A Proposal to Rightsize VXCERN with a Scaleable, Mixed Architecture VMS Cluster

Julian J.Bunn
CN/DCI/DEC

8–November–1993

1 Introduction

With a count of around 2500 different users in a week, and up to 300 concurrent logins at prime-time, VXCERN[1] rates as one of the most popular services run by the Division. This is due in part to its rich installed software base, to its definitive VMS version of the CERN Program Library, and also to its excellent network connectivity.

Traditionally, VXCERN has always been grossly overloaded, any increase in capacity being rapidly absorbed by user demand. It is interesting to speculate about what additional capacity is required in order to satisfy the users' appetites. In this paper, we describe an upgrade which we believe will finally achieve the goal of meeting the demand with a scaleable configuration, at a very moderate cost.

2 Why Rightsize ?

Despite the commonly held view that there is no very long term future for large scale services based on VMS, the Division has publicly committed to running a VXCERN service for the lifetime of the LEP experiments (at least until 1998). The VAX mainframe and associated HSC hardware is rapidly becoming outdated and is starting to look more and more expensive from the point of view of maintenance costs. The implications of *not* upgrading the existing hardware are financially very unattractive in the medium term[2].

The rightsizing proposed in this paper will leave us with a Mixed Architecture VMS Cluster (MAVC) of VAX and Alpha workstations. A VAX component is retained since VAX binary compatibility will be essential for some applications during the transition phase to Alpha. We describe in detail the configuration of the target MAVC, and show costs for hardware and software. The substantial maintenance savings that will be achieved by removing the old equipment are enumerated. **The major cost saving is achieved by replacement of the RA-series disk subsystem by modern, cheap SCSI units.** We also describe some quantitative results from the first phases of implementation of the new equipment.

The equipment purchases can largely be funded by the saving in 1994 maintenance costs achieved by removing the old equipment at the Christmas Shutdown at the end of 1993.

3 Rightsizing Strategy

The upgrade of VXCERN is towards a **scaleable** configuration of machines, at as small a cost as possible. Experience with the CORE Project indicates that UNIX/RISC workstation clusters can successfuly replace traditional mainframes. The availability both of Alpha RISC workstations running VMS, and of VAX workstations offering unprecedented CPU power[3], means that VMS workstation clusters are now also a cost-attractive alternative to the VAX mainframe.

Based on this premise, the next section describes in detail a new VXCERN configuration, which will provide the user community with a decent service, and which in the future will allow us to cheaply and easily add (or withdraw) CPU and storage capacity as the popularity and requirements for VMS dictate. The plan is for a CORE-like Cluster of VAX and Alpha workstations, equipped with SCSI disk arrays, and clustered across a DEChub 900 attached to the main GIGAswitch.

---

[1] Until recently a VAX 9000-410 and a VAX 8250.

[2] It is well known that maintenance costs for "out of catalogue" DEC hardware are punitive.

[3] The VAXstation 4000 model 90 rates at around 8 CERN Units, and a base configuration costs roughly 15000 SFR.

3.1 Target Configuration

In the table below, we present the hardware and software configuration and costs of the upgrade equipment. We indicate which pieces of equipment are already installed in the Cluster, and which are on order.

Table 1: Upgrade configuration and costs

Item	Cost in SFR	Comments
1 VAX workstation[1]	30000	Installed
1 Alpha "Sandpiper" workstation[2]	30000	Installed
1 Alpha "Sandpiper" workstation[2]	34500	Ordered
1 Alpha "Flamingo" workstation[3]	37000	Installed
Maintenance on 4 workstations	0	Maintenance starts end 1994
72 GBytes SCSI disk[4]	149000	Ordered
Maintenance on SCSI disks	0	No maintenance required
DEFTA FDDI boards for 4 stations	15000	Ordered/Installed
DEChub 90, FDDI equipment[5]	5000	Ordered
Unlimited VMS licenses for 3 Alphas	65000	Two installed, one ordered
Language licenses for 4 workstations	8000	Three installed, one ordered
System software (e.g. Multinet)	5000	Probably less
Software maintenance on the above	0	Included in Campus
Miscellaneous disks, cables, hardware	30000	E.g. 5 GByte of disk purchased with[3]
Totals	408500	Estimated Cost in 1993

[1] VAX Station 4000 model 90, rated at 8 CERN Units. 128 MBytes memory. Previously CO Group office equipment.

[2] DEC 3000-400 Alpha AXP workstation, rated at 17 CERN Units. 128 MBytes memory.

[3] DEC 3000-500 Alpha AXP workstation, rated at 20 CERN Units. 128 MBytes memory. Purchased at a special price.

[4] RZ 28 2GByte disks, fully housed and cabled in StorageWorks enclosures, with cabling.

[5] The DEChub90 supports up to 8 lines

When reading the table, the following points should be borne in mind:

- There is no maintenance charge for new equipment for the first year after purchase.
- It is proposed to pay no maintenance on SCSI disks; they are reliable, and several spares would be kept. (In any case, SCSI disks bought ex. DEC are covered by a 5 year guarantee. DEC now measure the MTBF of their RZ28 2GByte disks as being well in excess of 350,000 hours.)
- We assume that the DEC Campus scheme will be in place by the start of 1994: this will remove any software costs from consideration.
- It is arguable whether maintenance is necessary on reliable items like workstation servers. However, we would feel much more comfortable having it.

- No account has been taken of possible trade-in discounts, or special price negotiations. There is clearly a possibility of selling the current VXCERN equipment, once it is liberated by the upgrade. However, after contacting brokers in the U.K. we expect to recuperate very little monnies from such a sale, and DEC are certainly not interested in a buy-back.

The new configuration features disk capacity more than double that available today on VXCERN. The scalability of the cluster allows us to add or withdraw individual nodes as requirements dictate. This might be particularly useful if an experiment wishes to add its own node to the cluster, or if we wish to eventually accomodate part of the VXENG community on dedicated nodes in the cluster. Access to individual nodes can of course be restricted to certain user groups, if necessary. The hardware is all re-usable if the cluster were ever to be dismantled.

The total CPU power in the cluster is around 60 CERN Units, over six times the power available today on the 9000, and this could of course be augmented or decreased by adding or removing workstations.

3.2 Cost of Retaining the Current VXCERN Hardware

In the table below, we detail the cost for retaining the existing VXCERN hardware into the year 1996. We have assumed a conservative 10% increase in maintenance costs after 1994. We do not include the TA90 subsystem.

Table 2: Costs per year to retain VAX 9000, peripherals and software

Item	1994	1995	1996	Total for Period
VAX 9000 maintenance	120000	132000	132000	384000
HSC Disk controller maintenance	82500	90750	90750	264000
RA Series disks maintenance	41000	45100	45100	131200
Totals	243500	267850	267850	779200

3.3 Manpower Costs

We believe that the number of staff currently engaged in support for the VXCERN (and VXENG) clusters will continue to be sufficient to support the proposed configuration. We do not expect that the upgrade will liberate any of those staff for other activities.

4 Staged Rightsizing, and Current Status

The rightsizing is proceding in stages. This allows a re-evaluation of the system at each stage in the light of its performance. It also allows valuable experience to be gained in tuning the systems appropriately as a function of the Cluster size. **We are planning on having finished the rightsizing by Christmas 1993**, although system tuning will continue well into 1994.

4.1 Cluster Configuration

At present two of the Alpha workstations, and the VAXstation, are installed in the VX-CERN Cluster. The cluster traffic is routed over FDDI. We are progressively reducing the allowed number of logins to VXCRNA, in order to squeeze the user community onto the Alpha systems (and we are increasing the login limits on those machines). We are awaiting a **multi-user license for the VAXstation**. Before we have this license, we cannot remove VXCRNA as this would leave the Cluster without a VAX node. We have added the old VXCRNZ machine (the ex-Cray gateway) into the Cluster purely as a trick to enable preservation of disk shadow sets. This machine is not available for logins, and will be removed from the Cluster when the VAX 9000 is removed.

4.2 Interactive Response Time

Presently, around two thirds of VXCERN users are still on VXCRNA, with the remainder using the Alpha machines. Measurements of the interactive trivial response time on the Alpha machines as a function of the number of concurrent interactive users show that performance is still very good on each machine at least up to a limit of 70 simultaneous sessions. The table below shows the data for the Alpha workstations and for the VAX 9000. The figures given are the average times to complete a series of trivial commands in a dcl script. Data were collected continuously over a period of several weeks. Errors on these data are typically around 10%. It should be kept in mind that in the current configuration, the 9000 is serving the disks to the rest of the Cluster.

Table 3: Trivial Reponse Time vs. Number of Interactive Users

Number of Users	20	30	40	50	60	70	250
Alpha 128 MBytes	61	71	86	95	114	109	no data
Alpha 64 MBytes	67	86	100	150	no data	no data	no data
VAX 256 MBytes	83	111	116	139	167	233	2700

We thus expect good response in the right-sized VXCERN cluster for a total of at least 4*70 = 280 interactive users, compared with very poor response on the VAX 9000 with 250 simultaneous users today.

4.3 SCSI Disks

Our intention is to **attach all user SCSI disks on the Flamingo machine (AXCRNC)** in the Cluster. This machine has the required number of SCSI busses. We prefer to do this rather than spread the disks evenly around the Cluster nodes, as in that case the whole Cluster can become unavailable when any one node fails. Our plan is to make expedient image copies of all the existing RA disks onto SCSI counterparts, once the SCSI disks have arrived and have been installed. The disks ordered have a capacity of 2 GBytes, which corresponds to over 4 million VMS blocks. The largest RA disks we have in the Cluster today have a capacity of just over 3 million blocks. For certain key disks, such as the CERN disk, the CERN Program Library disk, the Scratch and Stage disks, we will take the opportunity of increasing considerably the available space. In addition, each node in the Cluster will contain an internal page, swap and dump file disk. This will mean that, in the unlikely event of one of the nodes "smoking", we can replace it without having to re-install system software.

There have been a total of 36 disks ordered in two "lots". Unfortunately, **delivery dates for the disks have slipped into mid-December**. If we are unable to persaude DEC to deliver sooner, then the resulting delay might jeopardize the feasability of finishing the rightsizing at Christmas.

4.4 Tapes and Staging

We are expecting to be able to remove all the TA90 cartridge drives from the Cluster. The only tape drives remaining in the Cluster will then be those in the Exabyte robot, currently used for backup/restore. **All user tape access will thus be through the TPREAD/TPWRITE/RFIO chain being developed by F.Hemmer under the CORE umbrella**. This will require a substantial increase in the size of the disk$stage pool on VX-CERN. If this software is not available in production by Christmas then we will be unable to remove the VAX 9000 from the Cluster: we would have to move one or more TA90 drives into a peripheral controller on the VAX. Users on the Cluster are already denied access to the new SMCG cartridge robot. System backups will continue to be done using the Exabyte robot, and there are plans to offer some user slots in this device, if it can be moved to the Vault.