SLAC-PUB-5292
August, 1990
(M)

Cross System Extensions (CSE) Experience

Ted Y. Johnston

Stanford Linear Accelerator Center *
Stanford University, Stanford CA 94309

ABSTRACT

Cross System Extension (CSE) provides VM/XA systems with the ability to share minidisks and spool in a loosely coupled environment. CSE will also cooperate with the VM/HPO Inter System Facility (ISF) in sharing minidisks between VM/XA and VM/HPO. -- This paper will discuss the use of CSE for migration from HPO to XA, reliability of CSE, and some--operational considerations when running with it.

Presented at Share 75, New Orleans, LA, August 12-17, 1990

* Work supported by Department of Energy contract DE-AC03-76SF00515

Introduction

The Stanford Linear Accelerator Center is a laboratory funded by the U.S. Department of Energy to perform High Energy Physics Research. High Energy Physics is the study of-the basic particles that make up all matter and requires large and sophisticated detectors that generate vast amounts of data (around 2 terabytes per year for a modern detector). The central computer facility at SLAC as well as providing general time sharing services to the laboratory has as a primary function providing the computing horsepower to reduce and analyze this data. The central computers are a 3090 200E and a 3081K sharing about 80 gigabytes of DASD (see appendix A for configuration). All tapes are also shared between the two systems.

SLAC has been running shared DASD systems since 1974 when ASP (the predecessor of JES3) was installed. In 1983 SLAC participated in a joint study with IBM and installed Multi-Access Spool (MAS) from Yorktown. MAS was the prototype for ISF. In 1988 we converted to SSI from VM/CMS Unlimited, and in 1990 made XA with CSE our production system.

SLAC's mix of work is somewhat different from the average VM shop. The SLAC systems run close to 100% cpu utilization most of the time. Over 80% of the cycles go to batch work. Consequently, all of the interactive users are on the front end 3090 and the 3081 only runs batch, All batch work-is done in VMs managed by our batch monitor. These VMs like all other VMs at SLAC run CMS — we do not run any guests.

I am going to talk about four different aspects of CSE:

1. CSE as a conversion and migration tool
2. CSE Stability and Performance
3. Functionality
4. Operational Considerations

Conversion

The SLAC physicists frequently have very complex environments with around 20 disks accessed at the same time. "Magic" execs establish their environment and provide capabilities. Because of this, it is very important that minidisk links are respected across the complex. Since 1983 we have provided the same protection for LINK in a multi-cpu as in a single cpu environment. We felt that the XA testing must also be able to coexist with HPO production in the same manner. Because of capacity problems we bought a second hand 3083 (very cheap!) as the XA development

machine. ISF and CSE minidisk sharing are compatible, but we were running VM/CMS's SSI. So we ordered ISF and retrofitted its disk sharing into SSI and then installed CSE on XA. This provided full protection of the minidisks and allowed users to test using any disks that they desired without concern. We ended up with the following test environment:

1. Fully protected shared DASD.

2. The same CP directory in XA and HPO (the XA directory might be an hour or so out of date).

3. Ability to have spool files move from the XA system to the production system for printing.

4. Ability to submit-batch work on any cpu to run on any cpu.

5. Fully integrated running of 3420 and 3480 tape jobs on XA.

6. Terminal access either by Passthru or TCPIP.

We think that using a separate system with CSE was a very effective conversion strategy. With the exception of data base applications, we did not have to set up special disks for people. We did not have to give them special instructions about how to do their work. We did not have to spend time in getting our production HPO system running under XA. We had much greater flexibility in having XA outages than if we had our production system running under it. We frequently had XA outages over lunch to install new software, something we couldn't even have thought of, if our production system was running under XA.

Stability and Performance

In January we ran an open system test of XA on the 3090 and 3081 from Saturday through Monday. There were 7 crashes during this test 3 of which were caused by CSE. All of these problems were corrected, and on February 21 we made XA the production system. In the first 15 days after the cutover there were 6 crashes 3 of which were caused by CSE. From March 8 to the present there have been no system crashes caused by CSE.

In addition to the early system stability problems, we have also had problems with the 3800 support. One of the crashes during the system test was caused by use of the delayed purge queue. The APAR is still open. Although, the system no longer crashes there are still CSE side effects that have not been resolved. Since XA issues a clear print command to the 3800, the delayed purge queue is not required. SLAC has installed a local fix to bypass the use of the delayed purge queue and its side effects.

A number of other CSE related problems have been fixed by IBM. SLAC has a local fix installed for a problem with losing spool files when a system restarts.

CSE uses PVM to provide the communication path between systems. We have had very few problems with PVM. We are running version 1 release 4 and are using the intercpu IUCV support that is part of that release. Occasionaly, if the backend system crashes, it is necessary to restart PVM on the front end system to resynchronize the systems. PVM uses about 4.5 more cpu hours per month with CSE than it did previously.

Functionality

CSE provides the following major functions:

1. Shared Spool
2. Shared Minidisks
3. Transparent Query, MSG, and SMSG

Dirmaint release 4 augments CSE's multi-cpu support by providing -- synchronized directories on all machines.

CSE misses several valuable functions that were part of SSI and affect running a multi-cpu complex:

1. It has no multi-cpu VMCF support. Many years ago we wrote our own non-transparent support that uses RSCS CTC support to communicate. However, this only works for processes that code for the non-transparent interface.

2. It has no transparent multi-cpu IUCV support. The support that exists in PVM is almost not documented. Processes such as OCO database systems from various vendors that use IUCV cannot be supported properly since interface changes are required.

3. It does not provide a general ability to execute CP commands on any processor. SSI provided the SSI command which allowed a CP command to be broadcast to all cpus or to any specific CPU and the responses returned to the issuer.

4. It does not provide TOD clock synchronization such as is provided by SSI.

Operational Considerations

Once there are multiple systems interconnected one needs to know where one is running. Software such as program products may be licensed on specific cpus, other things may be associated with logical CPUs. We wrote a single exec to provide a set of functions for determining system configuration such as: what logical CPU am I running on, what physical CPU am I running on, what logical CPUs make up the complex, what physical CPUs make up the complex, and what logical machine is MASTER. Under SSI from VM/CMS Unlimited the logical CPUid of the machine was returned as an extra token on "CP QUERY USER userid". We extended CSE to also provide this information for compatibility and increased functionality.

When we started we used different nodeids and special execs to get the same named virtual machine running on various processors. SSI required that all userids in the complex be unique, consequently, we changed all service VMs to have unique names. Under CSE we have continued to keep the names unique as it is much less confusing and the service VMs do not end up in the exclusion list. The exclusion list defeats shared spool for the VMs in the list and that is more of a hastle than unique names. Service VMs (such as SMART) where one will exist on each processor are named SMARTA, SMARTB, etc. The "A" or "B" refer to the names of the logical machines where they run. In the case of machines like SMART both machines will share the same minidisks via LINKs in the CP directory. The only virtual machines in the exclusion list are the two PVM machines (which is required by CSE). The system operator is supposed to be in the exclusion list, but because of a bug when we first brought up CSE, our operators (PROPVMA and PROPVMB) were not in the list. Although the bug has been fixed, we have not installed the fix, as we see no need to put these machines in the exclusion list.

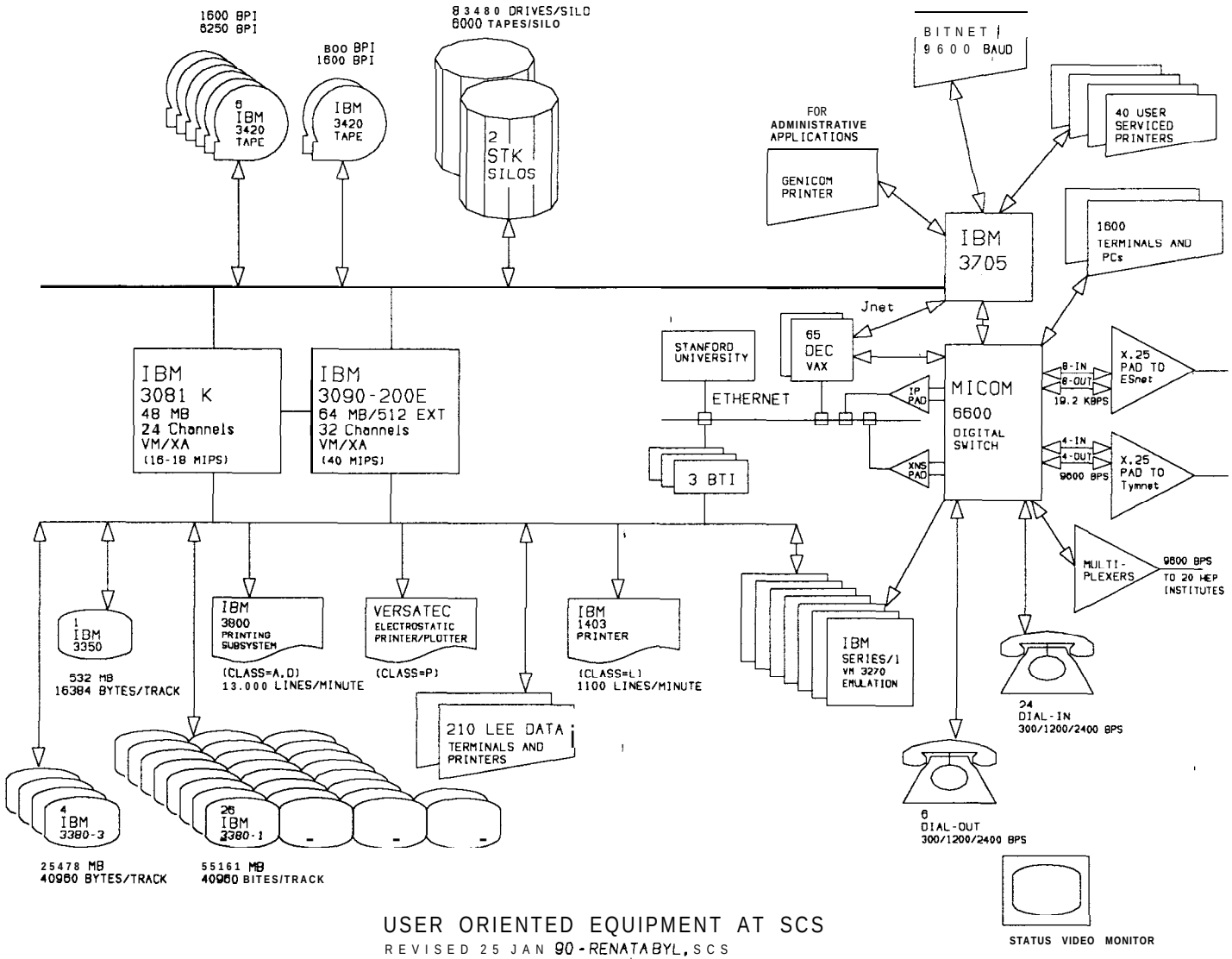
The entire central complex is externally a single node known as SLACVM (the nodename returned by IDENTIFY is SLACVM everywhere in the complex). A single RSCS on one of the processors services all RSCS traffic to the outside world and printers. Only one RSCS is needed because of global MSG, SMSG, and shared spool. Although, CSE requires that the cpus be named differently that is only visible to the CP Q USERID command and in the lower right hand corner of a 3270 screen.

An area of concern was managing DCSSs and NSSs. Under HPO we shared the saved system area between systems, but CSE does not support sharing of NSS files. Since NSSs can be dumped to tape and we have an automated tape library, we were able to easily setup virtual machines to dump the new NSSs from the front end system and restore them on the other system.

IBM says that you should not do minidisk caching in a shared DASD environment. That is generally true, but not completely. Since our backend system is a 3081 it cannot cache anything. Consequently, minidisks that are only updated on the front end system are safe to cache. Minidisks like the S and Y disk and group disks are only updated on the front end system; so we have enabled caching for those minidisks. We have changed the default to not cache all minidisks. Because of the caching we have not had to go to an alternate copy of the Y disk under XA such as we had under HPO.

Conclusion

CSE provides adequate function for SLAC and is stable. We miss some of the functionality of SSI. If we were running a more symmetrical system with connected users on all systems, life would not be as easy.



USER ORIENTED EQUIPMENT AT SCS

REVISED 25 JAN 90 - RENATA BYL, SCS

Appendix A