The Lattice QCD community needs both computing power and data sharing: the APE behind the GRID

A. Lonardo, A. Michelotti, D. Rossetti
I.N.F.N – sez. Roma I – P.le A.Moro, 3
00185 - Rome - Italy
alessandro.lonardo@roma1.infn.it
andrea.michelotti@roma1.infn.it
davide.rossetti@roma1.infn.it

Abstract

The introduction of new computing paradigms, namely Grid and Cluster computing, is exerting unusual pressure on traditional supercomputing environments to reach a certain level of integration. Grids are thought to become the infrastructure, which matches the processing power needs of near future High Energy Physics (HEP) experiments, like the CERN Large Hadron Collider (LHC). In fact, some of the new technologies developed within Grid initiatives are also appealing as pretty good solutions to some long-standing needs of the Lattice Quantum Chromo-Dynamics community. We report on the planned grid activities of the APE group.

1. Introduction

Historically the early ’80s saw a great deal of activity in the field of numerical simulations of discretized Lattice Quantum Chromo-Dynamics (LQCD). The computing power necessary to obtain even acceptable results in the so-called *Quenched* approximation to LQCD, where a simplified dynamics is considered − as opposed to the complete or Full LQCD − was beyond that available at that time. Somehow, the viability of this kind of computations was also linked to the realization of prototypes of custom super-computers [1, 2].

During the ‘90s, more ambitious efforts [3-8] gave unexpectedly good results. The physics program focused on increasing statistics in Quenched LQCD simulations and on first attempts at implementing Full LQCD algorithms.

As the technology provides for increasing computing power, the more refined Full LQCD model might be simulated with enough statistics to get relevant results. Anyway, going to Full QCD simulations with reasonable lattice sizes requires order of magnitudes more power than the approximated quenched model, to the point that the − stochastically generated − configurations of the gauge fields become a real treasure. This improved theory is necessary to study the low energy hadronic spectra. At the same time, we expect that very large lattice simulations will be carried out in the future with the quenched theory in order to study the physics of the weak interaction of heavy hadrons.

In a typical scenario, order-of-a-day time is necessary to produce both a de-correlated Full LQCD configuration and a Quenched LQCD propagator set. They have to be carefully stored as to be “measured” later on, calculating different physical observables. Given the large lattice sizes, this latter approach is going to be computationally cheaper than producing them on the fly, as it was done in the past for the propagators. Some sort of distributed storage facility is needed as gauge field configurations and propagators have to be shared among researchers. Furthermore a simple yet efficient security facility should be implemented to limit and/or control access to the data sets.

A similar yet more primitive approach has already been explored in the past by the SESAM and TχL collaboration [9], using various size APE100 machines and one Cray T3E, both in Italy and in Germany.

It is interesting that the size of the raw data sets, of the order of hundreds of Gigabytes per simulation, is comparable to those of HEP experiments. Somehow the data *throughput* is comparable. That is why we originally turned to evaluating DataGRID.

In the next two sections, we expose some details of the INFN apeNEXT experiment and of the CERN grid initiative, DataGRID. The fourth section is devoted to the proposed apeNEXT-DataGRID testbed, followed by some conclusion remarks.

2. The apeNEXT experiment

Custom super-computing viability is related to the peculiar characteristics of LQCD algorithms: complex numbers algebra, locality – as it is a first-neighbor interaction – and large data sets – as the lattice is 4D. –

The latter means that LQCD uses memory cache unfriendly algorithms, which hurt cache-based CPU architectures, typical in commodity HW, to the point that only the memory-CPU bandwidth is the key factor. This bandwidth is usually small compared to the CPU ability to crunch data and in fact mitigated by cache hierarchies – L1, L2 and so on. – Custom made FPU's, lacking the leading-edge silicon technology, may resort to accelerating in silicon the very critical algebra and balancing the architecture to match the LQCD critical numerical needs.

The APE family of super-computers is an example of the custom computing efforts that originated in the 80's. Its third generation and latest representative is APEmille [10, 11, 12], which sports up to 2048 processors, for 1 Tflops of peak SIMD performance and 64 Gbytes of memory. Today the project is considered finished with two large 128 Gflops installations in production for six months and a lot of HW is being manufactured and deployed just now – i.e. two new 64 Gflops systems are up and running. –

The APE group is the actual entity behind all the aspects of both the design and implementation of the APE super-computers. Within this very group both software --- compiler and operating system technologies --- and hardware --- VLSI and complex PCB design --- expertise is present. Moreover, the group is moving from the Italian-German collaboration, which is responsible for the APEmille supercomputer [10], to a wider European collaboration [13,14] as suggested by the recent ECFA documents [15]. The enlarged group is pretty well in the design stage of the apeNEXT supercomputer, the next APE machine, which is targeted toward the design goal of several Teraflops of computing power installed throughout the LQCD European community within 2005.

Last but not least, in the past the APE group has been one of the INFN origins of high-level technology transfer to the industry. We hope to follow on this task at a new, unprecedented European level, in a role somewhat similar, we hope, to the CERN one; just think of CERN big contribute to the IT industry for WEB technology.

3. The CERN DataGRID Project

Since the emergence of the Grid computing paradigm, the HEP community has plans to employ it for its purposes. In fact, the needs of tomorrow large HEP experiments at CERN [16] are: storage of huge data sets, computing power to process them and their sharing among distributed research communities. The availability of a first generation of technologies, such as the Globus Toolkit [17], is a great opportunity to start a follow-up activity to further extend them according to HEP needs. Within this framework the CERN DataGRID [18] project shines for being a rather comprehensive implementation of the Grid concept, from the software middleware up to the hardware infrastructure and testbed applications.

CERN DataGRID work packages

Middleware	
1	Grid Work Scheduling
2	Grid Data Management
3	Grid Monitoring Services
4	Fabric Management
5	Mass Storage Management
Infrastructure	
6	Testbed and Demonstrators
7	Network Services
Applications	
8	HEP Applications
9	Earth Observation Applications
10	Biology Applications
11	Dissemination
12	Project Management

DataGRID is an international effort, backed partly by the EU partly by the CERN partners. The huge estimated effort is partitioned into 12 Work Packages (WP), which are listed in the table above.

Among the contributors, INFN is fully committed in many aspects of the middleware software development and in the deployment of by-products through a special national experiment, INFN-Grid [19]. It is driven in parallel with CERN DataGRID with the aim to coordinate the INFN effort in it and to build up the Italian computing infrastructure for the future LHC experiments. In particular, INFN is responsible for WP1.

4. DataGRID and apeNEXT integration

Today there is a common agreement about the physics program of the European Lattice QCD community for the next years [14,15]. In the following table we try to lay down some estimated figures:

	APEmille	apeNEXT
Time frame	2000-2003	2003-2006
Lattice size	50^3 x 100	100^3 x 200
Propagator size	500 Gbytes	10 Tbyte
Number of configurations	50	100

Total data set size	25 Tbyte	1 Pbyte

Note that the “*number of configurations*” figures are very conservative. They have to be considered just as hints.

The involvement of the APE group in INFN-Grid activities focuses on the Data Management work package (WP2), for which INFN is not directly responsible. Its deliverable is a middleware for data management, which addresses the topics of fast data transmission, replication, synchronization and security.

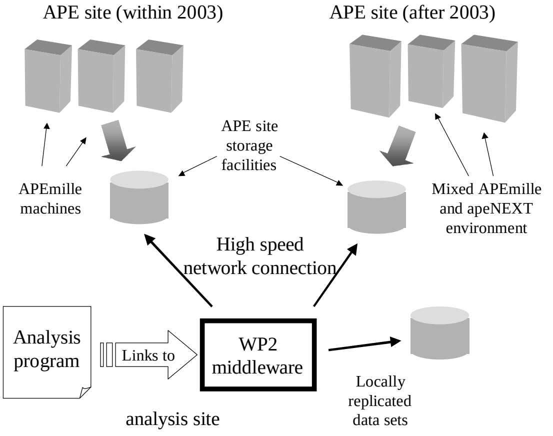
The suitable place to deploy this technology is the foreseen European LQCD collaboration. Indeed all of the main apeNEXT sites (ANS) will be equipped with:

- A pool of apeNEXT super-computers.
- A local multi-Terabytes storage facility for Full QCD configuration storage.
- A high-speed network connection to allow for configuration sharing.
- Optionally, a PC cluster to carry on measures.

Our technical point is that DataGRID data management technology provides the necessary infrastructure to transparently link all the ANS’s. Software programs to analyze the Full QCD configurations might use the WP2 middleware to read the configurations wherever they resides.

4.1. The architecture of a prototype apeNEXT testbed

As of now, the APE group has been officially involved in the design of an INFN-Grid testbed [19]. The figure below depicts a reasonable scenario for us. We plan to choose two large APEmille sites of today, which will be upgraded to apeNEXT machines as soon as available, and to link them with a fast WAN connection. We will let them produce and permanently store a reasonable amount of configurations. Then we will provide the necessary software tools to let a pool of physicists write down and test their analysis programs. For computational power demanding analysis we could provide a medium-sized PC cluster.



The configurations produced in each APE site will be stored at high bandwidth in the local storage facility. We plan to customize the WP2 software and wrap it in easy-to-use library routines. These routines should make easy to uniformly access local as well as remote files, affectively masking their proper location.

Files belonging to the same configuration – gauge field raw binary data, physical parameters and machine topology used in the simulation – will be archived in a unique data set identified by a tag containing timestamp, ANS ownership information and user specified data.

Providing this tag to the wrapper library, it will be possible to:

- Publish/Un-publish a configuration data set in the (unique) metadata catalog server. This will add/remove only the tag data set to the server and not the configuration itself. The data set will be digitally signed with the certificate of the ANS.
- Search for a configuration data set (local or remote) by metadata.
- Retrieve a configuration data set. This will be always possible for local configuration data set but a DataGRID security authentication phase will be necessary to download a remote configuration data set. Automatic replication could be used alternatively.

Generally speaking, our testbed has to support a number of operations on configuration sets:

- Production (on APEmille at first, then on apeNEXT)
- Storing on high-speed disk storage.
- Archiving, that is, automatic data migration to low-speed, inexpensive storage.
- Recovery of configurations from the archived to the stored state.
- Configuration analysis.
- Security is a moderate but present topic.
- Replication of configuration sets between the two ANS.

4.2. The APE group activities

The APE group activities are mainly focused onto WP2. As of now, we are well in the requirements analysis and production stage. We are actively collaborating with the people from the LHC experiments (CMS, Atlas, LHC-b, Alice) as well as from other non-HEP initiatives (Virgo, ESRIN) to find the best match for our needs. The main focus is on the data model of the different applications. As of now, HEP experiments applications seem to expose an extremely rich requirement list.

Complementarily such an activity is opening a good opportunity for us to deeply analyze some issues of the

LQCD, which were traditionally dealt with manually. For example, each LQCD group has its own policy for configuration sets maintenance. In the past, just using simple file and/or directory naming was enough. Even configuration back-up's were mostly driven by personal will. We think that in the enlarged LQCD community, it is necessary to turn to more sophisticated technologies, which enforce better policies for free.

Technically, we count on WP5 activities to have inputs as of the suggested SW/HW architectures [20-22] for the ANS storage facilities. In fact, while above we referred to ANS permanent storage, we foresee to ship most of the apeNEXT machines with further dedicated high performance, parallel I/O sub-system --- just like APE100 and APEmille. --- As we got to know by our long experience, large-scale simulation of the simplified, *quenched* theory needs a very large and fast temporary storage for swapping. These kinds of simulations are most interesting for the physics of the strong corrections to weak interaction processes --- weak decays. --- As a gross measure, an I/O bandwidth of 0.5-1 Mbytes/s per Gflops of processing power is considered necessary for the performance cost of swapping being negligible. In this area the APE group has developed a considerable experience in the last 10 years and we expect to be able to contribute back a lot to the DataGRID community.

By the same argument as above, it is expected that WP4 inter-networking choices are readily reusable to interconnect apeNEXT sites. Even in this field we expect the APE collaboration to contribute some of its high-speed, LVD link technology. It is well in our project to back-port the technology and implement it in a PC network board. This effort may be important for grids local interconnection.

Of course, all that translates to the necessity to properly *link* DataGRID technology to our future apeNEXT Operating System environment (NOS). In fact, NOS will be certainly designed with DataGRID middleware use in mind. In that sense, the NOS may be considered a client application of DataGRID middleware just as LHC experiment distributed data analysis programs.

Furthermore, we expect that in some apeNEXT sites PC clusters will be deployed and used as an alternative analysis engine. Even in this case, analysis programs will be linked against DataGRID middleware libraries mainly for data retrieval purposes.

4.3. Further developments

As a more ambitious project, we envision the possibility to transparently integrate apeNEXT machines in the future DataGRID job scheduling environment, just like a pretty standard computing facility. This way, apeNEXT would become a *grid computing resource*. In the end, it might be integrated in a EU-wise LQCD computing grid.

In this case, we need further stuff from DataGRID, not only WP2 middleware. Realistically, most of DataGRID middleware might be useful, especially Mass Storage Management and Fabric Management work packages.

5. Conclusions

We can conceptually draw a parallel between gauge field configurations production in LQCD numerical simulations and events production in particle physics accelerator experiments.

That is why we envision moving the technical solutions born in the latter environment to the realm of the former; it is somehow a case of solution-reuse. This effort may benefit both realms as the apeNEXT project has a shorter deadline and can be really be one of the first consumers of the DataGRID middleware.

Apparently, the main difference is that in the LQCD case, there are many experimental sites, which will be the apeNEXT super-computers installments; instead, for example, the LHC experimental site will be just one, that is, the CERN. In fact, the LHC case is more complex. The Raw experimental data will be collected in the CERN IT facility. These first data sets have to be processed in many stages to produce higher-level data [16]. Some processing might be carried out in the so-called Regional Centers, that is, the national LHC computing facilities. These processed data sets can be considered just like LQCD configurations, and are needed to produce further measurements.

As of now, the INFN-Grid project is well in the approval stage but preliminary funding has already been granted. Meanwhile, the EU funding of CERN DataGRID, while guaranteed, is still under positive discussion.

References

[1] APE original project: M. Albanese et al., "The APE computer: an array processor optimized for lattice gauge theory simulations", *Comp. Phys. Comm. 45*, 1987, p.345.

[2] Y. Iwasaki, T. Hoshino, T. Shirakawa, Y. Oyanagi, T. Kawai, "QCD-PAX: A parallel computer for lattice QCD simulation, *Comp. Phys. Comm. 49*, 1988, p.449

[3] C. Battista et al., *Int. J. High Speed Comput.* 5 (1993) 637.

[4] I. Arsenin et al., in: Proceedings of the CHEP'97, Berlin, 1997, p. 586

[5] N. Christ, "Proceedings of Lattice '99", Nucl. Phys. B (Proc. Suppl.), to be published and hep-lat/9912009.

[6] R. D. Mawhinney, "The 1 Teraflops QCDSP computer", *Parallel Computing 25*, North-Holland, 1999, p. 1281-1296

[7] A. Ukawa, in: *Proceedings of the CHEP'97*, Berlin, 1997, p. 595

[8] A. Ukawa (for the CP-PACS Collaboration), "Lattice QCD results from the CP-PACS computer", *Parallel Computing 25*, North-Holland, 1999, p. 1257-1289

[9] S. Gusken et al., "Lattice QCD with two dynamical Wilson fermions on APE100 parallel systems", *Parallel Computing 25*, North-Holland, 1999, p. 1227-1242, and references therein.

[10] APEmille project proposal: A.Bartoloni et al., *Nucl. Phys. B* (Proc. Suppl.) 42 (1995) 17.

[11] A. Bartoloni et al., “Progress and Status of APEmille”, Nucl.Phys.Proc.Suppl. 63 (1998), p. 991-993.

[12] R. Tripiccione, "APEmille", *Parallel Computing 25*, North-Holland, 1999, p. 1297-1309.

[13] F. Aglietti et al., “Proposal for a Multi-Tflops Computing Project”, Rome Preprint 1255/99.

[14] apeNEXT project proposal: R. Alfieri et al., “apeNEXT: A Multi-Tflops LGT Computing Project”, to be published.

[15] F. Jegerlehner et al., “Requirements For High Performance Computing for Lattice QCD: Report of the ECFA Working Panel”, Preprint ECFA/99/200.

[16] MONARC: Models of Networked Analysis at Regional Centers for LHC Experiments, http://monarc.web.cern.ch/MONARC/.

[17] The Globus project, http://www.globus.org/.

[18] CERN DataGRID project, http://www.cern.ch/grid, in particular http://grid.web.cern.ch/grid/papers/HEP-Grid-outline-version_1.htm.

[19] INFN-Grid collaboration, “A Computational and Data Challenge for future INFN Experiments; A Grid Approach”, INFN draft proposal.

[20] Wolfgang Hoschek, Javier Jaen-Martinez, Asad Samar, Heinz Stockinger, Kurt Stockinger, “Data Management in an International Data Grid Project”, to appear in *IEEE/ACM International Workshopon Grid Computing* (Grid'2000), 17-20 Dec. 2000, Bangalore, India.

[21] HPSS: High Performance Storage System, http://hpcf.nersc.gov/storage/hpss

[22] DPSS: Distributed Parallel Storage System, http://www-itg.lbl.gov/DPSS/.