КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2015 Т. 7 № 3 С. 559−563

Ки&М

СЕКЦИОННЫЕ ДОКЛАДЫ

УДК: 004.63

Distributed dCache-based storage system of UB RAS

E. Yu. Kuklin[1,3,a], A. V. Sozykin[1,3], A. Yu. Bersenev[1,3], G. F. Masich[2]

[1] Institute of Mathematics and Mechanics UB RAS, Sofia Kovalevskaya Str. 16, Yekaterinburg, 620990, Russia
[2] Institute of Continuous Media Mechanics UB RAS, St. Academ. Koroleva 1, Perm, 614013, Russia
[3] Ural Federal University, St. Mira 19, Yekaterinburg, 620002, Russia

E-mail: [a] key@imm.uran.ru

Received October 10, 2014

The approach to build territorial distributed storage system for high performance computing environment of UB RAS is presented. The storage system is based on the dCache middleware from the European Middleware Initiative project. The first milestone of distributed storage system implementation includes the data centers at the two UB RAS Regions: Yekaterinburg and Perm.

Keywords: network attached storage, parallel NFS, GRID, HPC

Распределенная система хранения УРО РАН на основе dCache

Е. Ю. Куклин[1,3], А. В. Созыкин[1,3], А. Ю. Берсенёв[1,3], Г. Ф. Масич[2]

[1] Институт Математики и Механики УрО РАН, Россия, 620990, г. Екатеринбург, ул. Софьи Ковалевской, д. 16
[2] Институт Механики Сплошных Сред УрО РАН, Россия, 614013, г. Пермь, ул. Акад. Королёва, д. 1
[3] Уральский Федеральный Университет, Россия, 620002, г. Екатеринбург, ул. Мира, д. 19

Представлен подход к созданию территориально-распределенной системы хранения данных для нужд среды высокопроизводительных вычислений УрО РАН. Система основывается на промежуточном программном обеспечении dCache из проекта European Middleware Initiative. Первая очередь реализации системы охватывает вычислительные центры в двух регионах присутствия УрО РАН: г. Екатеринбург и г. Пермь.

Ключевые слова: сетевые системы хранения, parallel NFS, ГРИД-технологии, параллельные вычисления

Supported by the grant of UB RAS 15-7-1-26 and by the RFBR grant 14-07-96001r_ural_a.

Citation: *Computer Research and Modeling*, 2015, vol. 7, no. 3, pp. 559–563.

Introduction

Works on creating a distributed high-performance computing environment based on GRID technologies are under way at the Ural Branch of the Russian Academy of Sciences. One of the main components of this environment is a distributed data storage system, which aims at integrating storage systems in the Ural regions [Goldshtein et al., 2013]. The system connects various resources, such as computing clusters, supercomputers and experimental setups of Ural institutes and universities. The participants of this storage system project are Institute of Mathematics and Mechanics in Yekaterinburg (computational recourses and storage system) and Institute of Continuous Media Mechanics in Perm (backbone networks).

Middleware selection

The approach to building territorial distributed storage system based on the dCache [dCache..., 2015] middleware from the European Middleware Initiative (EMI) [EMI..., 2015] project is presented. dCache is a distributed storage system focused on storing large amounts of experimental data. It can run on commodity hardware and allows the construction of storage facilities in hundreds of terabytes, with all the files in it logically organized into a single virtual file system tree. In addition, dCache assumes a simple extension of your storage by adding new nodes and can work with tape libraries. dCache supports a wide range of access protocols. Together with common standard protocols FTP, WebDAV, NFS, grid protocols SRM and GRIDFTP, as well as its own protocol dCap is used.

EMI includes three projects for building distributed storage systems: dCache, Disk Pool Manager and Storage Resource Manager. They applied in different projects to build GRID infrastructures, includes WLCG [WLCG..., 2015]. For UB RAS storage dCache was chosen, as it provides support for both GRID and Internet protocols, has a high quality documentation, as well as it is easy to install and administer. An additional reason for choice was the fact that there is a dCache-based store at the Joint Institute for Nuclear Research, and the only one distributed Tier1 center — Nordic Data Grid Facility, which storage nodes are located in different Nordic countries [Behrman et al., 2008].

Current stage

The first milestone of the implementation of distributed storage system running dCache 2.6 has been currently completed. Figure 1 shows its general plan. For data storing were selected servers by Supermicro. Now there are 4 servers running Scientific Linux 6.5 with usable capacity of 210 TB. We decided to store our data in the XFS file system by Silicon Graphics. It has shown good results in tests for read/write/access to data, and as well as EXT4, is native for Linux and actively developing file system. For example, our test using bonnie++ [Bonnie..., 2015] showed that there is no the distinct advantages between file systems (except random delete, what exactly our system does not imply). So, taking into account the turn of RHEL 7 to XFS, it was decided to leave the XFS.

The storage nodes are located in two computing centers: in Institute of Mathematics and Mechanics in Yekaterinburg (3 nodes) and Institute of Continuous Media Mechanics in Perm. The computing centers are separated by a 450 km distance and are joined by a dedicated communication channel. Channel performance provides by DWDM equipment of ECI-Telecom Company. Installed platforms allows transmission of two λ-channels with 10 Gigabit Ethernet technology. Setup and maintenance of backbone network is engaged by Laboratory of Telecommunication and Information Systems in Institute of Continuous Media Mechanics.

As shown at the storage system plan, Institute of Mathematics and Mechanics has at its disposal a supercomputer "URAN", which occupies the 9th position in Top50 CIS; Institute of Continuous Me-

dia Mechanics also has its own cluster. Connection to storage system was performed using the NFS protocol version 4.1 with Parallel NFS [Parallel NFS..., 2015]. On cluster's nodes mounted, as dCache structure suggests, one of the storage system servers (broker host). Parallel NFS allows direct connection between computational and storage nodes for data transfers, removing the traditional NFS-server bottleneck.

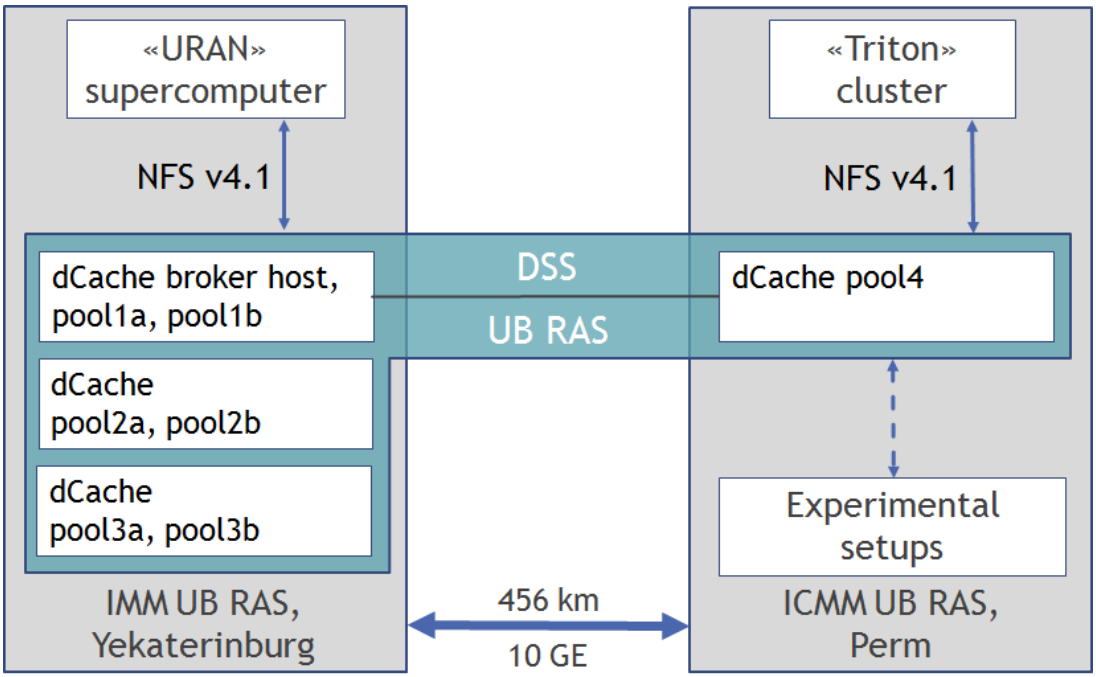

Fig. 1. General plan of UB RAS DSS

There are some unresolved for now drawbacks: how NSF will behave when transmitting data over a long distance, or how to optimize the OS network stack configuration for 10-Gigabit networks (the efficiency of Intel network cards on storage nodes leave much to be desired). Also, there are problems with mounting dCache on servers with a custom build kernel.

Network benchmarks

First of all, testing has shown that the evident bottleneck is the gigabit network between the storage system and supercomputer "URAN". 10-Gigabit equipment (with switch by Extreme Networks) for the internal network allowed fixing it, and significantly increasing the speed of data exchange. Figure 2 shows the results of the data recording via a dedicated channel between the data centers before and after attaching an additional storage node in Perm. It can be seen that the presence of the local server slightly improves storage system performance. These results were obtained with the IOR benchmark [IOR..., 2015].

Figure 3 displays an attempt to optimize the TCP/IP stack parameters on storage nodes. Our colleagues from Perm conducted tests using iperf [Iperf..., 2015] and various optimization algorithms. As can be seen, the best result was given by bic algorithm, but it could not overcome the threshold of 6 Gbit/s. This is expected has been given by Intel network cards installed on the servers.

dCache allows to create multiple copies of files that balance loading between the nodes and optimize the use of storage space. Figure 4 shows that automatic replication does not affect the system performance, so with the hardware RAID arrays, it will increase the reliability of the system, because we do not have at our disposal a tape library for archiving data.

Figure 2. Performance recording on DSS from ICMM client with and without an additional storage node in Perm

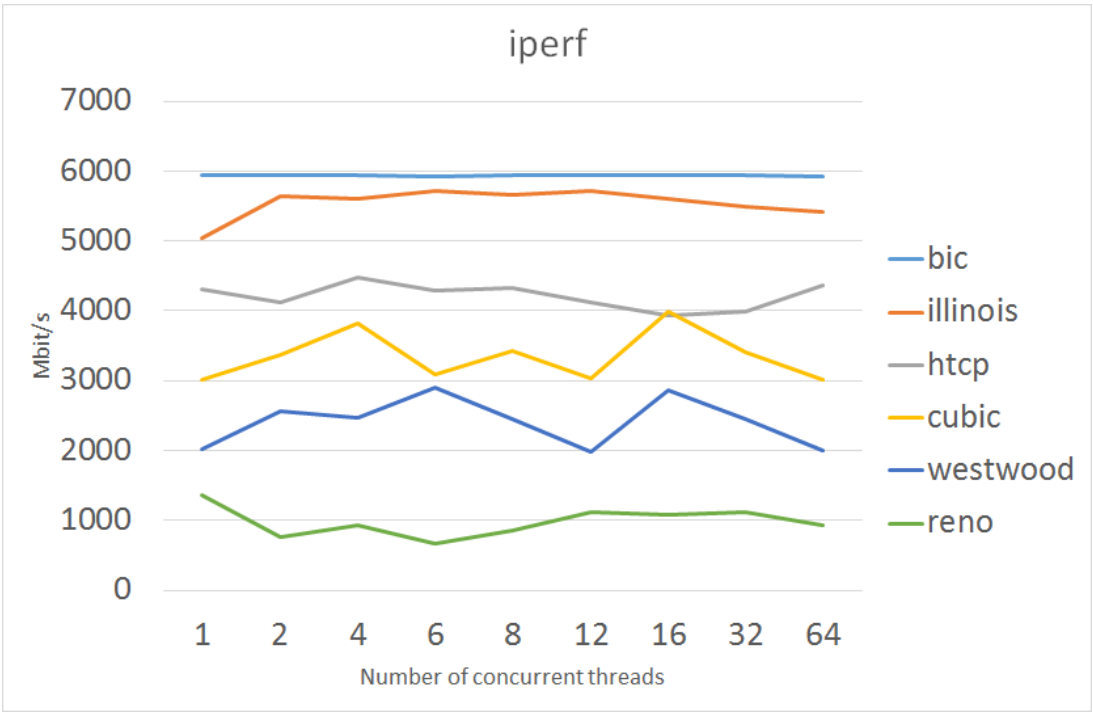

Fig. 3. Comparison of congestion control algorithms for TCP/IP networks

Conclusion

Now Institute of Mathematics and Mechanics conducted performance and fault tolerance benchmarks of considered storage system. Among them was investigated the effect of additional storage node and replication on the performance of the storage system. Also was selected congestion control algorithm to optimize the network parameters. In the nearest future we plan to pay more attention on security and monitoring. The next stage of implementation will be increasing of storage system

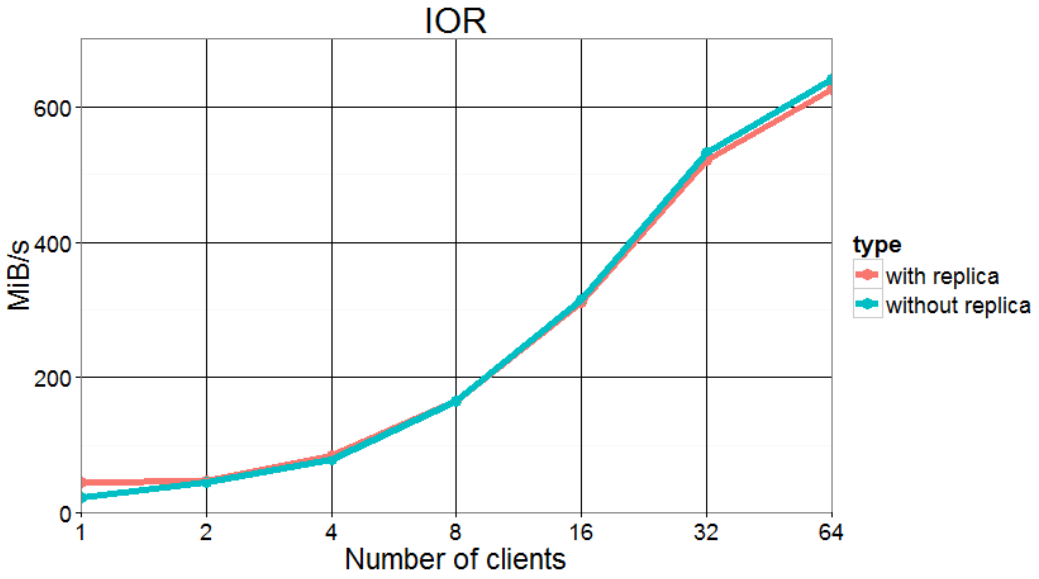

Fig. 4. Influence of replication on DSS performance

capacity and attaching the experimental setups in Institute of Continuous Media Mechanics in Perm and Ural Federal University. Data obtained from them will be recorded to the storage system and processed remotely by supercomputer "URAN", including the ability of process in real time and control the experiments. Further the connection of entire distributed computing environment to international GRID infrastructure Worldwide Large Hadron Collider Computing Grid is planned. We can provide for general use the resources of distributed storage system and computing clusters resources.

References

Behrman G. , Fuhrmann P. , Gronager M. and Kleist J. A distributed storage system with dCache // Journal of Physics: Conference Series, 119, 2008;

Bonnie++ [online] // — 2015. — URL: http://www.coker.com.au/bonnie++ (дата обращения: 16.01.2015);

dCache web-page [online] // — 2015. — URL: http://www.dcache.org (дата обращения: 16.01.2015);

EMI - European Middleware Initiative [online] // — 2015. — URL: http://www.eu-emi.eu (дата обращения: 16.01.2015);

Goldshtein M.L., Sozykin A.V., Masich G.F., Masich A.G. "Computing resources of UB RAS. Status and prospects." // Parallel Computational Technologies (PCT'2013): proceedings of the international scientific conference. Chelyabinsk, publishing center of SUSU, 2013. P. 330-337;

IOR [online] // — 2015. — URL: http://sourceforge.net/projects/ior-sio (дата обращения: 16.01.2015);

Iperf [online] // — 2015. — URL: https://iperf.fr (дата обращения: 16.01.2015);

Parallel NFS [online] // — 2015. — URL: http://www.pnfs.com (дата обращения: 16.01.2015) ;

WLCG — Worldwide LHC Computing Grid [online] // CERN, Switzerland — 2014 — URL: http://wlcg.web.cern.ch (дата обращения: 16.01.2015);