Sunday, May 3. 2009

Just a short comment about Cluster computing

I have a reader, who is really convinced that clusters are the solution for all problems of the world and that RAS is futile. In this arcticle i will talk about the "cluster for all problems". I just want to make a short comment here about such comments.

There is an important law in computing: Amdahls law mandates that the maximum performance of a highly parallel system is determined by the speed of the slowest component.

Now, just as a short information. The MPI latency of Myrinet is 2,6 microseconds, the MPI latency of Infiniband is 1.07 microseconds. 10GB Ethernet? In the range of 50 microseconds (although there are developments to reduce this). On the other side inside a really large non-clustered system: The memory latency of a Sun SPARC Enterprise M9000/64 is 437 nanoseconds to 532 nanoseconds.

And there we are in viewing distance of the problem why the idea of this reader that cluster is the solution of all problems is nonsense. As soon as you have to use the interconnect the process of transmitting the data becomes the slowest part in the equation. 1.07 microseconds are 1070 nanoseconds. At 1 GHz the length of one cycle is one nano seconds. At 3 GHz a third of a nanosecond. So you see ... in the latency of the interconnect you could do a lot of work. It´s really easy to see, that the connecting network is the slowest part in the equation.

When you look at the workloads for clusters, they are always corner cases. Rendering ... almost no communication between the nodes, the interconnect isn´t really a part of the processing. Many HPC tasks are similiar in nature or have at least a moderate amount of communication. Map-Reduce ... low amount of transported data as it´s a shared nothing distributed data architecture. But you can look at the problem like you want ... such architectures are cornercases. You don´t use Map/Reduce for SAP or Oracle Financials, only a few people render films.

I know that there are just clusters in the HPC Top 500 list. But that doesn´t mean that the people just use clusters for HPC. The point it: Linpack (the benchmark behind the list) is cluster-friendly. In my opinion it´s even too friendly to special purpose hardware with small memory footprint, but that´s a different discussion). Many problems in science aren´t problems you can compute on such a cluster. That´s the reason why there are a number large SMP/ccNUMA systems in HPC. They aren´t large enough to make it into the list, but on their workload the compute at least faster than most systems on the Top500 list.

Webserver Clusters are a great shared nothing architecture. Fileserver may be a great shared nothing architecture/no interconnect load as well with the advent of pNFS. But at the end ... corner cases.

It´s the same with Oracle RAC. Oracle RAC runs reasonably well on data sets where you can seperate the data access patterns (e.g. ZIP 1-10000 on first node, ZIP 10001 to 20000 and so on) but at a truly random pattern on the database it sucks, as it has to transport the cache quite often over the interconnect. By the way, is there nobody else out there who thinks it´s strange not to find a single Oracle result in the TPC-E benchmark ?

It´s not that Sun is the only computer company who thinks this way: IBM has its pSeries, HP has the Superdome ...

And there is my point: Blinded by the success of clusters in some areas some people (and this reader of my blog is one of them) believe that they will be an solution for all the inherent problems of cluster computing and cluster systems are the solution for all problems. .That´s simply not true. The world isn´t that simple like some people want to suggest. IT doesn´t consist only out webservers, rendering and Google-like computing. Everyone thinking otherwise should broaden it´s view on IT over the borders of his own problem sets.

Posted by Joerg Moellenkamp in English, The IT Business at 16:11

Latency is distance. 1 Ghz at lightspeed are 30cm, 3 Ghz at lightspeed are 10cm.

Consistency requires 2PC or a slower protocol. 2PC requires two round trips in a successful transaction. The size of a consistent

system at 3GHz is therefore limited to a cube of one inch in size. Make it larger, and lose either consistency or speed.

That is why distributed systems are hard (Ye cannah' change the laws of physics, Captn! (Unless you are a Q)).
Anonymous on May 3 2009, 16:31