Using String Matching for Deep Packet Inspection

pp. 23-28

Po-Ching Lin, Ying-Dar Lin, Tsern-Huei Lee, and Yuan-Cheng Lai

String matching has recently proven useful for deep packet inspection (DPI) to detect intrusions, scan for viruses, and filter Internet content. However, the algorithm must still overcome some hurdles, including becoming efficient at multigigabit processing speeds and scaling to handle large volumes of signatures.

Before 2001, researchers in packet processing were most interested in *longest-prefix matching* in the routing table on Internet routers and *multifield packet classification* in the packet header for firewalls and quality-of-service applications. However, DPI for various signatures is now of greater interest.

Quantitative Retrieval of Geophysical Parameters Using Satellite Data

pp. 33-40

Yong Xue, Wei Wan, Yingjie Li, Jie Guang, Linyian Bai, Ying Wang, and Jianwen Ai

A reliable atmospheric remote-sensing monitor uses physical or statistical models for which the parameters must be retrieved quantitatively. However, such retrieval is data-intensive. High-resolution, wide-range, and long-duration observations produce several terabytes of data each day.

Processing such massive volumes of data into scientific aerosol products involves addressing several computational problems. Processing Level 1B data for Level 2 aerosol products for a single day requires 13 Gbytes total to achieve full coverage of China's main land surface.

Accelerating Real-Time String Searching with Multicore Processors

pp. 42-50

Oreste Villa, Daniele Paolo Scarpazza, and Fabrizio Petrini

String-searching algorithms are at the core of search engines, intrusion detection systems, virus scanners, spam filters, and content-monitoring systems. Fast string-searching implementations traditionally have been based on specialized hardware like FPGAs and application-specific instruction-set processors, but the advent of multicore architectures such as IBM's Cell Broadband Engine is adding new players to the game.

The authors developed a parallelization strategy for the Aho-Corasick algorithm that achieves performance comparable to other results in the literature with small data dictionaries but exploits the Cell's sophisticated memory subsystem to effectively handle large dictionaries.

Analysis and Semantic Querying in Large Biomedical Image Datasets

pp. 52-59

Vijay S. Kumar, Sivaramakrishnan Narayanan, Tahsin Kurc, Jun Kong, Metin N. Gurcan, and Joel Saltz

Digital microscopy opens new opportunities to study a disease's tissue characteristics at the cellular level. Traditionally, human experts visually examine tissue and classify images, then make a diagnosis. This process is time-consuming and the sheer size of image datasets makes gleaning information from digital microscopy slides data-intensive.

The authors' work addresses problems in two areas: the processing of large digitized slides for analysis and the semantic query of annotated images and image regions in a large image dataset.

Hardware Technologies for High-Performance Data-Intensive Computing

pp. 60-68

Maya Gokhale, Jonathan Cohen, Andy Yoo, W. Marcus Miller, Arpith Jacob, Craig Ulmer, and Roger Pearce

Data-intensive problems challenge conventional computing architectures with demanding CPU, memory, and I/O requirements. Using benchmarks that draw on three data types—scientific imagery, unstructured text, and semantic graphs representing networks of relationships—the authors demonstrate that emerging hardware technologies to augment traditional microprocessor-based computing systems can deliver 2 to 17 times the performance of general-purpose computers on a wide range of data-intensive applications by increasing compute cycles and bandwidth and reducing latency.

ProDA: An End-to-End Wavelet-Based OLAP System for Massive Datasets

pp. 69-77

Cyrus Shahabi, Mehrdad Jahangiri, and Farnoush Banaei-Kashani

By design, developers optimize traditional databases for transactional rather than analytical query processing. These databases support only a few basic analytical queries with nonoptimal performance and, therefore, provide inappropriate tools for analyzing massive datasets.

Online analytical processing tools have emerged to address the limitations of traditional databases and spreadsheet applications. OLAP tools support complex analytical queries and handle massive datasets. The authors' ProDA system uses these tools to enable exploratory analysis of massive multidimensional datasets.