Status on the Serverization of UNICOS - (UNICOS/mk Status)

Jim Harrell, Cray Research, Inc., 655-F Lone Oak Drive, Eagan,
Minnesota 55121

ABSTRACT: *UNICOS is being reorganized into a microkernel based system. The purpose of this reorganization is to provide an operating system that can be used on all Cray architectures and provide both the current UNICOS functionality and a path to the future distributed systems. The reorganization of UNICOS is moving forward. The port of this "new" system to the MPP is also in progress. This talk will present the current status, and plans for UNICOS/mk. The challenges of performance, size and scalability will be discussed.*

1 Introduction

This discussion is divided into four parts. The first part discusses the development process used by this project. The development process is the methodology that is being used to serverize UNICOS. The project is actually proceeding along multiple, semi-independent paths at the same time.The development process will help explain the information in the second part which is a discussion of the current status of the project. The third part discusses accomplished milestones. These are points in our plans that we thought would prove significant portions of the technology. The milestones and their significance will be explained. Finally, the plans for future work will be discussed.

2 Development "Steps"

The first part of the project is to serverize UNICOS using CRAY EL and YMP systems to port the UNICOS features to the new system organization. The UNICOS feature code already works on these architectures, so the porting can proceed quickly. This allows us to concentrate on ensuring that the system architecture is kept consistent, and that the functionality of the UNICOS features is correct. As the features are added there is work to analyze the performance and make modifications to keep performance within acceptable boundaries.

The next step is to port UNICOS/mk or features in UNICOS/mk to the CRAY T3D. The CRAY T3D is being used because it is an available MPP architecture that can provide a good development environment until the CRAY T3E hardware becomes available.

Initially this concentrates on MPP machine-dependent modifications that have to be made in order to run UNICOS on an MPP system.There are also a number of MPP features that have to be added in order to provide required functionality, such as support for distributed applications. As the work of adding features and porting continues there is a testing effort that ensures correct functionality of the product.

Some of the initial porting can be and has been done in the simulator. However, issues such as MPP system organization, which nodes the servers will reside on and how the servers will interact can only be completed on the hardware. The issue of system organization is tightly linked to performance and scalability which are important factors to have evaluated as the port becomes more serviceable.

The results of early evaluations are being used to correct and tune the models of MPP system performance that have been built. We are then using the models to help us predict what development efforts should be undertaken to further enhance the system.

The last step in the current phase of the project is to port UNICOS/mk to the CRAY T3E. In the same way that the CRAY T3D port proceeds from the PVP work, this step follows the work on the CRAY T3D. This work is currently done in the CRAY T3E simulator. This port takes UNICOS/mk from the CRAY T3D and ports it to the CRAY T3E. Since the processor and MPP architectural constructs are essentially the same as the CRAY T3D, this work can proceed fairly quickly. In this step verification of functionality and performance/scalability evaluations are the most important aspects of the effort.

It is important to note that the initial functionality of the system, and each individual feature, proceeds through these steps independently. The initial base system has been the first through each step because it is needed as the base for all the other features. Each feature component of the system proceeds through the steps the higher level features following the more basic ones. Some of this work can and does proceed in parallel where ever it is possible.

3 Current Status

The status of the project is broken out by machine architecture. This reflects the methodology and also gives some insight about what feature content is likely to be in the near future.

3.1 Vector SMP Status

The initial port is complete. The system comes up multi-user and users can login, run processes, use the filesystem, etc. We are currently running development batch on an EL several hours a day. There are monthly Reliability Runs that verify functionality and allow us to ensure that there are no regressions.

The system currently has 167 working system calls out of UNICOS' 235. The network is up and almost all the use of the machines is across the net using TCP/IP. Basic tape support has been ported. We have begun to run applications, as tests. The application suites include DGauss, Fluent, and AIT. We will be increasing the number of hours of production and moving to larger machines as development progresses.

3.2 CRAY T3D Status

The initial port is in progress. We are able to run multi-user on a single CRAY T3D node, and have begun running several hours a day for user level porting and testing. The network is up and running. Again, in this environment, almost all the use of the machines is across the network using TCP/IP. The next step is to run UNICOS/mk on multiple nodes and start to distribute the OS servers across some of these nodes We expect to run distributed applications soon, and move to limited production in the next few months.

3.3 CRAY T3E Status

The initial port of the microkernel and servers is underway in the simulator. We expect to be able to run simple programs soon to verify that the initial port is complete. As the work on the CRAY T3D MPP features progress these will be added to the CRAY T3E port.

4 Accomplishments

In late 1994 we completed two important project milestones. The first was demonstrating Direct I/O on a CRAY T3D. Direct I/O showed that UNICOS on the CRAY T3D could access and run from a directly connected Model E IOS. This proved that the hardware connection from CRAY T3Ds to the IOS was functional, and also showed that the UNICOS I/O support was correctly modified to use the CRAY T3D IOG and CRAY T3D addressing.

Both raw and cached I/O were demonstrated. The demonstration was done by booting UNICOS into a single CRAY T3D PE and using standard UNICOS commands to access the filesystem on the Model E IOS disks. The root filesystem was also on the same disks so the entire system was running without the CRAY C90 used to boot it up.

The second milestone was demonstrating multiple CRAY T3D PEs running UNICOS. This demonstration was important because it proved that the microkernel and UNICOS were capable of running on multiple PEs and supporting processes on any PE. It also demonstrated that either cached or raw I/O could be directed to the correct processes I/O buffer independent of the PE address.

5 Work in Progress and the Future

We are currently looking at the issues related to MPP system organization and the related work involved in performance, size, and scalability. In this work the server layout, or organization of the servers, on the Operating System (OS) nodes has started. This work is directly related to scalability. As mentioned previously, we are using models to help us determine where the bottlenecks in scalable performance are located.

Related to OS server layout are the decisions about what services will be made available in the compute nodes and what this does to both size and performance of the compute node and system. Adding services increases the size, and many services would be used infrequently. However, we want to ensure that I/O performance is adequate to support the application and processor power. This may lead to providing some simple I/O capabilities.

For MPP I/O we have been investigating how to make available the CRAY T3D and T3E hardware performance enhancements. One particularly interesting hardware feature is the MPP centrifuge which allows data blocks to be "gather/scattered" from a number of PEs. We are working on a design of a set of listio(2) extensions (which we call distio) that would allow a distributed application to issue a single I/O request with data buffer addresses from a number of PEs.

The read version of the request would access the data on a disk device and then, using the centrifuge in the CRAY T3E hardware, distribute chunks of the data to each PE's data buffer. The advantage is that the application's PEs can make fewer OS requests for data and continue processing.

The work that will be accomplished over the rest of this year is to complete the CRAY T3D port and demonstrate a working prototype. We will complete the CRAY T3E port and be prepared for CRAY T3E hardware when it is powered up.

We will continue to look for opportunities to take advantage of MPP capabilities so that we can deliver the best performance on those machines. We will continue to maintain commonality with the vector architecture machines because this is both our development base, and prepares UNICOS/mk to replace UNICOS on the next vector machines. Finally, we will be working towards shipping UNICOS/mk with the first CRAY T3E.