computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

THE CHALLENGE OF MANAGING AND ANALYZING BIG DATA

Hermann Heßling

Berlin University of Applied Sciences (HTW)
Wilhelminenhofstr. 75, D-12459 Berlin, Germany
hessling@htw-berlin.de

Abstract: The amounts of data produced in science are growing exponentially. Traditional methods for storing and maintaining the enormous flood of data seem to be no longer sufficient anymore. The complexity of the data that will be distributed more and more worldwide, is going to constitute a considerable challenge for their analysis. According to Alex Szalay there soon will be produced so many data that they cannot even be stored and maintained anymore. The data have to be analyzed in real time in order to extract the relevant information. An outline of the project Large Scale Management and Analysis (LSDMA) is given. The status of our research group on distributed real-time computing is reviewed. Finally, a novel approach to time-dependent image processing based on local thermodynamical methods is presented.

Keywords: Big Data, Real-time computing, Time-dependent image processing.

1. INTRODUCTION

Experiment and theory are pillars for scientific discoveries. With the emergence of computers new insights originated in simulations. In the meantime, data-intensive computing is considered as the fourth pillar for scientific discoveries [1].

In many scientific areas the resolution power of experimental devices is increasing strongly leading to a steadily increasing flood of data. The Large Hadron Collider (LHC) at CERN is used to explore the realm of elementary particles. The produced data rate of the order of 1 PB/s is reduced considerably in the subsequent workflow. Finally, of the order 25 PB are stored per year. The data are distributed worldwide and can be analyzed by thousands of scientists using the world largest grid computing system [2]. The radio telescope Square Kilometre Array (SKA) will allow to explore the Universe with thousands of antennas located in South Africa and Australia. First experimental results are expected in 2019. When finished in 2023 SKA will have to store of the order of 1 Exabytes per day [3].

"Soon we cannot even store the incoming data stream" (Alex Szalay [4]). Traditional approaches for analyzing and maintaining data seem to be no longer sufficient anymore. New ideas and tools are needed.

Data-intensive research is sometimes under attack. Nobel laureate Sidney Brenner criticizes most biology as low-input, high-throughput, no-output biology [5]. The LHC discovery of the Higgs particle showed what Big Data is capable of.

2. LSDMA

In the joint research and development (R&D) project "Large Scale Data Management and Analysis" (LSDMA) several Helmholtz centers and German universities are supporting scientists in analyzing, maintaining, and archiving their data [6]. Developing a one-size-fits-all solution is not feasible because of the heterogeneous requirements of the scientific communities. Therefore, the participating communities are divided into five "Data Life Cycle Labs" (DLCL).

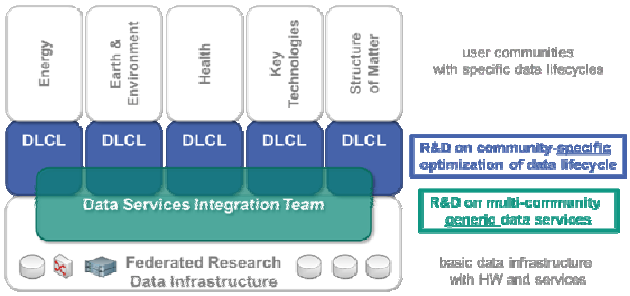

Fig. 1 – Organization of the LSDMA project.

The general goals of the DLCLs are supporting scientists in
- organizing their data and metadata,

- establishing easy access and use of local, national and international infrastructures for data storage, data processing and data archiving, and
- standardizing data management techniques in the scientific communities and their data life cycles.

The Data Services Integration Team (DSIT) develops generic services based on requests from the DLCLs:

- Federated identity management
- Federated data access
- Meta data catalogue
- Archive services
- Monitoring, modeling, optimization
- Data intensive computing

Due to lack of space it is not possible to review all LSDMA activities. The following selection is fairly subjective.

Based on the results of a DSIT workshop on AAI (Authentication, Authorization, Infrastructure) [7] it was decided to develop a distributed identity and authorization service based on Shibboleth.

dCache is a system for storing and retrieving huge amounts of data [8]. It manages approx. 50 % of the LHC data. Several access protocols are supported, e.g. GridFTP, NFS 4.1 (pNFS) and WebDAV. Cloud computing may provide a more flexible access to scientific data. For that reason, dCache is extended by the Cloud Data Management Interface (CDMI) [9]. As a cloud application an online storage is created called dBoX, to be used by the scientific community at DESY. The data are stored at DESY Hamburg and, thus, are subject to German law. The authentication is based on X.509 certificates and user/password. dBoX can be accessed via mobile devices [10].

A lot of research in the DLCLs is based on image processing. For example, the embryogenesis of vertebrates is studied with microscopes that are able to record 3D images with high spatial and temporal resolution. A tracking of the evolution of individual cells requires high data acquisition rates. By developing time-efficient algorithms and using high parallelization, data sets of the order of 10 TB from a single probe can be processed in less than a day [11], see Fig. 2.

3. DISTRIBUTED REAL-TIME COMPUTING

Data reduction in real-time will become increasingly important. For example, in photon science ultrashort X-ray flashes are used to explore nanostructures of probes in material sciences, chemistry and biology. Due to principal limitations in the experimental setup only a few percent of the data are of ~~a~~ sufficient quality for a later analysis [12]. In other words, only a fraction of the data should be stored and it is advantageous to identify useless data as early as possible. A natural strategy is to parallelize data analysis tools. In photon science diffraction images are taken, i.e. the relevant information may be smeared over the whole image. Consequently, parallelized image processing has to rely on the exchange of intermediate results. However, an efficient exchange of intermediate results is a challenging task, as large parallel systems are usually built of subsystems connected by heterogeneous networks and the memory may be distributed between the subsystems.

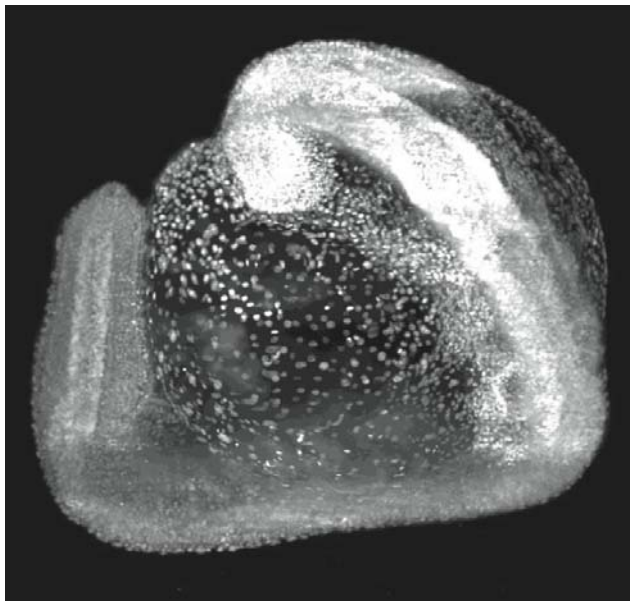

Fig. 2 – A Zebra fish embryo after 24 hours [11].

GriScha is a chess program running on distributed environments, e.g. a grid or cloud computing system [13]. It is an ideal test bed for exploring various aspects of distributed real-time computing. A Gatekeeper installs a pilot job on each participating worker node (WN), see Fig. 3. When running on a worker node the pilot job starts a chess engine client and, thereby, establishes a network connection to the external MasterNode. Worker nodes are protected by firewalls and cannot be accessed from the outside. Pilot jobs are used to overcome firewalls in that these do not block outbound connections, in general. The MasterNode distributes the game tree to the worker nodes which analyze their subtree for some period of time (default 15 s) and, then, send their best move back to the MasterNode. The MasterNode selects the best of all moves offered by the worker nodes. The chess engine on the worker nodes is very simple and evaluates a chess position like a beginner. No external intelligence is used, i.e. no data base of chess openings, no endgame tablespaces, no database of chess games. The essential question is:

can you make many beginners stronger than a master? The challenge is to establish a collective communication in huge distributed environments.

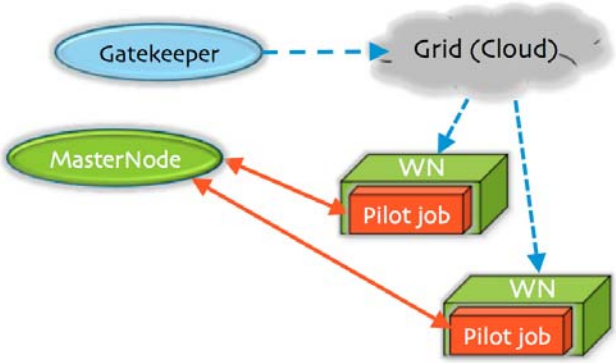

Fig. 3 – GriScha: chess in the grid.

The communication between the MasterNode and the worker nodes is based on SIMON, an extension of the RMI network protocol that allows a server to invoke methods on a client using the same socket connection [14].

In order to explore the capabilities of SIMON for exchanging large amounts of data, random message strings were send from a client over a LAN to a server. It turns out that there are jumps in the response time at some specific message lengths, see Fig. 4. A discontinuous response behavior is hardly acceptable in real-time computing.

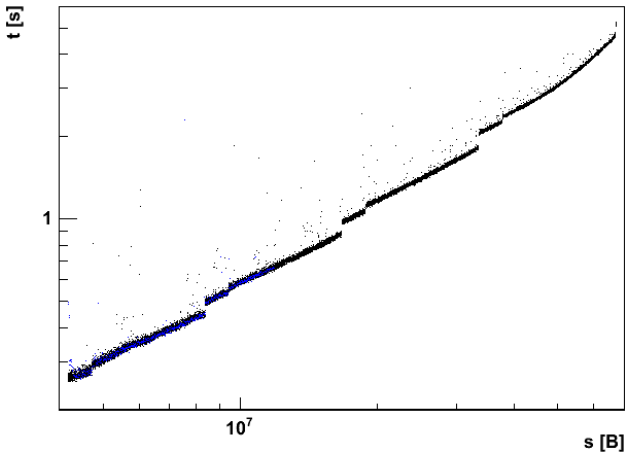

Fig. 4 – Message length s versus response time t.

The origin of the discontinuities is unclear. They depend neither on details of the network, operating system, versions of the Java virtual machine and SIMON [15], nor on the garbage collection in Java [16].

Communication tools on huge networks are based on a decentralized organization and use Peer-to-Peer (P2P) protocols. As an alternative to SIMON the open source P2P protocol JXTA was implemented in GriScha [17]. However, the future of JXTA is unclear as Oracle announced its withdrawal from this project.

After some moves a previously obtained game position may reappear again. To improve the playing strength of GriScha moves already analyzed should not be reanalyzed a second time. Instead, they should be stored in a shared memory accessible by worker nodes. When a worker node realizes that a certain move is noted it can cut out a subtree and analyze in more depth the remaining part of the game tree. JuxMem is a data-sharing grid service based on JXTA [18]. However, the future of JuxMem is also unclear. In Ref. [19] a design of a grid-compatible shared memory is suggested that is currently being realized.

The communication protocol XMPP is used to exchange messages in near real-time. The time needed for simultaneously sending messages from clients to a server is increasing linearly with the number of clients even for very short messages of only one Byte and is significantly longer compared to SIMON [20, 21]. Therefore, XMPP is of limited attractiveness for large distributed real-time systems.

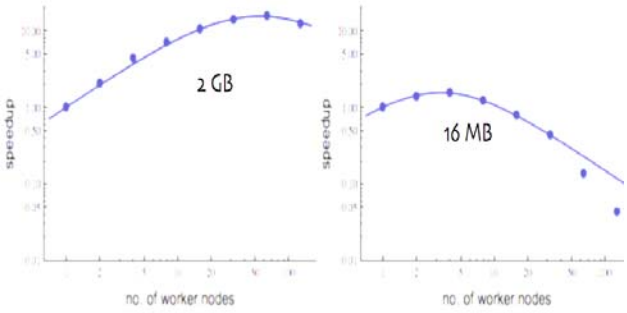

Fig. 5 – Speedup for message exchange: total message size N = 2 GB (left), N = 16 MB (right). The curve is a least-square fit to the data using Amdahl's law (1).

Speedup is an essential quantity in parallel computing. It is defined by the ratio $S(p)=T(1)/T(p)$ where $T(p)$ is the execution time of a tool running in parallel on p devices. According to Amdahl's law the speedup is limited by the sequential part of a tool since $T(p) = T_{seq} + T_{par}/p$. This formula has to be modified if latencies due to message exchange cannot be neglected,

$$T(p) = T_{seq} + T_{par}/p + (p-1)\, T_{msg}. \qquad (1)$$

Let us assume that the exchange of intermediate results needs a transport of N/n Bytes of data from n worker nodes to an external server. In Fig. 5 speedup measurement results for the grid system DECH with up to 128 worker nodes are shown. The worker nodes of DECH are located at research institutes and universities in Germany and Switzerland. The results

indicate that in the case of data exchange one should be aware of an upper bound for parallelization. Beyond this bound the system becomes inefficient as the speedup is going down.

4. TIME-DEPENDENT IMAGE PROCESSING

4D imaging is applied in various sectors. In medicine, for example, it is used to improve the success of a radiation therapy by aligning a radiation source along the motion of a tumor [22]. The upcoming European XFEL will produce up to 27,000 light pulses per second and new high-resolution detectors will open new research areas in photon science. It will be possible to take movies not only from chemical reactions but also from the motion of quantum objects like electrons [23].

In the following we present a novel method for identifying "regions of interest" in dynamically deformed objects that is based on statistical physics. For simplicity we concentrate on one-dimensional systems.

The left picture of Fig. 6 shows the motion of an object of initial length L. The object is compressed and after some time t it is decompressed. From the evolution of the "texture" of the object trajectories can be determined. In this case, the trajectories have the form

$$x_t = x_0 \cosh(gt) + p_o/(gm) \sinh(gt).$$

The momentum trajectories are given by $p_t = m\mathrm{d}x_t/\mathrm{d}t = p_o \cosh(gt) + mgx_0 \sinh(gt)$. The initial momentum is proportional to the initial position, $p_o = ax_o$.

The transformation defined by

$$X(x,p,t) = [x + t\, p/m]\cosh(gt) - [p/(gm) + gtx]\sinh(gt)$$
$$P(x,p,t) = p\cosh(gt) - mgx\sinh(gt).$$

is canonical.

Proof (sketch). The assertion follows from the fact that there is a generating function F(X,P,t) such that the Poincaré-Cartan 1-form

$$p\,\mathrm{d}x - H\,\mathrm{d}t = P\,\mathrm{d}X - P^2/(2m)\,\mathrm{d}t + \mathrm{d}F(X,P,t)$$

is fulfilled where $H = p^2/(2m) - (m/2)\, g^2 x^2$ is the Hamilton function.

From the Poincaré-Cartan 1-form follows that the trajectories with respect to the new coordinates are straight lines $X_t = x_0 + (p_0/m)\, t$.

The entropy of a 1-dim. ideal gas consisting of N particles enclosed in a volume V reads (W. Pauli)

$$S = k_B N\, [\, ½ \ln(E/N) - \ln(N/V) + s0\,]$$

where $E = ½\, N\, k_B\, T$ is the internal energy and T the temperature. For a local description we introduce the densities $s=S/V$, $n=N/V$, and $e=E/V$. Then the temperature

$$T(e,n) = (2/k_B)\; e/n$$

and the entropy density

$$s(e,n) = k_B\, n\, [\, ½ \ln(e/n) - \ln(n) + s0\,] \qquad (2)$$

can be represented in terms of the particle density n and energy density e.

The texture of an object is given by the initial gray values of an image of the object. We interpret the texture as a state function $\rho(x_0, p_0)$ describing the statistical distribution of the initial values x_0, p_0. The state function is positive, $\rho(x_0,p_0) \geq 0$, and normalized, $\int \rho(x_0, p_0)\, \mathrm{d}x_0\, \mathrm{d}p_0 = 1$. The state function of the left picture of Fig. 6 has the form

$$\rho(x_0, p_0) = f(x_0)\; \theta(x_0)\,\theta(L-x_0)\; \delta(p_o - a\, x_o)$$

where $\theta(x)$ is the Heaviside step function and $\delta(x)$ the Dirac function. The shape of the texture is given by

$$f(x_0) = [(3/2)^4 + x_0(L-x_0)\,(x_0 - L/2)^2]/[L(3/2)^4 + L^5/120].$$

In statistical physics the time evolution of the mean value of an observable $A(x,p,t)$ can be determined in the Heisenberg picture

$$\langle A_t \rangle = \int A(x_t,p_t,t)\; \rho(x_0, p_0)\; \mathrm{d}x_0\; \mathrm{d}p_0.$$

Consider the point-like observables

$$n_{x^*}(x) = \delta(x - x^*),\; e_{x^*}(x,p,t) = 1/(2m) P^2(x,p,t)\, \delta(x - x^*).$$

Their expectation values $e_{x^*,t}$ and $n_{x^*,t}$ are inserted into Eq. (2) in order to obtain the local entropy $s_{x^*,t} = s(e_{x^*,t}\,,\, n_{x^*,t}\,)$.

The two local maxima of the particle density follow the trajectories and are most prominent when the object is maximally compressed, see Fig. 6 (left panel). The maximum at the bottom is larger than the local maximum at the top. The local entropy shows also two maxima, see the yellow regions in Fig. 6 (right panel, the blue colors represent smaller values). However, the entropy is maximal in the top region where the curvature of the trajectories is stronger. In other words, the local entropy may be useful for rating textures.

Fig. 7 (left panel) shows that the total entropy is not constant. As long as the object is compressed the entropy is decreasing.

The energy density $e_{x^*,t}$ and the particle density $n_{x^*,t}$ are conserved. The resulting continuity equations can be used to derive a balance equation for the local entropy that, in turn, allows to extract the entropy production σ [24]. As can be seen in Fig. 7 (right panel) the entropy production changes its sign in certain subregions of the system and in the course of time.

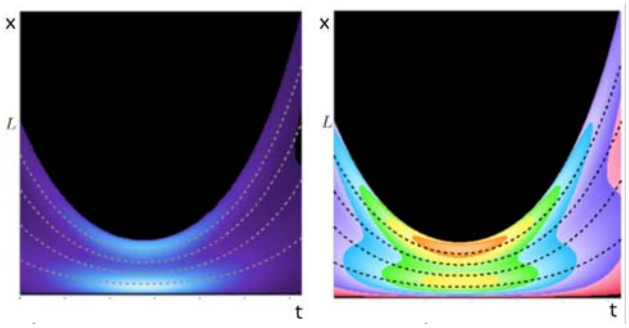

Fig. 6 – Particle density (left) and local entropy.

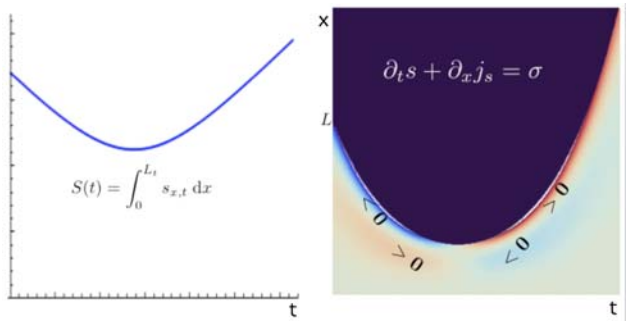

Fig. 7 – Total entropy (left) and entropy production.

5. CONCLUSION

LSDMA is a project on Big Data in science. It follows a dual approach by developing community-specific and generic services. To handle the rapidly growing deluge of data real-time analysis of Big Data will become increasingly important. In distributed real-time computing data exchange may limit the degree of parallelization.

A novel method for specifying "regions of interest" in time-dependent image processing is presented which uses methods of statistical physics. It is applied to a simple 1-dimensional example. A generalization to higher dimensions and more involved dynamics with time-dependent Hamilton functions is feasible.

6. ACKNOWLEGEMENTS

Stimulating discussions with many members of the LSDMA project are gratefully acknowledged.

7. REFERENCES

[1] T. Hey, S. Tansley, and K. Tolle (Eds.), *The fourth paradigm, data-intensive scientific discovery*, Microsoft Cooperation, 2009. Available at http://research.microsoft.com/en-us/collaboration/ fourthparadigm.

[2] http://wlcg.web.cern.ch

[3] P. E. Dewdney, *SKA1 system baseline design*, SKA-Tel-SKO-DD-001 (2013).

[4] A. Szalay, *Extreme data-intensive computing in science*, at: 1st International LSDMA Symposium *The Challenge of Big Data in Science*, Karlsruhe, Germany (25 September 2012).

[5] E. C. Friedberg, an Interview with Sydney Brenner, *Nature Reviews Molecular Cell Biology*, (9) (2008), pp. 8-9.

[6] http://www.helmholtz-lsdma.de.

[7] LSDMA IDM-Workshop, DESY, Hamburg, Germany (March 11, 2013).

[8] http://www.dCache.org

[9] J. Weschenfelder, *CDMI for dCache*, Internal report, University of Applied Sciences (HTW), Berlin, 2013. (in German)

[10] L. Blöcher, and T. Schubert, private communication, 2013.

[11] A. Kobitskiy, G. U. Nienhaus, J. C. Otte, M. Takamiya, U. Strähle, J. Stegmaier, and R. Mikut, *Light Sheet Microscope (LSM)*, in: R. Stotzka (Ed.), *Data Life Cycle Lab. Key Technologies, Big Data in Science, Status 2013*. Available at http://digbib.ubka.uni-karlsruhe.de/volltexte/1000037134.

[12] A. Barty, The coming deluge of data from XFEL sources, *LSDMA Workshop*, DESY, Hamburg, Germany, March 12, 2013.

[13] H. Heßling, Real-time grid computing: recent results on a pilot job approach, DESY Computing Seminar, University Hamburg, Germany, February 7, 2011.

[14] SIMON (Simple Invocation of Methods over Networks), http://dev.root1.de/projects/simon.

[15] P. Eckert, *Optimization in Java: Client-Server Communication*, Internal report, University of Applied Sciences (HTW), Berlin, 2013. (in German).

[16] P. Eckert, *Garbage Collection in Java*, Internal report, University of Applied Sciences (HTW), Berlin, 2013. (in German).

[17] C. Lehmann, *Studies on Using Peer-to-Peer Techniques in Grid Computing*, Master Thesis, University of Applied Sciences (HTW), Berlin, 2013. (in German).

[18] http://juxmem.gorge.inria.fr.

[19] K. Kochan, *Distributed Tree Search in GriScha*, Internal report, University of Applied Sciences (HTW), Berlin, 2013. (in German).
[20] L. Bortfeld, *Real-time Communication in Grid Computing based on XMPP*, Internal report, University of Applied Sciences (HTW), Berlin, 2013. (in German).
[21] P. Stewart, *Real-time Communication in Grid Computing based on XMPP*, Internal report, University of Applied Sciences (HTW), Berlin, 2013. (in German).
[22] J. Erhard, and C. Lorenz (Eds.), *4D Modeling and Estimation of Respiratory Motion for Radiation Therapy*, Springer, Heidelberg, 2013.
[23] G. Dixit, J. M. Slowik, and R. Santra, *Proposed imaging of the ultrafast electronic motion in samples using X-ray phase constrast*, Physical Review Letters, (110) 13 (2013), 137403.
[24] D. Kondepudi, and I. Prigogine, *Modern Thermodynamics. From Heat Engines to Dissipative Structures*, John Wiley, 1998.

Hermann Heßling studied Physics at the Universities of Münster, Göttingen and Hamburg. He received the Ph.D. (Dr. rer. nat) in Theoretical Physics and was appointed a postdoctoral research fellow at Deutsches Elektronen-Synchrotron (DESY), Hamburg (1993-1996).

Subsequently, he continued his work with a computer communications and networking company and accepted in 1999 an offer from the University of Applied Sciences Hof as a Professor of Operating Systems. Since 2000 he has been Professor of Applied Informatics at the University of Applied Sciences HTW Berlin. His research areas include distributed real-time computing and Big Data.