Wpływ organizacji danych na rozproszony system pamięci masowej

Autoreferat pracy doktorskiej

Leszek Gryz

styczeń 2012

Wraz z rozwojem technologii informacyjnych ilość wytworzonej informacji rośnie w postępie geometrycznym [8]. W 2009 roku wytworzono prawie 800 trylionów bajtów, co stanowi 62% wzrost względem poprzedniego roku. W 2010 wytworzono o 50% danych więcej, tj. prawie 1200 trylionów bajtów [7].

Praktycznie wszystkie nowo powstałe dane są przechowywane w postaci cyfrowej. Bardzo szybko wzrastająca ilość wytworzonej informacji przekłada się na podobny wzrost popytu na systemy przechowujące dane, w szczególności na systemy przechowujące kopie zapasowe danych. Popyt ten jest dodatkowo wzmacniany przez wymogi prawne nakazujące przechowywanie pewnych danych przez wiele lat [10, 2]. Jednakże obecnie stosowane rozwiązania nie są przystosowane do efektywnego zarządzania tak szybko rosnącą i dużą ilością danych.

Tradycyjnie do długotrwałego przechowywania cyfrowych danych wykorzystuje się rozwiązania oparte na bibliotekach taśm magnetycznych. W ostatnich latach na popularności zyskują także rozwiązania bazujące na magnetycznych dyskach twardych. Dyski twarde zorganizowane są w formy DAS (ang. Direct-Attached Storage), NAS (ang. Network Attached Storage) i SAN (ang. Storage Area Network).

DAS jest najprostszą formą organizacji dysków. Kilka z nich jest bezpośrednio zainstalowanych w serwerze. To powoduje, że serwer staje się pojedynczą *wyspą pamięci masowej* (ang. *storage island*), ponieważ dane nie są dzielone pomiędzy serwerami. Takie rozwiązanie jest nieskalowalne i nieefektywnie wykorzystuje przestrzeń dyskową. Ze względu na problemy wynikające z zarządzaniem wieloma DAS, takich rozwiązań nie stosuje w przypadku dużej ilości danych.

NAS integruje wiele dysków, tworząc jedno urządzenie, które poprzez

sieć komputerową udostępnia interfejs systemu plików. Dzięki temu dane mogą być dzielone przez wiele serwerów. NAS są bardziej skalowalne niż DAS, ale w dalszym ciągu tworzą pojedyncze wyspy pamięci masowych, ponieważ skalowalność pojedynczego NAS jest ograniczona przez liczbę dysków jaką pojedynczy serwer NAS może obsłużyć. Aby przechować duże ilości danych potrzebnych jest wiele systemów NAS. Administratorzy statycznie alokują dane do NAS, co prowadzi do niskiej wydajności i nieefektywnego wykorzystania przestrzeni dyskowej. Ponadto każdy serwer NAS jest pojedynczym punktem awarii.

SAN wykorzystuje sieć komputerową do połączenia wielu dysków, które mogą być podzielone na wiele logicznych jednostek. Serwery obsługują swoje własne systemy plików na tych logicznych jednostkach. Jednak ze względu na brak współdzielenia jednostek pomiędzy serwerami, w dalszym ciągu dochodzi do powstania wysp pamięci masowej. Ponadto wstępny przydział przestrzeni dyskowej do systemu plików skutkuje jej nieefektywnym wykorzystaniem.

Biblioteki taśm magnetycznych osiągają rozmiary rzędu petabajtów. Stanowią jednak nieelastyczne rozwiązanie, ponieważ wymagają dużych inwestycji początkowych. Tym samym uniemożliwiają rozpoczęcie działania z małym systemem i stopniowe powiększanie go w miarę rosnących potrzeb. Ponadto biblioteki taśmowe mają bardzo długi czas dostępu, ponieważ żądana taśma musi być załadowana do czytnika i przewinięta w odpowiednie miejsce. Kolejną wadą jest wymaganie składowania taśm z danymi w specjalnie do tego przystosowanych pomieszczeniach, zapewniających odpowiednią temperaturę i wilgotność powietrza. Rozwiązania oparte na taśmach magnetycznych wymagają od administratora dużego zaangażowania w proces tworzenia kopii zapasowych i zarządzania nimi.

Powyższe rozwiązania mają kilka dodatkowych wspólnych wad. Żadne z nich nie zapewnia elastycznej odporności na awarie. Rozwiązania dyskowe standardowo bazują na schematach RAID-5 [11] lub RAID-6 [15], nie chroniąc przed utratą danych w sytuacji, gdy awarii ulegnie więcej niż odpowiednio jeden lub dwa dyski. Natomiast biblioteki taśmowe w ogóle nie są odporne na awarie. Zniszczenie jednej taśmy prowadzi do utraty danych, często także danych znajdujących się na innych taśmach. Aby temu zapobiec, typową praktyką jest tworzenie nadmiarowych kopii danych, co powoduje nieefektywne wykorzystanie zasobów. Co najważniejsze, powyższe rozwiązania nie eliminują duplikujących się danych. Taka cecha bardzo zwiększyłaby efektywność wykorzystania nośników danych użytych do przechowywania kopii zapasowych, ponieważ kopie zapasowe tych samych

danych różnią się od siebie w bardzo niewielkim stopniu.

Niedostosowanie istniejących systemów do zmieniających się potrzeb powoduje, że w dziedzinie przechowywania danych muszą powstać nowe rozwiązania przeznaczone do zarządzania ich bardzo dużą i szybko rosnącą ilością. Trendy rynkowe, takie jak ciągły wzrost pojemności dysków twardych, pojawienie się relatywnie tanich wielordzeniowych procesorów oraz szybkie sieci komputerowe, umożliwiają budowę nowego typu systemów - skalowalnych, *rozproszonych systemów pamięci masowych* (RSPM) (ang. DSS *distributed storage systems*). RSPM cechują się automatycznym zarządzaniem, efektywnym wykorzystaniem przestrzeni dyskowej poprzez deduplikację danych i brak statycznego alokowania danych oraz wysoką dostępnością i odpornością na awarie.

Spełnienie wymagań RSPM bezpośrednio zależy od sposobu organizacji danych systemu. Problemem jest to, że niemożliwe jest zaprojektowanie takiej organizacji danych, która umożliwiałaby systemowi optymalne spełnienie każdego z jego wymagań z osobna. Dzieje się tak, ponieważ dla wielu par wymagań organizacja danych, która zwiększa stopień spełnienia jednego wymagania obniża stopień spełnienia drugiego. Trzeba więc znajdować rozwiązania, które łagodzą te konflikty. W rozprawie opisane są takie konfliktujące się wymagania i powody, dla których organizacja danych nie może spełniać ich jednocześnie. Następnie opisana jest proponowana przez nas nowatorska organizacja danych, która uwzględnia te konflikty. Skuteczność naszej organizacji danych potwierdza zbudowany przez nas komercyjny system RSPM zwany HYDRAstor [4], w którego prace badawcze, projekt i realizację byłem zaangażowany od samego początku.

HYDRAstor, według naszej wiedzy, jest jedynym system realizującym wszystkie wymagania stawiane systemom RSPM. System HYDRAstor został uhonorowany trzema nagrodami. W 2007 roku otrzymał nagrodę produktu roku w kategorii systemów służących do tworzenia kopii zapasowych i odzyskiwania danych po awarii nadaną przez "TechTargets Storage magazine" i "SearchStorage.com". W 2008 roku system otrzymał nagrodę jako najbardziej innowacyjny produkt nadaną przez "Network Products Guide", a także nagrodę nadaną przez "2008 American Business Award" za łatwo zarządzalny system pamięci masowej cechujący się nieograniczoną skalowalnością, wysokim bezpieczeństwem i niezawodną deduplikacją danych. Jest także najszybszym system zapewniającym deduplikację danych [12].

W rozdziale pierwszym opisana jest motywacja dla budowania rozproszonych systemów pamięci masowej i zdefiniowany jest problem badawczy.

W rozdziale drugim zdefiniowany jest rozproszony system pamięci ma-

sowej wraz z klasyfikacją funkcjonalną tych systemów. Podana jest też motywacja z powodu której skupiliśmy się tylko na systemach dedykowanych dla tworzenia kopii zapasowych i archiwów danych cyfrowych.

W rozdziale trzecim opisane są wymagania RSPM, których spełnienie zależy od właściwej organizacji danych.

W rozdziale czwartym opisane są pary wymagań RSPM takich, że organizacja danych zwiększająca stopień spełnienia jednego wymagania zmniejsza stopień spełnienia drugiego wymagania.

W rozdziale piątym opisana jest nowatorska organizacja danych powstała w wyniku naszych badań i twórczego wykorzystania najnowszych technologii takich jak jednokierunkowe funkcje mieszające (np. SHA-1 [6]), adresowanie na podstawie zawartości (ang. CAS, content-addressable storage) [1, 13, 16], rozproszone tablice mieszające [5, 9, 14] i kody korekcyjne typu "erasure codes" [3]. Ponadto rozdział ten zawiera dyskusję, w jakim stopniu udało się zrealizować wymagania i w jaki sposób proponowana organizacja danych uwzględnia konfliktujące się wymagania.

W rozdział szóstym opisana jest organizacja danych konkurencyjnych systemów pamięci masowych.

Rozdział siódmy jest podsumowaniem ze wskazaniem dalszych, potencjalnych kierunków prac.

Reasumując, w rozprawie:

- Zidentyfikowane zostały wymagania rozproszonych systemów pamięci masowych, których spełnienie wymaga odpowiedniej organizacji danych.
- Zidentyfikowany został problem, polegającym na tym, że istnieją pary wymagań, które konfliktują ze sobą poprzez organizację danych (jak opisane powyżej).
- Zaproponowana została organizacja danych spełniająca wymagania rozproszonych systemów pamięci masowych i uwzględniająca problemy konfliktujących się wymagań.
- Przeanalizowano w jakim stopniu proponowana organizacja danych spełnia wymagania i z jakim skutkiem rozwiązuje problemy konfliktujących się wymagań.

Bibliografia

[1] EMC Centera: content addressed storage system, January 2008. http://www.emc.com/centera.

[2] United States of America 107th Congress. Public Law 107-204: "Sarbanes-Oxley Act of 2002". July 2002.

[3] Johannes Blömer, Malik Kalfane, Marek Karpinski, Richard Karp, Michael Luby, and David Zuckerman. An xor-based erasure-resilient coding scheme. Technical Report TR-95-048, International Computer Science Institute, August 1995.

[4] Cezary Dubnicki, Leszek Gryz, Lukasz Heldt, Michal Kaczmarczyk, Wojciech Kilian, Przemyslaw Strzelczak, Jerzy Szczepkowski, Cristian Ungureanu, and Michal Welnicki. HYDRAstor: a Scalable Secondary Storage. In *FAST '09: Proccedings of the 7th conference on File and storage technologies*, pages 197–210, Berkeley, CA, USA, 2009. USENIX Association.

[5] Cezary Dubnicki, Cristian Ungureanu, and Wojciech Kilian. FPN: A Distributed Hash Table for Commercial Applications. In *Proceedings of the Thirteenth International Symposium on High-Performance Distributed Computing (HPDC-13 2004)*, pages 120–128, Honolulu, Hawaii, June 2004.

[6] Donald E. Eastlake and Paul E. Jones. US Secure Hash Algorithm 1 (SHA1). RFC 3174 (Informational), September 2001.

[7] John Gantz and David Reinse. The Digital Universe Decade - Are You Ready? *IEEE Trans. Parallel Distrib. Syst.*, May 2010. Sponsored by EMC Corporation.

[8] John Hantz, Christopher Chute, Alex Manfredi, Stephen Minton, David Reinsel, Wolfgang Schlichting, and Anna Toncheva. The diverse and exploding digital universe: an updated forecast of worldwide information growth through 2011. In *An IDC White Paper sponsored by EMC*, March 2008.

[9] P. Maymounkov and D. Mazieres. Kademlia: A peer-to-peer information system based on the xor metric. In *In Proceedings of IPTPS02*, Cambridge, USA, March 2002.

[10] European Parliament. Directive 2006/24/EC "On the retention of data generated or processed in connection with the provision of publicly available electronic communications services or of public communication networks". March 2006.

[11] David A. Patterson, Garth Gibson, and Randy H. Katz. A case for redundant arrays of inexpensive disks (RAID). In *Proceedings of the 1988 ACM SIGMOD international conference on Management of data*, SIGMOD '88, pages 109–116, New York, NY, USA, 1988. ACM.

[12] W. Curtis Preston. Target deduplication appliance performance comparison. http://www.backupcentral.com/mr-backup-blog-mainmenu-47/13-mr-backup-blog/348-target-deduplication-appliance-performance-comparison.html, October 2010.

[13] Sean Quinlan and Sean Dorward. Venti: A new approach to archival storage. In *FAST '02: Proceedings of the Conference on File and Storage Technologies*, pages 89–101, Berkeley, CA, USA, 2002. USENIX Association.

[14] Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, and Scott Schenker. A scalable content-addressable network. In *SIGCOMM '01: Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 161–172, New York, NY, USA, 2001. ACM.

[15] SNIA. *Common RAID Disk Data Format Specication*. SNIA Storage Networking Industry Association., March 2009. Version 2.0, Revision 19.

[16] Benjamin Zhu, Kai Li, and Hugo Patterson. Avoiding the disk bottleneck in the Data Domain deduplication file system. In *FAST'08: Proceedings of the 6th USENIX Conference on File and Storage Technologies*, pages 1–14, Berkeley, CA, USA, 2008. USENIX Association.