

## Cluster computing: CPU count isn't the whole story

JOHN WEIGANT, Geotrace Technologies, Houston, Texas, U.S.

**“We** have 1024 CPUs in our cluster.”

“This imaging job will be run on our 2000 CPU Linux cluster.”

“We have over 3000 CPUs dedicated to seismic processing.”

How many times have oil company representatives heard lines like these? How many times have those of us in the seismic processing industry delivered lines like these? Most importantly, does it really matter how many CPUs you have?

With the advent of low-cost Linux clusters, seismic processing options such as large-scale prestack depth migration and wave-equation migration became feasible on a commercial level. In turn, in the never-ending cycle of one-upsmanship that goes on among contractors, the number of CPUs in a company's Linux cluster became a new metric. But is this really the right measure of “processing power” or throughput capabilities?

For the sake of this particular argument, let's ignore some of the more unscrupulous ways that CPU count has been abused in the past to bolster the appearance of capacity: Counting every PC and laptop CPU, adding this number to the number of CPUs in the 10-15 year old machines unplugged in the corner and combining this sum with the actual 256 usable Linux processors available and making a statement like, “We have over 1000 CPUs in our facility.” Let's just talk about real cluster capacity and how it should be measured.

One of the truly relevant numbers at which everyone is trying to arrive when they talk about number and clock speed of CPUs is floating point performance. This metric is measured in trillions of floating point operations per second, or teraflops. There is a theoretical maximum number for a cluster that can be calculated by multiplying a processor's cycles per second by the number of floating point operations performed in a cycle by the number of processors. This gives you what is referred to as the  $R_{\text{peak}}$  value on the Top500 Supercomputer list, a Web site for tracking and detecting trends in high-performance computing. This list, updated twice annually, can be found at [www.top500.org](http://www.top500.org).

However, the real measure of your cluster's capability is its maximum achievable number of teraflops, or  $R_{\text{max}}$  on the Top500 list. This number is not only measuring how many CPUs are in your cluster, but how well they communicate with each other. The ratios between  $R_{\text{peak}}$  and  $R_{\text{max}}$  for clusters near the top of the Top500 vary from 0.45 to 0.88. Some installations much lower in the list achieve better ratios while some achieve much worse. Another way to look at this is as a measure of a system's efficiency. Is the cluster CPU bound, or is it waiting on the network that is connecting the CPUs?

The application used to rank clusters for the Top500 list is not a seismic application. Some seismic applications, by their very nature, are “embarrassingly parallel” and the node-to-node communication requirements are minimal. There is, however, still the need to read and write data. This is another potential pitfall for cluster efficiency. If a com-

puter's storage and networking cannot keep up with an application's I/O demands, the system will not remain CPU bound and will therefore run at less than optimal efficiency.

Shot-domain wave-equation depth migration is an example of an application that is very CPU intensive with minimal I/O and communications requirements. A more I/O-intensive application is a Kirchhoff PSTM. As the available Linux processors become faster, the challenges of preventing networking and disk I/O from becoming bottlenecks to efficiency become greater.

So far, we have touched on hardware configurations and applications. The third leg in the tripod supporting successful cluster computing for seismic applications is operations. Processing systems that have been ported to run on large clusters are often inefficient to use when it comes to job submission and monitoring. This is usually because they have always been somewhat inefficient, but now the problem has been greatly multiplied by the fact that systems are now composed of thousands of processors. It is not unusual for data processors to have to set up, submit, and monitor hundreds, or even thousands, of jobs to complete a PSTM or PSDM on a large marine survey.

So, the next time that someone is bragging about the number of CPUs they have, ask the following questions:

How many teraflops is that? How are your nodes connected? What kind of storage are you using? How efficiently does your cluster run? How do you submit and monitor your jobs? Can I watch the job submission process? Can I watch the system monitor while my jobs are running? Can I meet your IT people?

These are the real capacity indicators. **TJE**

Corresponding author: [jweigant@geotrace.com](mailto:jweigant@geotrace.com)