Dissertação de Mestrado

Estudo e Implementação de Códigos Corretores de Erros no Sistema de Arquivos Distribuído do Hadoop

Celina d' Ávila Samogin
Instituto de Computação — UNICAMP

13 de novembro de 2012

Orientadora: Profa. Dra. Islene Calciolari Garcia

Agenda

Motivação

- Este trabalho é uma contribuição para *software* livre em sistemas distribuídos.
- Armazenamento de arquivos é um componente essencial na computação de alto desempenho.
- Códigos Corretores de Erro (*Erasure codes*) introduzem redundância e tem sido utilizados em sistemas para alcançar confiabilidade e redução do custo de armazenamento.

Motivação

- Alguns sistemas que utilizam códigos corretores de erros:
  - *NASA's Deep Space Network* no envio e na recepção de sinais e dados de telemetria (*downlinks*) vindos de veículos espaciais (*very distant spacecrafts*) e para enviar telecomandos (*uplinks*) para veículos espaciais [1, 2, 5, 17];
  - *Delay and Disruption Tolerant Networks*, redes de sensores e redes *peer-to-peer* [4, 6, 9, 10, 12, 14, 20];
  - armazenamento de grande volume de dados [3, 11, 15, 16, 18, 19, 21], como também o sistema de arquivos distribuído do Hadoop (HDFS) [7].

Motivação

- O HDFS, por padrão, implementa alta disponibilidade dos dados via replicação simples dos blocos de dados. Esta abordagem acarreta um alto custo de armazenamento para garantir que os dados estarão sempre disponíveis.
- Esforços iniciais nessa linha foram feitos utilizando técnicas de *Redundant Array of Independent Drives* (RAID) [7, 13] e mais recentemente do algoritmo Reed-Solomon (RS) [8].

Motivação

- RAID
- Reed-Solomon

Objetivos deste trabalho

- avaliação de desempenho, ganhos, e custos de diferentes estratégias de códigos corretores de erro;
- implementação de otimizações ou extensões para o código que atualmente implementa Reed-Solomon, tentando melhorar, principalmente, a parte de distribuição de blocos;
- implementação de novos algoritmos (e.g., Tornado codes) e extensão da interface atual para aceitá-los;
- integração do código atual com o HDFS.

Agenda

Motivação

- Software Livre, Open-source
- Sistemas Distribuídos

Agenda

Referências Bibliográficas I

Silvio A. Abrantes.
Códigos Corretores de Erros em Comunicações Digitais, chapter 1,2,3,4,5,6,7,8,9, pages 17–600.
FEUP edições, Porto, Portugal, 2001.

G. M. Almeida.
Códigos corretores de erros em hardware para sistemas de telecomando e telemetria em aplicações espaciais.
Master's thesis, Pontifícia Universidade Católica do Rio Grande do Sul - Faculdade de Informática, Porto Alegre, Brasil, março 2007.

T. Anderson, M. Dahlin, J. Neefe, D. Roselli, R. Wang, and D. Patterson.
Serverless network file systems.
Technical report, University of California at Berkeley, Berkeley, CA, USA, 1998.

Referências Bibliográficas II


R. Bhagwan, K. Tati, Y. Cheng, S. Savage, and G. M. Voelker.
Total recall: system support for automated availability management.
In *Proceedings of the 1st conference on Symposium on Networked Systems Design and Implementation - Volume 1*, pages 25–25, Berkeley, CA, USA, 2004. USENIX Association.


A. R. Curtis.
Space today online - communicating with interplanetary spacecraft.
URL=http://www.spacetoday.org/SolSys/DeepSpaceNetwork/DeepSpaceNetwork.html. Acessado em 03 de maio de 2010.


Marcelo Sampaio de Alencar.
Telefonia Celular Digital, chapter 6, pages 191–231.
Editora Érica, São Paulo, Brasil, 2004.

Referências Bibliográficas III


Apache Software Foundation.
Hdfs-503 - implement erasure coding as a layer on hdfs.
URL=https://issues.apache.org/jira/browse/HDFS-503. Acessado em 08 de maio de 2010.


Apache Software Foundation.
Mapreduce-1969 - allow raid to use reed-solomon erasure codes.
URL=https://issues.apache.org/jira/browse/MAPREDUCE-1969. Acessado em 20 de agosto de 2010.


A. Haeberlen, A. Mislove, and P. Druschel.
Glacier: highly durable, decentralized storage despite massive correlated failures.
In *NSDI'05: Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation*, pages 143–158, Boston, MA, USA, 2005. USENIX Association.

Referências Bibliográficas IV

Y. Houri, M. Jobmann, and T. Fuhrmann.
Self-organized data redundancy management for peer-to-peer storage systems.
In *IWSOS '09: Proceedings of the 4th IFIP TC 6 International Workshop on Self-Organizing Systems*, pages 65–76, Berlin, Heidelberg, 2009. Springer-Verlag.

J. Kubiatowicz, D. Bindel, Y. Chen, S. Czerwinski, P. Eaton, D. Geels, R. Gummadi, S. Rhea, H. Weatherspoon, W. Weimer, C. Wells, and B. Zhao.
Oceanstore: an architecture for global-scale persistent storage.
In *ASPLOS '00: Proceedings of the ninth international conference on Architectural support for programming languages and operating systems*, ASPLOS-IX, pages 190–201, New York, NY, USA, 2000. ACM.

Referências Bibliográficas V


C. T. Oliveira, M. D. D. Moreira, M. G. Rubinstein, L. H. M. K. Costa, and O. C. M. B. Duarte.
Mc05: Redes tolerantes a atrasos e desconexões.
In *Anais do 25o Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, Belém, Pará, Brasil, May 2007.


D. A. Patterson, G. Gibson, and R. H. Katz.
A case for redundant arrays of inexpensive disks (raid).
In *SIGMOD '88: Proceedings of the 1988 ACM SIGMOD International Conference on Management of Data*, pages 109–116, New York, NY, USA, 1988. ACM.


R. Rodrigues and B. Liskov.
High availability in dhts: Erasure coding vs. replication.
In *Peer-to-Peer Systems IV*, pages 226–239. LNCS, 2005.

Referências Bibliográficas VI


Y. Saito, S. Frølund, A. Veitch, A. Merchant, and S. Spence.
Fab: building distributed enterprise disk arrays from commodity components.
In *Proceedings of the 11th international conference on Architectural support for programming languages and operating systems*, ASPLOS-XI, pages 48–58, New York, NY, USA, 2004. ACM.


F. Schmuck and R. Haskin.
Gpfs: A shared-disk file system for large computing clusters.
In *Proceedings of the 1st USENIX Conference on File and Storage Technologies*, FAST '02, Berkeley, CA, USA, 2002. USENIX Association.


R. W. Sniffin.
Telemetry data decoding.
URL=http://deepspace.jpl.nasa.gov/dsndocs/810-005/208/208A.pdf. Acessado em 03 de maio de 2010.

Referências Bibliográficas VII


M. W. Storer, K. M. Greenan, E. L. Miller, and K. Voruganti.
Pergamum: replacing tape with energy efficient, reliable, disk-based archival storage.
In *FAST'08: Proceedings of the 6th USENIX Conference on File and Storage Technologies*, FAST'08, pages 1:1–1:16, Berkeley, CA, USA, 2008. USENIX Association.


M. W. Storer, K. M. Greenan, E. L. Miller, and K. Voruganti.
Potshards: a secure, recoverable, long-term archival storage system.
Trans. Storage, 5:5:1–5:35, June 2009.


Z. Wilcox-O'Hearn B. Warner.
Tahoe: the least-authority filesystem.
In *Proceedings of the 4th ACM international workshop on Storage security and survivability*, StorageSS '08, pages 21–26, New York, NY, USA, 2008. ACM.

Referências Bibliográficas VIII


H. Xia.
Robustore: a distributed storage architecture with robust and high performance.
PhD thesis, University of California at San Diego, La Jolla, CA, USA, 2006.
AAI3225997.