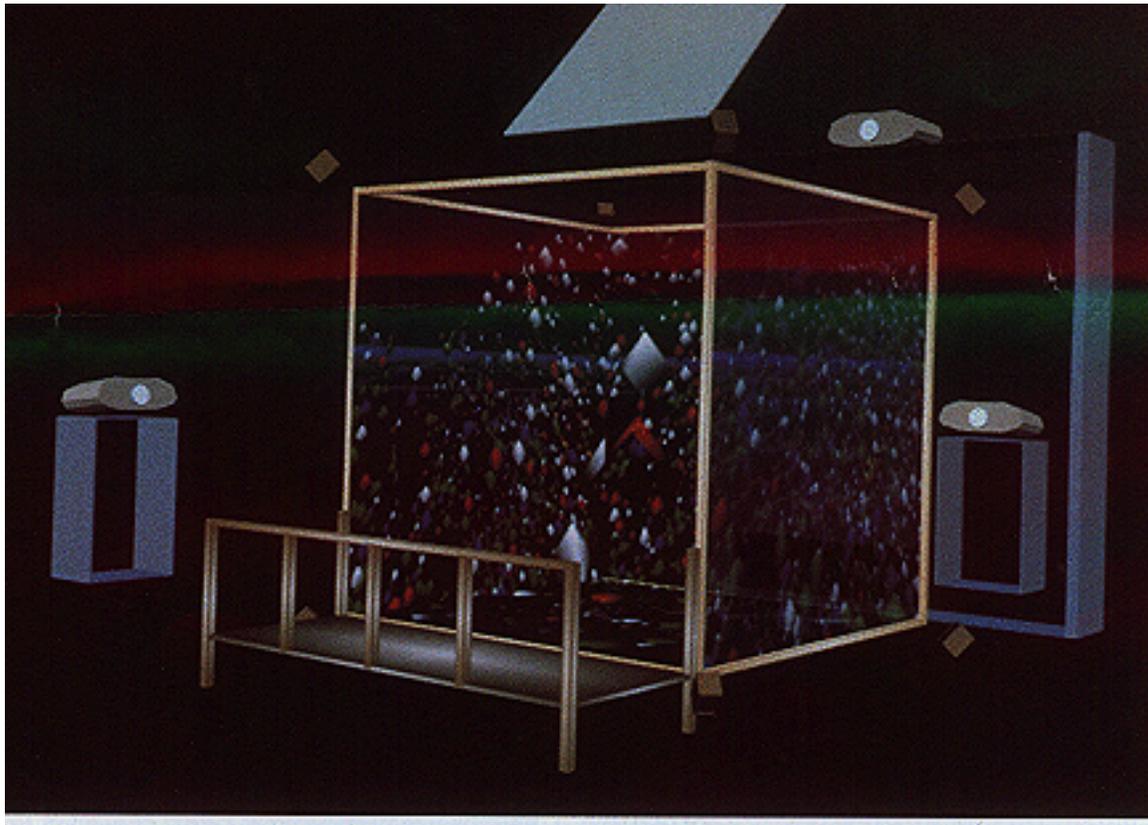
[Next Article](#)

[Contents](#)

[Main Menu](#)

[NAS Home](#)





[Next Article](#)[Contents](#)[Main Menu](#)[NAS Home](#)

# Linearizing Nested Loops

by *George Myers*

One measure of the performance of systems used for scientific research is the number of floating point operations performed per second. Users of **vonneumann**, the NAS CRAY C90, are achieving averages of between 250 and 300 megaflops (MFLOPS) per CPU. Each CPU is rated at 1 gigaflops (GFLOPS), so there is still some room for improvement. This information is extracted from the Cray Hardware Performance Monitor (HPM).

Another number that comes from the HPM data that is particularly revealing is the average vector length. One strength of the Cray systems is the vector processing units, which are capable of processing as much as 128 elements concurrently. The longer the vector sent to the vector units, the faster the processing. Vector length on **vonneumann** averages between 60 and 80 elements.

One simple technique for increasing the vector length in Cray Fortran is to "linearize" nested loops. This is accomplished by "over extending" the first dimension of a multidimensional array. For example:

```
real a(50,40), b(50,40), c(50,40)
do j = 1, 40
do i = 1, 50
a(i,j) = b(i,j) * c(i,j)
enddo
enddo
```

can be written as:

```
real a(50,40), b(50,40), c(50,40)
do i = 1, 50 * 40
a(i,1) = b(i,1) * c(i,1)
enddo
```

This increases the vector length from 50 to 2000, producing a speedup from 230 MFLOPS for the nested loops to 345 MFLOPS for the linearized loop. Be careful to test the results when converting loops in this fashion--the loop must have no dependencies.

The compiler can do this for you through the Dependency Analyzer, **fpp**. To use **fpp**, invoke the Cray Fortran Compiling System (cf77) by typing:

## cf77 -Zv prog.f

The **-Zv** flag instructs cf77 to invoke **fpp**, which, by default, converts loops as illustrated in the example above.

When linearized loops are not possible, make every attempt to index the inner loop on the longest array dimension, provided that this does not cause a data dependency.

More techniques to improve job performance on the NAS CRAY C90 will appear in future issues of *NAS News*. (See also "[Improving I/O Performance on the NAS CRAY C90](#)")



*George Myers, NAS HSP consulting group lead, joined the NAS Program in 1993. He has a broad range of computing expertise, including systems and software support analysis at Control Data Corp. Myers works for Sterling Software, and has provided user consulting at Ames for the last six years.*

[Next Article](#)

[Contents](#)

[Main Menu](#)

[NAS Home](#)

[Next Article](#)

[Contents](#)

[Main Menu](#)

[NAS Home](#)

# Users, Developers Converge at PTOOLS Consortium

*by Marcia Redmond*

The first general meeting of the Parallel Tools (PTOOLS) Consortium was held at NASA Ames Research Center, Mountain View, CA, June 8-10, 1994. Parallel tool users and developers from organizations that included Cambridge University, IBM, Fujitsu, Sandia National Laboratories, and Denmark's Institut fuer Informatik, gathered to discuss current parallel tools projects.

The meeting was sponsored by the Ames Information Sciences Division, the NAS Systems Division, and Recom Technologies Inc.

Throughout the three days, small group sessions were held, giving participants the opportunity to interact closely on topics such as:

- New Approaches to Parallelization (led by Sanjay Bhansali, Washington State University, and Rangaswamy Jagannathan, SRI International)
- Debugging Heterogeneous Systems (John May and Ming Hao, Hewlett-Packard)
- Managing Partitioning/Scheduling (Daniel Scales, Stanford University; Wolfgang Nagel, California Technical Institute/Research Center Juelich [Denmark]; and Tao Yang, University of California, Santa Barbara)

Other topics included "Scalable Parallel Unix Commands," presented by Rusty Lusk, Argonne National Laboratory, and "Distributed Array Visualizer," by Al Malony, University of Oregon.

A highlight of the event was the panel session "Enabling Parallel Applications: What Tool Support do Third-Party and Industrial Software Developers Need?"

The [PTOOLS Consortium](#) brings together representatives from the federal, industrial, and academic sectors to address issues about parallel users' needs in software tool support and how user feedback can be incorporated effectively into the tool development cycle. The goal of the Consortium is to provide a forum where developers and users can work together to identify tool needs and revise tools in response to those needs.

There are no proceedings for this event. For more information on software tools for parallel computing and how to become involved in the PTOOLS group, contact Jerry Yan at Mail Stop 269-3, NASA Ames Research Center, Moffett Field, CA 94035 or send email to [jerry@ptolemy.arc.nasa.gov](mailto:jerry@ptolemy.arc.nasa.gov).



*[Marcia Redmond](#) is the technical training coordinator for the NAS Program. She also provides editorial support for NAS News as part of NAS Technical Publications, and provides User Interface support by giving computer room tours.*

[Next Article](#)

[Contents](#)

[Main Menu](#)

[NAS Home](#)

[Next Article](#)

[Contents](#)

[Main Menu](#)

[NAS Home](#)

# Pat Elson, User Liaison `Extraordinaire`

*by Elisabeth Wechsler*

It seems that Pat Elson has been preparing for her job as NAS User Interface Manager all her life--and that's fortunate--because she brings a broad array of skills and experience to her role as liaison for approximately 1,400 NAS users.

"I see the job as being an advocate for our users," Elson said. "I try to help them find their way and to get them working." In that capacity, she deals with all NAS departments, as well as with experts from other government sites. A naturally empathetic and approachable person, Elson admits that patience, stamina, and a sense of humor are important in meeting the demands of her job.

Elson usually is the first person a new user talks with, and often there's a question, problem, or misunderstanding to be worked out. She estimates that she has contact with at least 30 NAS users a week through electronic mail and on the phone, and many more visitors to the NAS Facility in her ancillary role as tour conductor for groups ranging from the nation's top scientists to local fourth graders.

"If someone has a problem, we try to figure out what went wrong, then work to remove the roadblocks," Elson said.

Sometimes users are concerned about their proposal requests, and Elson said she tries to explain the criteria that NAS reviewers look for: "The projects must have technical merit, be of national significance, and be appropriate for NAS resources." In addition, she emphasized that "about 90 percent of NAS projects involve aeronautics."

Bicycling is her "passion" and only a recent foot injury keeps her from logging up to 200 miles a week. She belongs to several bike clubs, including Western Wheelers. Although Elson sometimes commutes the 12 miles to NAS on her bike, she prefers long-distance rides. In June 1993, she participated in the Los Angeles Grand Tour, riding 200 miles in one day. To prepare for long rides, Elson said that she could spend 40-50 hours a week behind handlebars.

## Crewed for Race Across America

This year she crewed for local cyclist Seana Hogan, who set a women's world record in the Race Across America--even after being hospitalized for dehydration during the race. Elson accompanied the ten-person, three-vehicle support team, providing Hogan with food, liquids, replacement bicycle parts, and--above all--encouragement for the nine-day, virtually nonstop endurance test. The race began July 30 in

Irvine, CA, and ended August 7 in Savannah, GA. Traditionally, neither the competitors nor their crews get much sleep during the race.

It's hard to know whether Elson's experience in helping her cycling competitor pace herself while also giving encouragement under brutally stressful circumstances makes Elson a highly effective User Interface Manager-- or vice versa. In any case, such skills certainly benefit both her professional life and her intense avocation.

As Elson sees it, she's been involved in training or customer service in some form since college days. Among her life trophies are: swimming instructor; residential treatment counselor for pre-teens; manager of a photo lab; number cruncher for an insurance underwriter; police dispatcher on the swing shift (3:00 - 10:00 a.m.); alarm monitor, crisis control person, supervisor, and trainer for a manufacturer of security devices; computer programmer and instructor; and past president of a local Toastmasters group.

In her off-duty hours, if Elson isn't "doing bicycles," she's probably enrolled in a seminar or community college course. "I'm an inveterate student," she admits. Recently, she received a certificate from the University of California Extension at Santa Cruz for completing a course in Program Management.

## **An Avid Student**

Her academic accomplishments include a Bachelor of Arts degree in psychology from Michigan State University; several Associate Arts degrees in art, architecture, and business administration from community colleges; a teaching certificate from Los Angeles City College; "lots" of computer classes at the University of California at Los Angeles; and a one-year internship in computer graphics at Foothill College, Los Altos Hills, CA.

Elson is married to David, a computer programmer, and lives in Santa Clara, CA, with "no children or pets." She's worked for Sterling Software at NAS since 1989, first as a member of the FAST (Flow Analysis Software Toolkit) team, and then moved to her present position in 1992. Elson was one of four NAS recipients of the 1994 Space Acts Award (*see the [July-August 1994](#) issue of NAS News*) for her work on PLOT3D.

In addition to serving as user advocate, Elson manages the submission process and publication of the *NAS Technical Summaries*. (Everyone allocated time on NAS supercomputing systems is required to submit a one-page summary of their work at the end of the operational year.) For the 1992-93 operational year, 133 summaries were selected for publication by a NAS internal review team.

One of Elson's goals to help implement "the paperless future" at NAS, as a sign at the entrance to her office attests. At this stage, she is focusing on the process of collecting and publishing the technical summaries online.

## Organizes Conference Efforts

Elson also organizes the NAS effort at conferences, such as Supercomputing '93 (Portland, OR) and '94 (Washington, D.C.), and the NAS User Interface Group meeting, held at the NAS Facility each January.

As busy as she is, Elson can always take a moment to laugh, share a kind word, or tell a funny story. For protection against taking life too seriously, she keeps a squirt gun and a collection of *MAD* magazines handy as home therapy after an especially grueling day.



*NAS User Interface Manager Pat Elson*

[Next Article](#)

[Contents](#)

[Main Menu](#)

[NAS Home](#)

[Next Article](#)

[Contents](#)

[Main Menu](#)

[NAS Home](#)

# NAS News

## Sep - Oct 1994, Vol. 2, No. 5

**Executive Editor:** Marisa Chancellor

**Editor:** Jill Dunbar

**Senior Writer:** Elisabeth Wechsler

**Contributing Writers:** Jean Clucas, Clayton Guest, Ed Hook, George Myers, Marcia Redmond

**Image Enhancements:** Chris Gong

**Other Contributors:** Ryan Border, Steve Bryson, Bob Ciotti, Jim Crow, Dave Cooper, Pat Elson, Jeanne Fouts, Sandy Johan, John Lekashman, Terry Nelson, Alan Powers, Bill Saphir, Dave Schmitz, John West

**Editorial Board:** Marisa Chancellor, Jill Dunbar, Chris Gong, Serge Polevitzky, Pamela Walatka, Elisabeth Wechsler, Rita Williams



# NEWS

Volume 2, Number 5

September - October 1994

## NAS To Join Ames Information Systems Directorate

by Elizabeth Westaler

A reorganization in progress at Ames Research Center is expected to have little impact on NAS users, except to make additional technologies more accessible, according to Dave Cooper, Chief of the NAS Systems Division since 1991 and a 10-year NASA veteran.

Effective October 1, Cooper will head Information Systems, one of four consolidated Ames directorates. The other directorates are Aerodynamics, Space, and Center Operations. Previously, Ames was divided into eight major entities.

### Focus on "Best" Technologies

The purpose of the reorganization has been to better focus the Center's activities on its customers, on the technologies it does best, and to make the organization more efficient and cost effective, Cooper explained.

NAS will become part of Information Systems, yet remain a distinct entity. "NAS is a very important organization—and only to Ames—but also to the NASA Office of Aeronautics," Cooper said. "Because of its impor-

tance and visibility, we've seen fit to keep the NAS organization as a separate division."

Information systems will include all computers and interconnections at Ames, most of the Office of Aeronautics supercomputing activities, artificial intelligence and expert systems, plan-of-the-element of computational sciences such as neuroimaging, robotics, and visualization technology. This organization will manage the Ames part of the High Performance Computing and Communications Program, including the Information Infrastructure Technology and Education and National Research and Education Network, as well as the Health Open System Trials and Surface Movement Advisor projects.

### Information Systems To Be Model

"We have a lot of activities and firm commitments with solid milestones on these programs," Cooper said. "What I envision for Information Systems is that we press the leading edge of these technologies and that we take and infuse those technologies into the rest of the aeronautics program at Ames,

and also Space and Center Operations. Once we've done that, I hope that we'll be able to serve as a model center for NASA to learn information systems or technologies can change the way we do business. I believe these changes can make us considerably more cost effective and efficient in what we do," he said.

Cooper noted that while a few functions within the NAS organization may be downsized as a result of increased budget cuts, "we'll try to minimize the impact of the reorganization on NAS users. The result will be transparent to our customers, except that there will continue to be more emphasis on technology transfer," he said.

### Chancellor Named Acting Chief

Marcia Chasickis, currently Deputy Chief, was named by Cooper to be Acting Chief of the NAS Division, also effective October 1, while the position of Chief is advertised nationally. It could take up to a year before this and other senior executive positions at Ames are permanently staffed, he said.

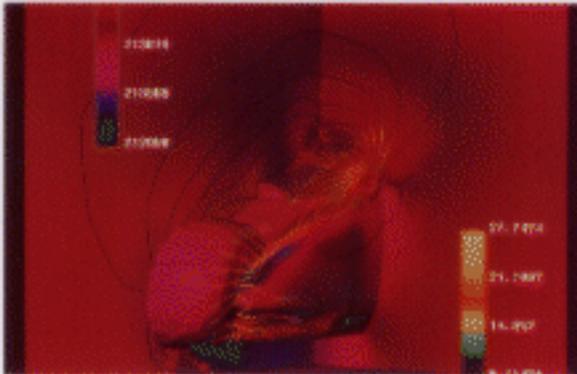
Continued on page 2

## SIGGRAPH Highlights Virtual Reality, Interactive Media

by Jean Lucas

Visitors to the SIGGRAPH conference in Orlando soon discovered the "VRROOM," or Virtual Reality Room, the most recent addition to this annual conference. The second floor of the conference was devoted to the VRROOM, which highlighted scientific applications of virtual reality, using several different approaches.

One section, the "Caves," consisted of dark rooms in which three Electrohome Marquis 8000 projectors cast stereo images onto the floor and walls, appearing to envelop users. Viewers, wearing stereographic LCD monochromatic glasses, wore hems wearing a location sensor that controlled the display. The Caves operated in "real" mode, showing a pre-computed dataset and "reactive" mode, in which users could interactively view a simulation running on an IBM SP2 or Silicon Graphics Inc. (SGI) Challenge symmetric multiprocessing servers. Audio was supplied



While researching the head of a high-speed train track in a cave? The surfaces represent cross-sections, iso-surfaces and boundary surfaces. The duplicate object is pseudo-colored with a coarse and is composed with two stereo lines rendered with cross-line stereo. The scene was created with a custom C++ data visualization toolkit that addresses such components of visualization technology. The scene probably of SIGGRAPH was created by Jerry Tene, George Washington University.

through multiple speakers and the imagery for each room was created by an SGI G45x with three Reality Engines.

A second approach to virtual reality was the "ROOM" Room," which provided virtual displays using Microsoft Direct-Ordered Monitor (DOM) technology by Falconview Inc. A ROOM monitor often uses a display and is supported by a jointed arm, which allows it

to be held close to the eyes and moved with the user.

The ROOM Room included the NAS Virtual Wind Tunnel, where users interacted with precomputed simulations of air flow around aircraft. Steve Dayton, NAS Applied Research Branch and Virtual Wind Tunnel team leader, Sandy Mohan (NAS Systems Develop-

Continued on page 6

### THIS ISSUE

Improving I/O Performance page 2

RAID Technology on C90 page 3

NX to MPI page 4