A Proposal to Upgrade VXCERN with a Scaleable Mixed Architecture VMS Cluster

Julian J.Bunn
CN/SW/DEC

Tim Whibley
CN/CO/DEC

2–November–1993

1 Why Upgrade VXCERN?

With a count of around 2500 different weekly users, and 250 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.

Unfortunately, the system has been grossly overloaded, making the quality of the service less than satisfactory. In addition, the VAX mainframe and associated HSC hardware is rapidly becoming outdated and is starting to look more and more expensive from the point of view of maintenance costs.

Despite the possibility of there being no long term future for large scale services based on VMS, the Division should not be complacent about running second-rate services like this. Moreover, at the 1992 POW meeting it was stated that the Division is committed to running a VXCERN service for the lifetime of the LEP experiments. In this case, the implications of *not* upgrading the existing hardware are financially very unattractive in the medium term[2].

This paper proposes an upgrade for the VXCERN Cluster into a Mixed Architecture VMS Cluster (MAVC) of VAX and Alpha workstations, at a moderate cost. We need to retain a VAX component since VAX binary compatibility will be essential for some applications (like the VAX version of the Program Library) during the transition phase to Alpha. We describe in detail the configuration of the target MAVC, and provide cost estimates for hardware and software, taking into account the substantial maintenance savings that will be achieved by removing the old equipment. Since the original version of this paper was written, several hardware and software components of the new configuration have already been purchased and put into operation: we describe some quantitative results so far.

The proposed upgrade is cost-neutral over two, or at most three, years, starting from 1994. The major cost saving is achieved by replacement of the RA-series disk subsystem by modern, cheap SCSI units. In fact, the equipment purchases can be funded by the saving in maintenance achieved by removing the old equipment at the Christmas Shutdown at the end of 1993.

2 Proposed Upgrade

The upgrade of VXCERN should be 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 be successful in replacing 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 also a cost-attractive alternative to the VAX mainframe.

Based on this premise, the next section describes in detail an upgraded VXCERN configuration, which will provide the user community with a decent service, and which in the future will allow CERN 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. A short history of the VXCERN service may be found in Appendix A

[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.

2.1 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. 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.
- Maintenance of the FDDI interfaces is included in the maintenance cost of the servers.
- 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.

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	603000	28000

[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

The configuration has the following features:

1. Its total **disk capacity** is more than double that available today on VXCERN.
2. Access to individual Cluster nodes can be restricted on a group basis (which satisfies the **"Group Server"** concept).
3. All the **hardware is re-usable** elsewhere (except maybe the VAXstation), if it is decided to end VMS support in the Centre, since the Alpha machines can run OSF or Windows/NT, and the SCSI disks can be attached to most other workstations.
4. User-groups may buy their own VMS machine and SCSI disks to add to the Cluster, so benefitting from a **centrally managed and backed-up** environment.
5. If the demand for VAX hardware disappears, then the VAX station(s) may be removed.
6. The total cluster CPU power is around 60 CERN Units, **over six times the power** available today on the 9000.
7. The CPU power could be augmented or decreased by adding or removing workstations.

2.2 Cost of Retaining the Current VXCERN Hardware

In the table below, we detail the cost for retaining the existing VXCERN 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
Software maintenance	80000	88000	88000	256000
Totals	323500	355850	355850	1035200

2.3 Manpower

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

3 Staged Rightsizing

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 Current Status (November 1993)

At present two of the Alpha workstations, and the VAXstation, are installed in the VXCERN 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. 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. 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 Alphas 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 (we see no significant difference between the Alpha with 64 MBytes of memory and the Alpha with 128 MBytes

memory, up to a limit of 70 interactive users.) 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%.

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

Machine	10	20	30	40	50	60	70
Alpha	55	54	59	54	53	36	52
VAX 9000		83	111	119	139		

We thus expect good response in the right-sized VXCERN cluster for a total of 4*70 = 320 interactive users, compared with very poor response on the VAX 9000 with 250 simultaneous users today. 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.

We are expecting to be able to remove all the TA90 cartridge drives from the Cluster. 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 VXCERN. It seems likely, however, that we will have to purchase an STK desktop 3480 drive for those users who, for some reason, simply have to mount tapes locally on a cluster machine. 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.

5 Appendix A - The VXCERN Service

In April 1985, a DEC VAX 8600 machine was put into service in the Computer Centre for use by physicists from the four LEP experiments. This machine, rated at around one CERN Unit, quickly became a popular platform for code development, as a source for the latest VMS versions of the Program Library, and as a springboard for network connections worldwide, amongst other features. Since then, due to constantly increasing demands, the VXCERN service has been upgraded several times, and until recently comprised a VAX 9000 model 410 (rated at ten CERN Units), and an ageing VAX 8250 (rated at a fraction of a CERN Unit, and used as a gateway to the Cray). Nevertheless, increases in the CPU capacity of the service have in general only provided a brief respite before renewed saturation.

5.1 Load Balancing

Part of the VMS clustering software allows load balancing between the participating nodes in the cluster: new users logging in to the cluster alias are connected to the machine that is least loaded at that time. At the time when VXCERN consisted of the VXCRNA(VAX 8650) and VXCRNB(VAX 8800) machines, this load balancing was found to operate very successfully. In addition, it is possible to devise schemes which allow a user to:

- log in to the least loaded node in the Cluster, or
- log in to the least loaded VAX node in the Cluster, or
- log in to the least loaded Alpha node in the Cluster, or
- log in to a specific node in the Cluster,

depending on the user's requirements.