



# Beyond 50 thousand billion



© CEA

## Jean Gonnord

Head of the digital and IT simulation project, CEA/DAM.

## Pierre Leca,

Head department of Simulation and Information Sciences, CEA-DAM.

## François Robin

Deputy head of Simulation and Information Sciences department, CEA-DAM, and operational manager responsible for the TERA-10 project.

Performing more than 50 thousand billion operations a second, TERA-10 will be the most powerful computer in Europe. Profile of this very new supercomputer, designed and build by Bull for the Military Application Directorate (DAM) of the French Atomic Agency (CEA)

Since the French parliament ratified the international Comprehensive Nuclear Test Ban Treaty (CTBT) in 1998, the country's nuclear weapons are no longer tested. Nevertheless, these weapons continue to evolve... To ensure that they remain both effective and secure, computer simulations replaced the physical tests. Predictive models developed by researchers in the Military Applications Directorate (DAM) at the French Atomic Energy Agency (the CEA) are based on non-linear equa-

tions, which are solved by iterative calculations, with the correctness of the resulting predictions relying heavily on the size of the mesh. In other words, this kind of modeling requires very powerful computers. The need for processing power, also driven by improvements to the models being used, is growing at breathtaking speed: in essence, by a factor of ten every four years. This sustained pace has resulted in a new supercomputer being installed every four years. Tera-I, provided by Hewlett-Packard at the end

of 2001 gave way at the end of 2005 to Tera-10. This machine, built by Bull, should rank amongst the most powerful supercomputers in the world's Top 500.

Like its predecessor, Tera-10 features an SMP\* cluster architecture. Processors (the basic elements that carry our computation) may belong to one of two categories: scalar and vector. The scalar units carry out simple operations, generally on numbers which can be simply defined by their measured value: adding two numbers together, for example.



# operations/second !



**THE AUTHORS**  
In front of a line  
of cabinets that  
make up the Tera  
computer. © CEA

The vector processors carry out operations on objects that can be defined using a combination of several mathematical entities, such as the addition of two vectors each comprising 500 different elements. Vector processors are especially well suited to regular calculations, found frequently in computer simulation tasks: during the execution of an operation of this type, a vector processor can function at a speed close to its maximum performance (or 'peak\*'). By contrast, the same operation executed on a scalar pro-

cessor requires many independent operations (vector element by vector element), carried out at well below peak performance. Three types of architectures enable processors to be used in parallel: vector supercomputers, 'clusters' of shared memory scalar processors, and PC clusters (the computers we all have in our homes). Although it is inexpensive, this third kind of architecture is not suitable for environments where many different users carry out calculations that are very 'greedy' in terms of

the processor power and memory they require. Vector supercomputers are very expensive but prove more powerful (in scientific computation, they usually achieve more than 50% of peak performance). Clusters of scalar processors, which are at the heart of most mass-market computers, are certainly less powerful (delivering between 5% and 20% of peak performance), but are also significantly less costly. And this means that they offer by far the best price/performance ratio for the CEA/DAM's applications.

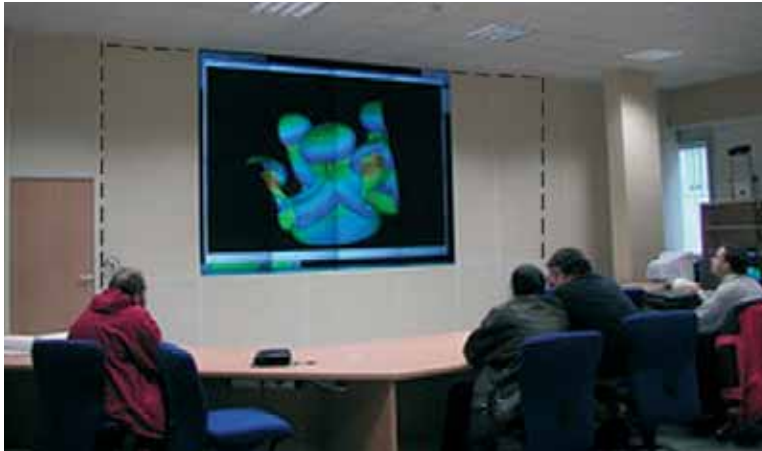
\*SMP, shared memory multiprocessor.  
\*Peak performance, optimum performance that a computer can achieve, albeit very briefly.





## SUPERCOMPUTER TERA-10

WITH TERA-10, a 5.5 x 3 meter wall display will replace the current screen (dotted line). © CEA



→ That is why, beginning with Tera-1, we chose to move away from the traditional scientific computing hardware we had been using up to that point (Cray vector computers) and opted for an SMP cluster architecture.

The basic computing unit consists of a group of sixteen processors sharing the same memory. These units (known as 'nodes') are assembled in parallel, each one carrying out part of the calculation. The nodes are interconnected and share common disk space. Some of them (I/O controllers) are dedicated to controlling the disk space, and this enables the computing nodes to access the disk space to record the results of their calculations or read the initial conditions needed to perform subsequent calculations.

The CEA's experiences gained while using Tera-1 enable the strengths and limitations of these types of supercomputers to be identified.

The first question was to discover whether this type of architecture could actually deliver the necessary computing power. This proved to be the case, on condition that it had access to applications that had been properly configured for parallel processing.

When an application runs on several hundreds or even thousands of processors, it has to effectively be broken down into as many small-

ler computational units as there are processors. These units are in regular communication with each other, so that the boarder results are cross-checked, which does consume processing time. This means this exchange has to be optimized as far as possible. In addition, the application is fragmented in such a way that every calculation on every section of the program runs for an almost identical length of time.

No one part should be effectively putting the brakes on the entire calculation. When applications are written to meet these parallel processing criteria, the performance of SMP clusters really comes into its own. Its overall performance also depends on the supercomputer's loading capabilities.

So the job schedules must be extremely carefully managed to ensure that the maximum number of processors are continually in use. This involves both small and major jobs being scheduled to start at the same time.

The use of an optimized queuing system, taking into account the relative priorities of different calculations, enabled Tera-1 to achieve productivity records beyond 80%. This is relatively unusual for these kinds of systems, where the norm is usually less than or equal to around 70%.

The chosen architecture for the Tera computers nevertheless has

some limitations. The biggest drawback is undoubtedly the I/O load due to the shared disk storage of the SMP cluster. The second limitation of these very large-scale systems is the high number of electrical components they contain. Even though, individually, they may be very reliable, none is immune from a failure. With Tera-1, we had around 20 of these a month. Most of these only resulted in a delay of a few hours in completing particular calculations (which often take many hundreds of hours to run). Bigger breakdowns – resulting in actual machine downtime – are thankfully fairly rare: typically less than once a quarter.

Under these conditions, it was essential to establish fault tolerance strategies at every level in the system. So, on Tera-1, critical hardware elements essential to the operation of the whole system (the machine interconnectivity network, part of the shared disk space...) were mirrored, while others were not, for reasons of cost. In case of a breakdown, the software can be called on. For example, if a particular processing node goes down, the calculation it was performing is lost. However, it is then relaunched from a 'restarting point', recorded at regular intervals by the applications running on the computer. Obviously, this risk of breakdown means that major maintenance and surveillance systems have to be in place. Some 15 people are employed in system surveillance at the CEA, with maintenance services being provided by the hardware manufacturer.

For Tera-10, the SMP cluster architecture tested on Tera-1 has been retained. This makes it much easier to transfer software from one platform to the other, however, it still has to be optimized for the new system, which is a heavy undertaking. To respond to current and future developments in production code

**\*Sustained performance:** average performance that a computer achieves on an actual application..



and limit the number of nodes in the cluster, a machine consisting of extremely powerful nodes was chosen (100 gigaflops/node), offering ten times greater levels of sustained performance\* than Tera-1. So, although the basic architecture remains the same, the number and nature of the components changes.

The 544 'computing nodes' that make up Tera-10 are each equipped with 16 processors, making a total of some 9,000 processors when the dedicated I/O nodes are taken into account. And if Tera-10's processing performance is ten times better than Tera-1, the disk space for its part has been multiplied by 20, to meet the extreme demands in terms of the quantity of results produced by the simulation software.

When it comes to the 'system' software, which manages the operation of the actual machine, open software was chosen, including Linux for the server operating system and Lustre for the global parallel file system. Bull is providing additional specific elements for both these software components in

order to optimize the calculations. Moving from Tera-1 to Tera-10 also involves many adaptations to the IT infrastructure. To ensure that the users get the most from the new power on offer, the supercomputer has to be networked and linked to data storage systems, interfacing and data visualization software, etc. This will enable Tera-10 to be connected to a 5.5 x 3 m image display wall, allowing groups of users to view the results of simulations together.

Currently, Tera-1 is processing between three and five terabytes of data a day. With Tera-10 the volumes will increase in almost direct proportion to the increase in processing power, necessitating the implementation – during 2005 and 2006 – of much more powerful network and data storage equipment. Internal networks have been consolidated and data storage systems increased. There are two levels of storage: data that is more than a year old, which is less likely to be required again in the future, is kept on a magnetic tape system (the most reliable and cost-effective

way of storing large volumes of data), stored in tape silos.

The most important data is duplicated and placed into secure zones. For short-term data storage, the magnetic tapes will be replaced by magnetic hard disks, which have come down enormously in price, driven by the home computing market. In this context, four petabytes (four million gigabytes) of disk space will be installed during 2006, for first level data storage.

A large new building (featuring 2,000m<sup>2</sup> of machine rooms) was specially built for Tera-1. It was designed from the outset to accommodate several generations of simulation hardware (computers and storage), including Tera-100, which is planned for 2009.

Nevertheless, each new version requires some adaptations. So, to move from Tera-1 which consumes around 0.6 MW of power to Tera-10 which requires around three times more than this, it has been necessary to install a new 2MW air cooler, three inverters (which provide battery-powered AC current in the case of electrical power cuts), not forgetting an extension of the fire detection/extinguisher network and the machine room cabling! Once the simulation code used for the design and verification of Tera-10 has been validated, Tera-1 will be switched off and dismantled. This is due to take place in mid 2006. The space freed up as a result will accommodate Tera-100 in 2009.

With Tera-10, the physics experts and designers at CEA/DAM will have access to ten times the simulation capacity provided by Tera-1. But the installation, and then the six-month optimization period required for a system of this scale, remains a challenge. Because here, as with any exceptional piece of equipment, both the technologies and human experience are being pushed to the outer limits. ■

**J. G., P. L. and F. R.**

**TERA-10 consuming three times the power of its predecessor, a new 2MW air cooler has been installed.** © CEA

