The HLRN System

Peter Endebrock
HLRN / RRZN

The HLRN (Hochleistungsrechner Nord, High Performance Computer North) is a joint effort of six North-German states. The HLRN has two installation sites, one in Hannover at the RRZN, the other in Berlin at the ZIB, but for the user it is intended to look like a single system. Each site is presently equipped with 13 (in the future 16) IBM p690 series 32–way SMP frames, and the two sites are coupled via Gigabit Ethernet. The HLRN offers high-performance computing service and support to the North-German universities and other scientific institutions. A "network of competence" with consultants from the participating institutions provides the support for the users.

The six North German states Berlin, Bremen, Hamburg, Mecklenburg-Vorpommern, Niedersachsen and Schleswig-Holstein decided in the late 1990s to start a joint venture to install a high-performance computer system. The reason was that none of the states was able to finance such a system by itself, but that combining efforts might make it possible to get a high-performance computer system for Northern Germany. That was supported by positive experience from previous cooperation of Berlin, Niedersachsen and Schleswig-Holstein in the NVV, the North German Vector Computer Cooperative.

The idea was (with some hesitation and after a lot of negotiation) approved by the German federal funding and review organizations, and the system was installed in the second quarter of 2002. The concept is that this should be a "single supercomputer at two sites". Identical systems (except for the file archiving system) have been installed in Hannover and Berlin, and for the user they are intended to look like a single system.

The present hardware equipment at each installation site consists of thirteen 32-way IBM p690 frames with Power4 processors connected by the SP2 ("Colony") switch. Three more of these frames will be installed by mid-2004 together with the "next generation" ("Federation") switch. In addition there are eleven p655 servers for I/O and archiving on each site. You can find details of the present configuration of one site in Figure 1.

The summarized data of the two sites are:

- 26 x 32 CPUs with 1.3 GHz, equivalent to a theoretical peak performance of c.4.4 Gflops
- Main memory 22 x 64 MB, 2 x 128 MB and 2 x 256 MB, together c.2.2 TB
- Total mass memory c.52 TB
- Total archive storage c.3 PB

The two sites are coupled by a dedicated fibre network with a bandwidth of 2.4 Gbit/s.

The user view of a "single system image" is still under development, but a lot of features in that direction have already been realized. The system and application software at both complexes are identical, except for minor differences due to licensing restrictions. Each user has a single user name and password at both sites, and has an identical home directory available. The directories for permanent and medium-term files are presently only available directly at the user's primary site, and they have to make file transfers to the secondary site, if necessary, on their own. One batch system (IBM LoadLeveler) manages jobs for both sites, with the possibility of sending jobs to a pre-selected site, or leaving that decision to the system. Actually, as mentioned before, the user has to arrange for some of their files to be available where the job executes, but we hope to improve that to a certain degree. Currently, only test jobs have been allowed to use CPUs from the two sites in one job, but it has been shown that that is possible in principle. As you can imagine, due to latency and transfer capacity, this will not be a sensible method of system usage in general, but it may prove to help with the solution of large loosely-coupled systems.

A challenge for the two installation sites was the cooperation of two computer centres each with a long history of their own, to develop into a single centre with a common policy. A consolidation phase was necessary because of different "cultures", but most of that is history now. Regular video conferences support the cooperation, and the combined manpower of the two sites was and is needed to support the system – one site wouldn't have had enough people to do it by itself.

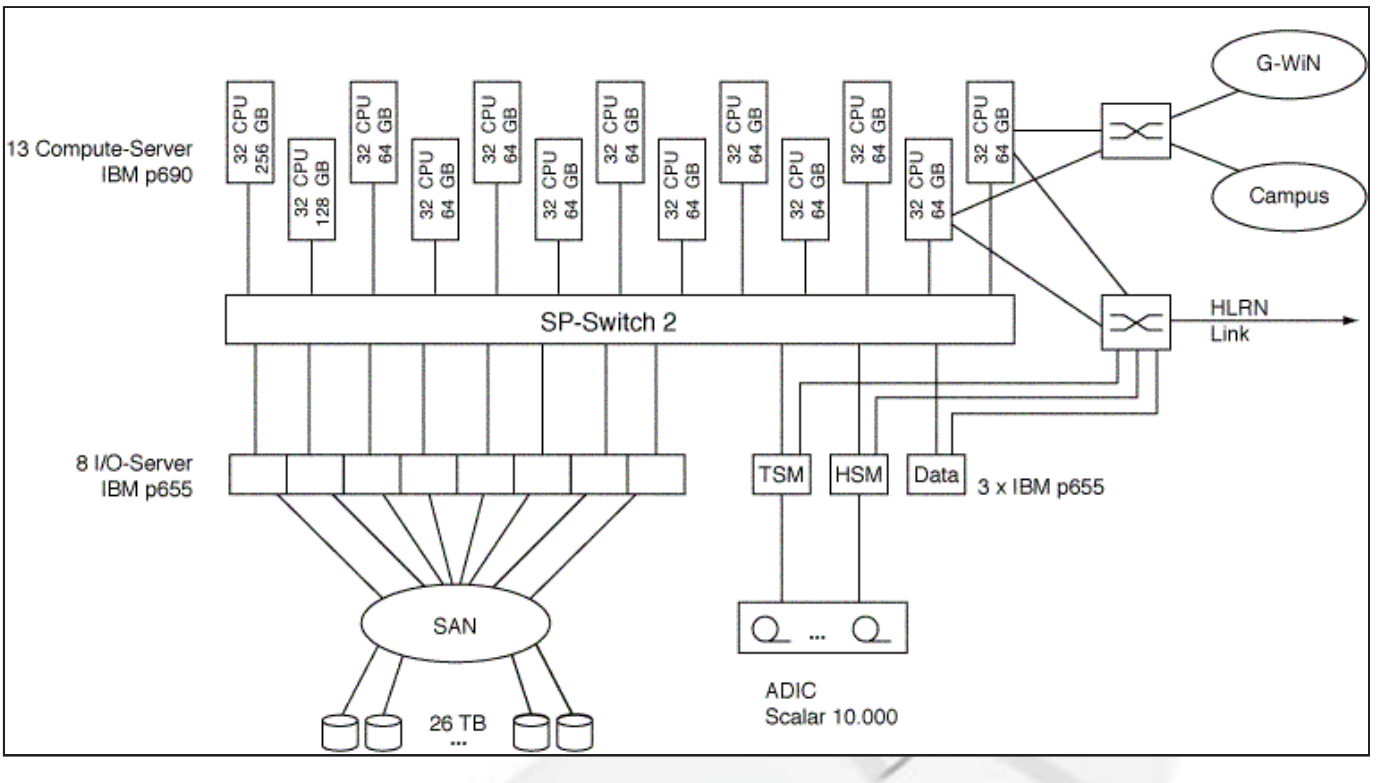

Figure 1: One half of the HLRN system (the one at Hannover; the archive system at Berlin is different)

The main application areas are environmental research, climate and ocean modelling, and various engineering applications like ship building or computational fluid dynamics, as well as fundamental research in physics, chemistry and life sciences. Many users have their own application codes, but standard application packages in different areas are also offered and used.

The initial user name application is relatively informal and does not require detailed argumentation. You get a small resource allowance to prepare your project application. The projects applying for using the system are reviewed by a Scientific Board that allocates resources on a quarterly basis, for a maximum of four quarters in advance. For the first application period, starting with the second quarter 2003, the total applications amounted to three times the available resources, an indication of how badly the HLRN system had been needed.

The system shows presently a usage of more than 60 percent of the available resources, without calculating downtimes. We consider that to be good for a large parallel processing system with a mixed workload and a preference for large projects.

User support is organized by a network of consultants distributed over the participating states and their local computer centres. Each project is assigned an individual

Figure 2: The installation phase of HLRN at Hannover

consultant, preferably with experience in the field of the project and geographically close to it. If that is not possible, teams of two consultants are assigned to a project, the "local advisor", and a remote "specialized" consultant.

In our opinion, the concept of cooperation between several federal states to install and run a "distributed supercomputer" is proving its feasibility and its advantages, and because of the long planning cycles we are already starting to prepare a request for a similar successor system.

If you want, you can find more details about HLRN at http://www.hlrn.de/.