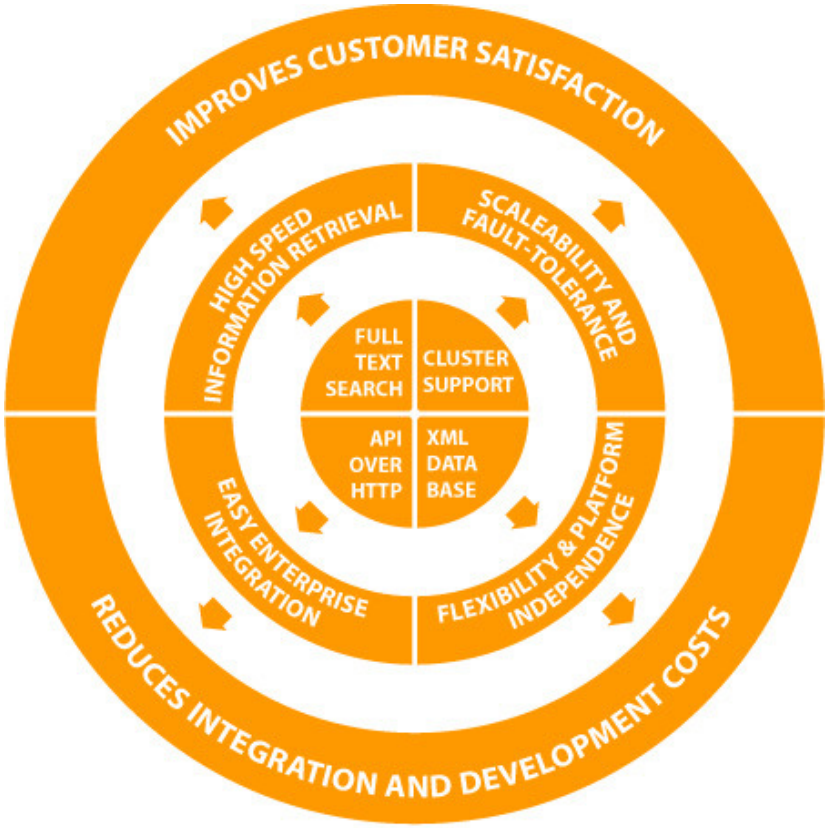
Siets Server - Architecture, Vision & Performance

www.siets.net

Riga, 2005

What is Siets - Architecture & Vision



Platform for building high speed full text search engines

Innovative architecture offers multiple improvements in how companies deal with large volumes of data at a fraction of the cost of competitors.

What does 'SIETS' mean?

- In Latvian 'siets' means 'sieve'- an old tool for sifting grains out of raw crops.

- **Siets Server** can sift through massive data sets in a second and retrieve all relevant information.

Main Competitive Features

- Flexible relevancy definition
- Enterprise level security
- Input/output flexibility through XML messaging
- Scales to terabytes of data in a cluster
- Stores any multi-language data (UTF-8)
- API over HTTP (Web services support)
- High indexing and search speeds
- Full-text, boolean, numeric search (>50 options)
- Centralized Web management

Figure 1a: Siets Server can be used to build scalable search engines in cluster configuration.

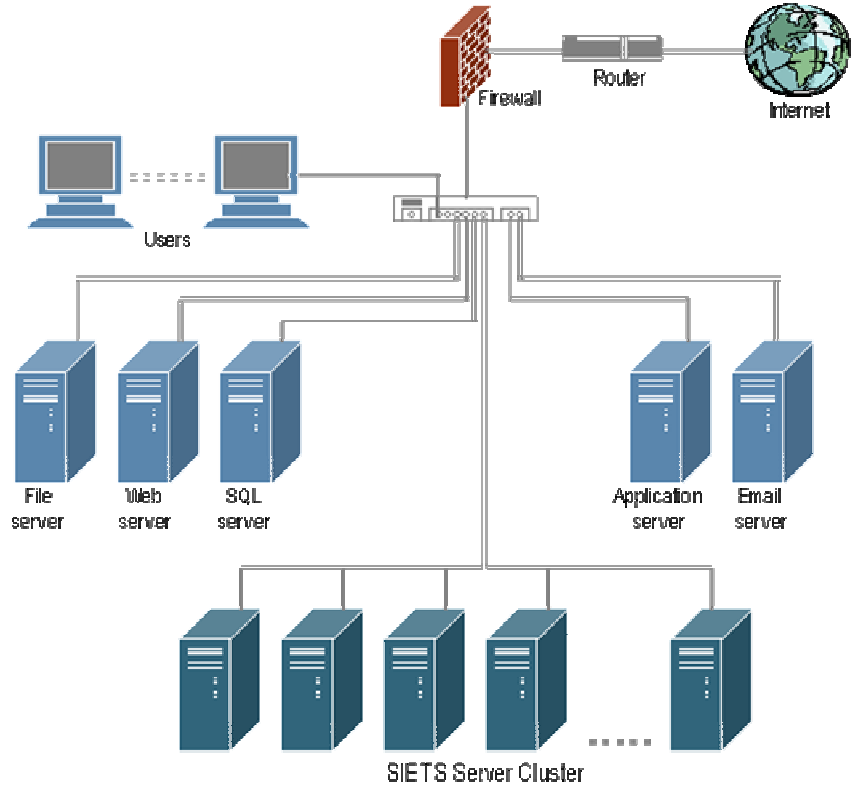
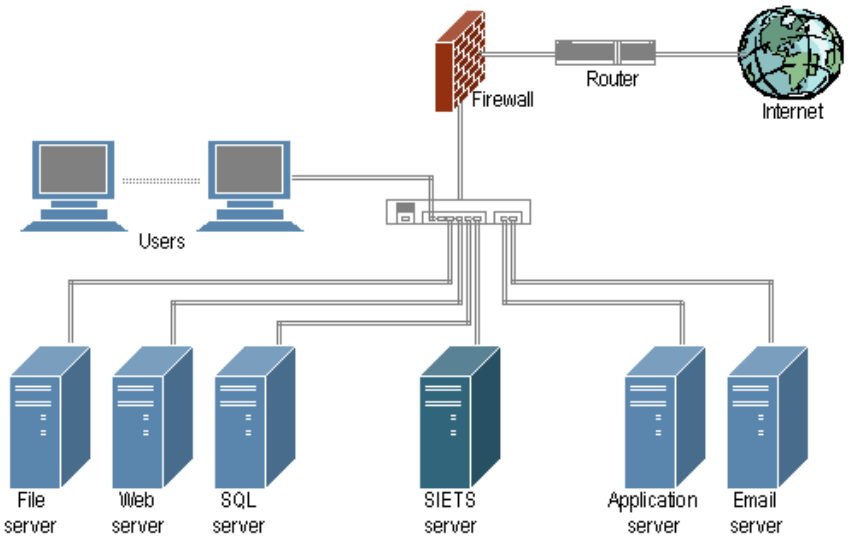


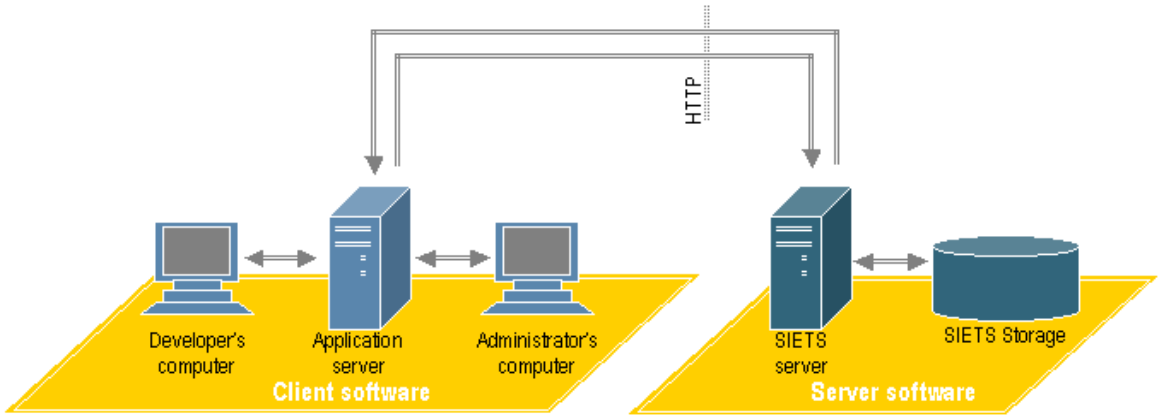
Figure 1b: Siets Server can be used in a corporate network as dedicated search appliance/server.



What are Siets technology advantages?

- **Fast Search** - one of the fastest search engines on the market (1500 full text queries per minute with disk access, up to 15,000 memory-cached).
- **Generic Scalability** – stores any XML content as distributed database on hundreds of servers and instantly combines most relevant search results.
- **Original Source Code in C** – search and indexing engine source code was developed in C and C++ for maximum performance.
- **Smart Use of Memory** - algorithms use all available RAM memory for data storage and caching minimizing mechanical disk access.
- **Sequential Indexing** - index reads/writes are performed sequentially instead of random disk access singinificantly improving data access times.
- **True Real Time Updates** - intelligent memory caching for updates provides real time modification for index, which is written to disk as needed.
- **Simplicity and Efficiency** - increase customer ROI and profitability through ease of depolyment of XML and Web standards-based Siets software.

Figure 2: Siets software is based on proven and robust client-server architecture.



Enterprise Level Quality

Four main qualities for enterprise search solutions:

- Search speed
- Indexing speed
- Real-time index modification
- Reliability

Siets Server was designed to achieve all of those four qualities into a single platform software.

Figure 3: Siets Server uses simple XML messaging (Siets API) and http protocol.

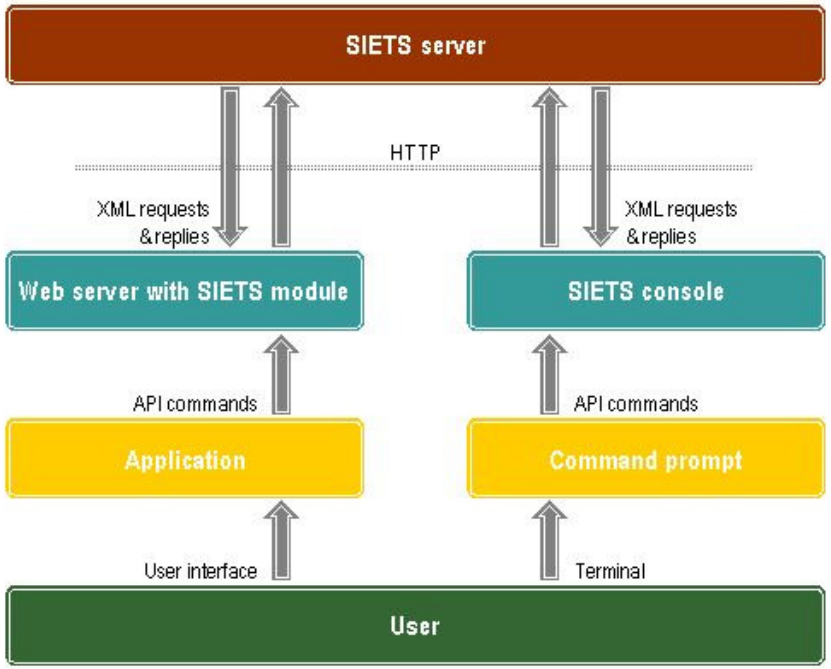
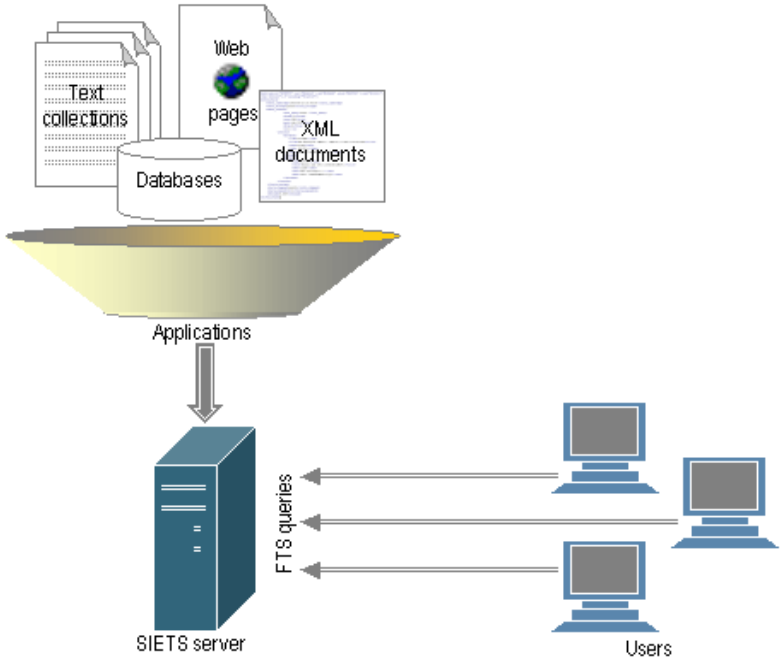


Figure 4: Users can use high performance full text search queries for finding any data.

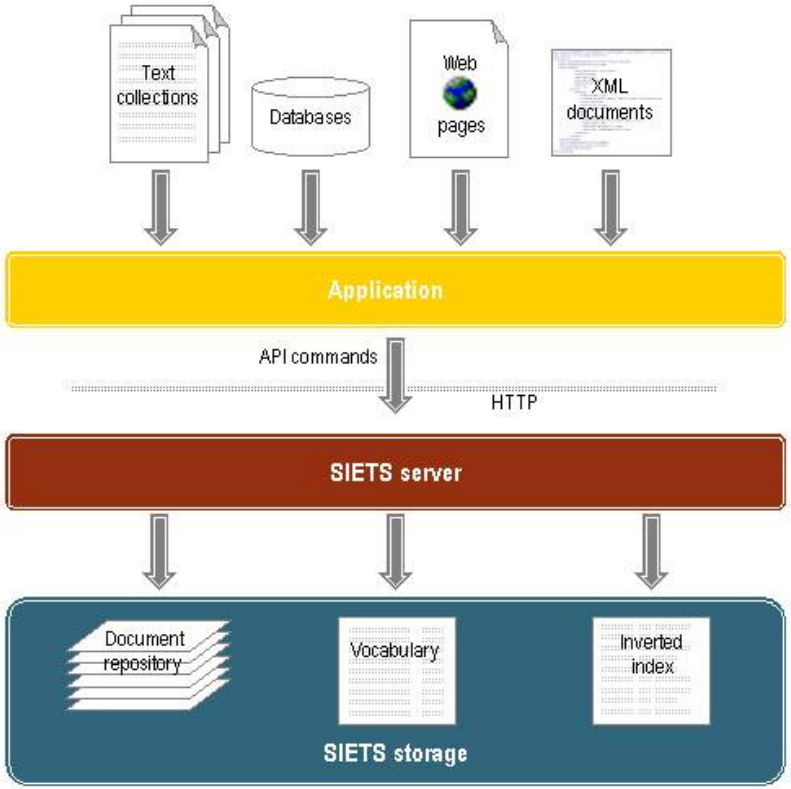


More Than 50 Search & Indexing Options

- sub-second full text search in giant databases
- word stemming, wildcards, synonym support
- multi-level Boolean (AND, OR, NOT) and phrase search
- spell check suggestions using fuzzy logic
- sorting by relevance, rating or date
- Web friendly search results, incl. text snippets and highlighting
- document search by similar content
- domain grouping, ingnored text support
- fielded search, proximity search, numeric range search
- automatic indexing of XML meta-markup
- multi-level XML-drill down for meta data
- definition of custom relevance rules for parts of documents
- location search with distance sorting etc.

For a full list of all options please see Features section on www.siets.net.

Figure 5: Siets Server automatically builds inverted index for all documents.



Developer's Benefits

- Easily integrates into any corporate environments and programming languages through XML based API messaging and firewall friendly HTTP protocol
- Siets API is more simple than Web services. Does not require document definition schemas.
- Stores multilingual data in the same storage using UTF-8 standard. Supports > 160 languages under Linux.
- Detailed Siets API developer's documentation and sample code in C, PHP, Perl, Java, .NET languages.
- Full version software download for evaluation

Figure 6: All data items are stored / retrieved as simple XML documents from Siets Server.

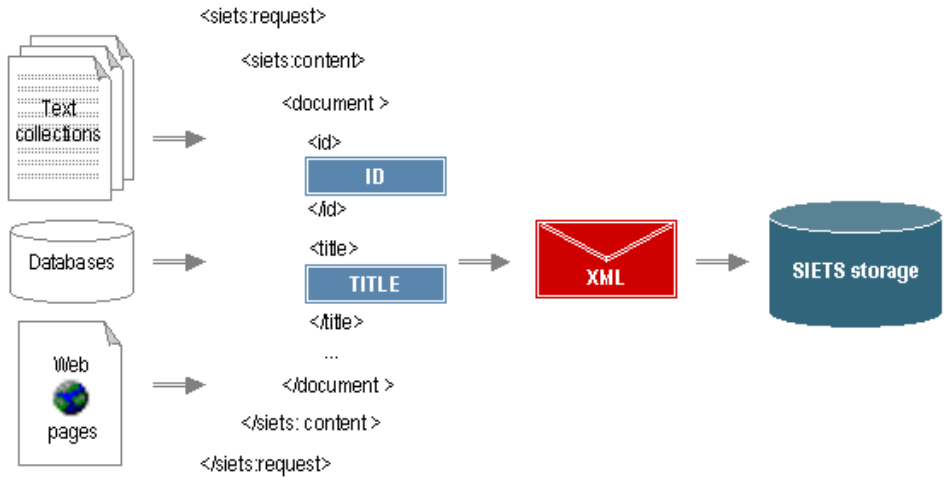


Figure 7: Siets Server can operate multiple document collections per one computer.

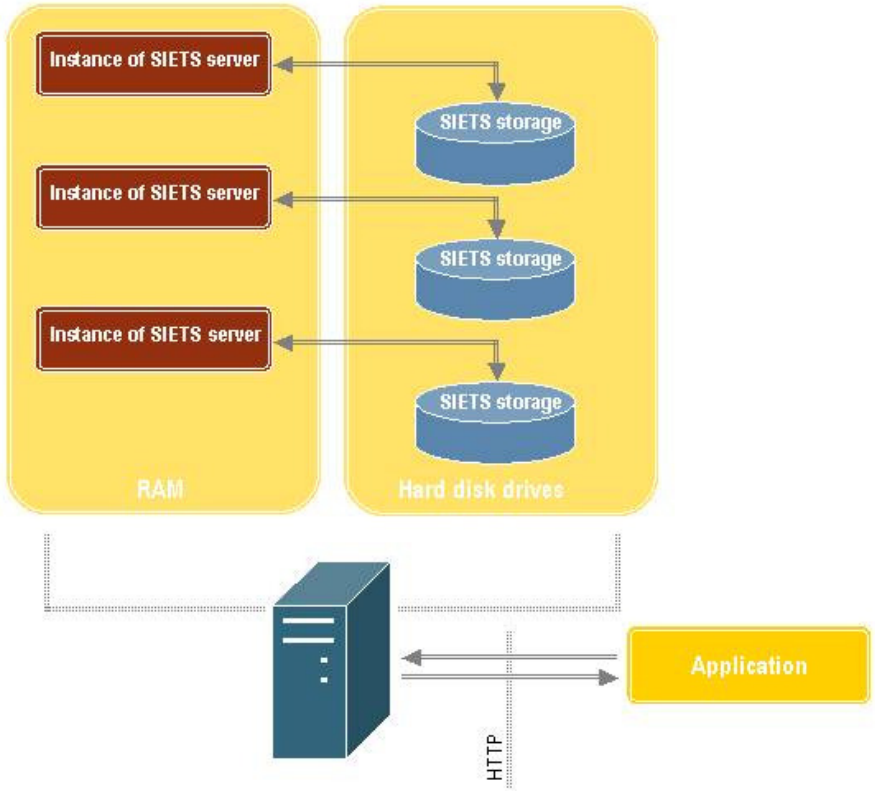


Figure 8: Siets Server can be used in cluster configurations for search in very large data sets.

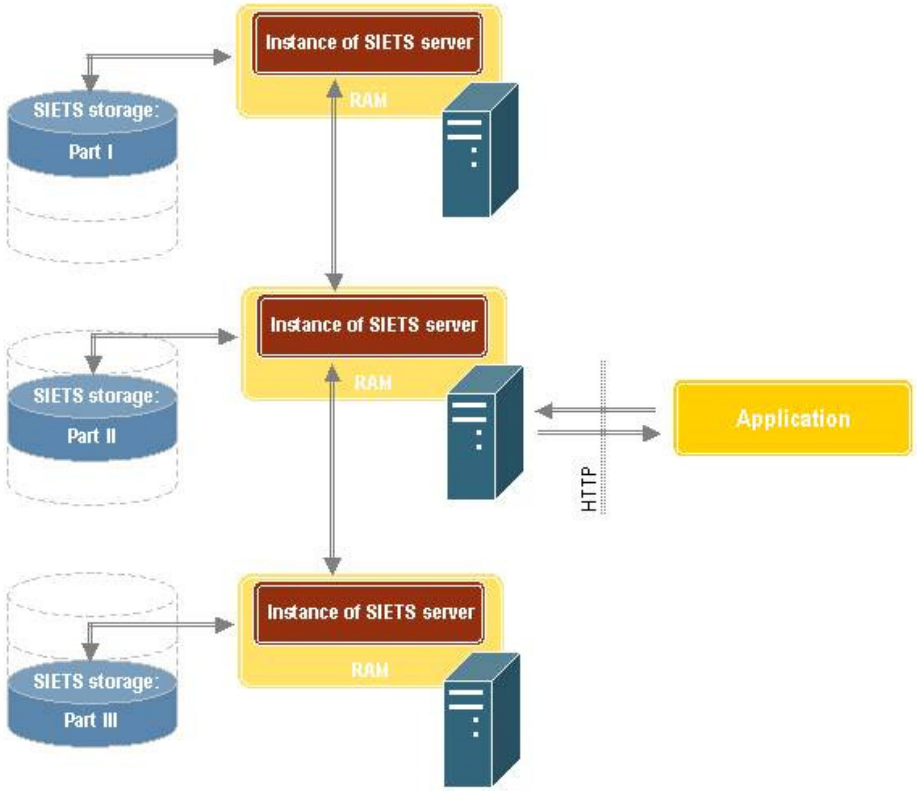


Figure 9: Siets stores documents in different languages and encodings in UTF-8 (more than 160 language encodings).

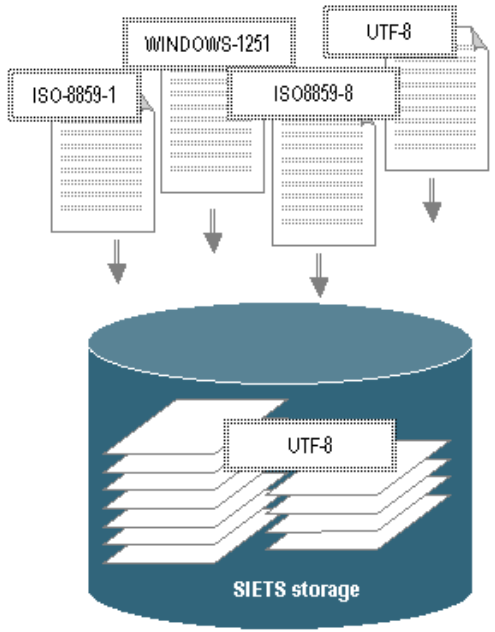


Figure 10: Storing documents in different encodings and return of results in UTF-8.

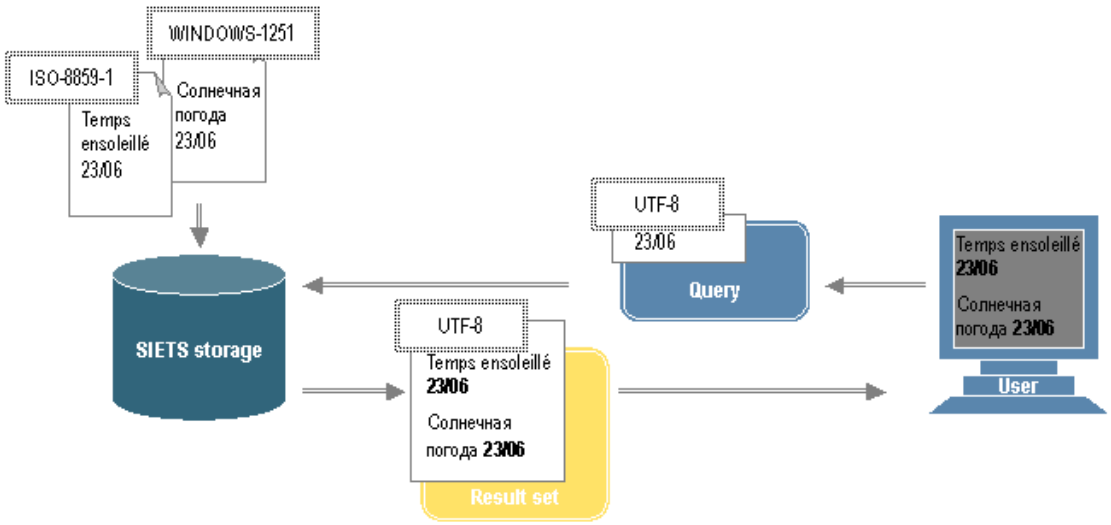
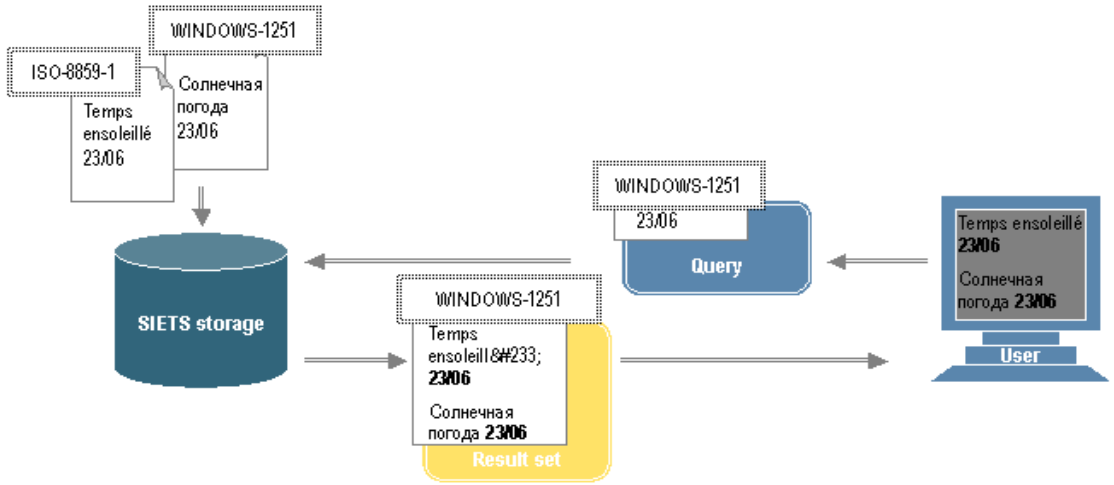


Figure 11: Storing documents in different encodings and return of search results in one specified encoding.



12: Applications can send and receive XML messages for indexing/querying directly to Siets Server - no client software needed.

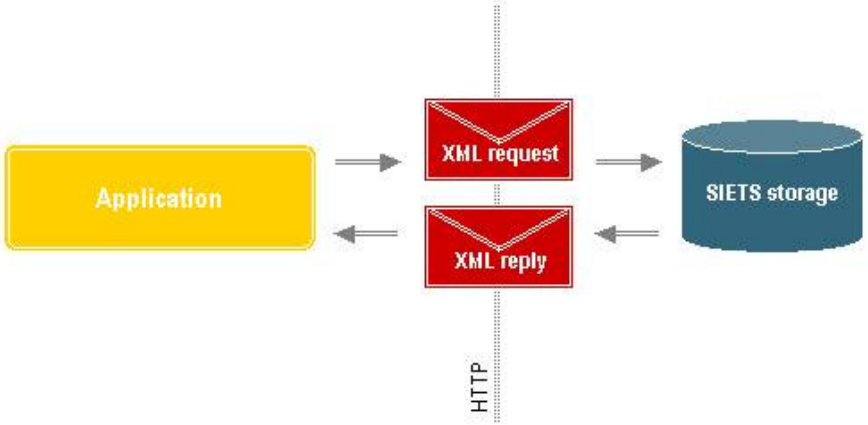


Figure 13: Applications can submit parameters and receive XML / XSLT formatted replies.

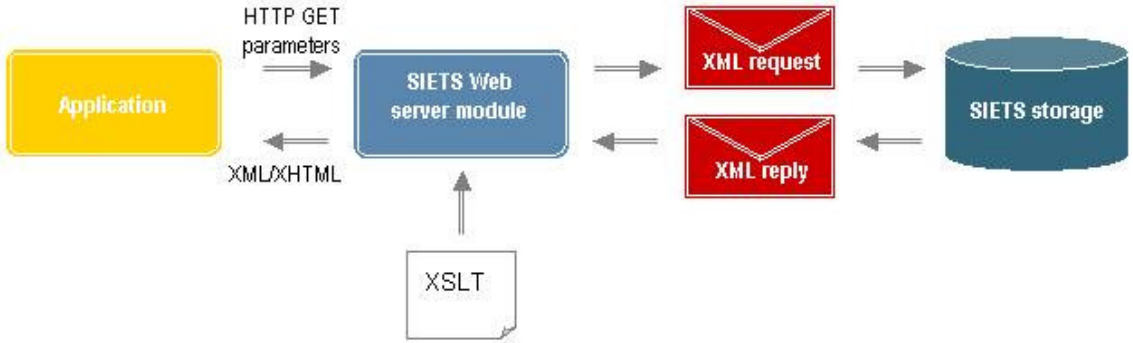


Figure 14: Typical interaction model between User application and Siets Server using Web services.

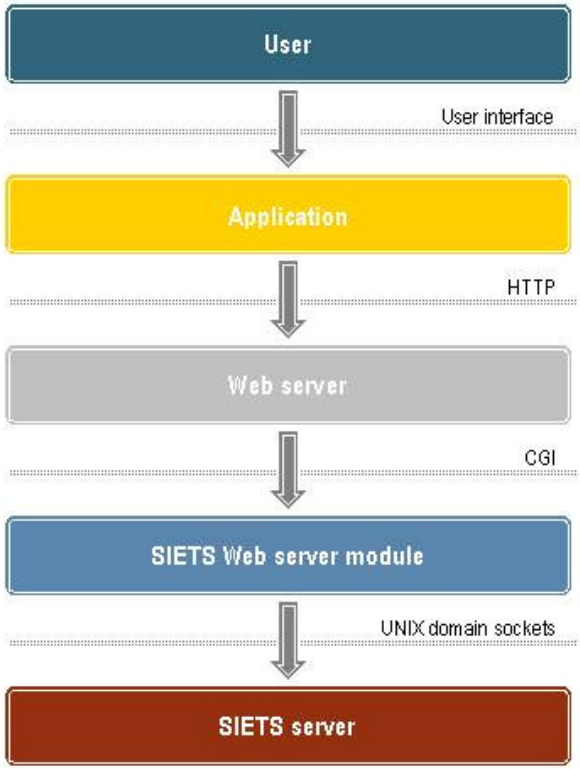


Figure 15: Every searchable document collection has its own Siets server instance running the Control demon, the Document Repository demon, the Vocabulary demon and the Inverted Index demon.

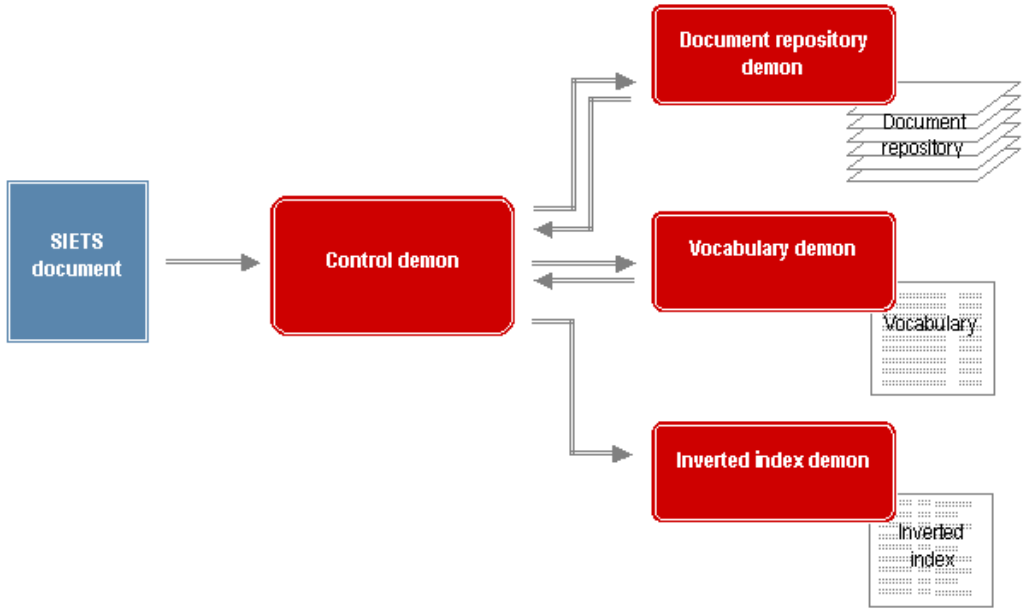
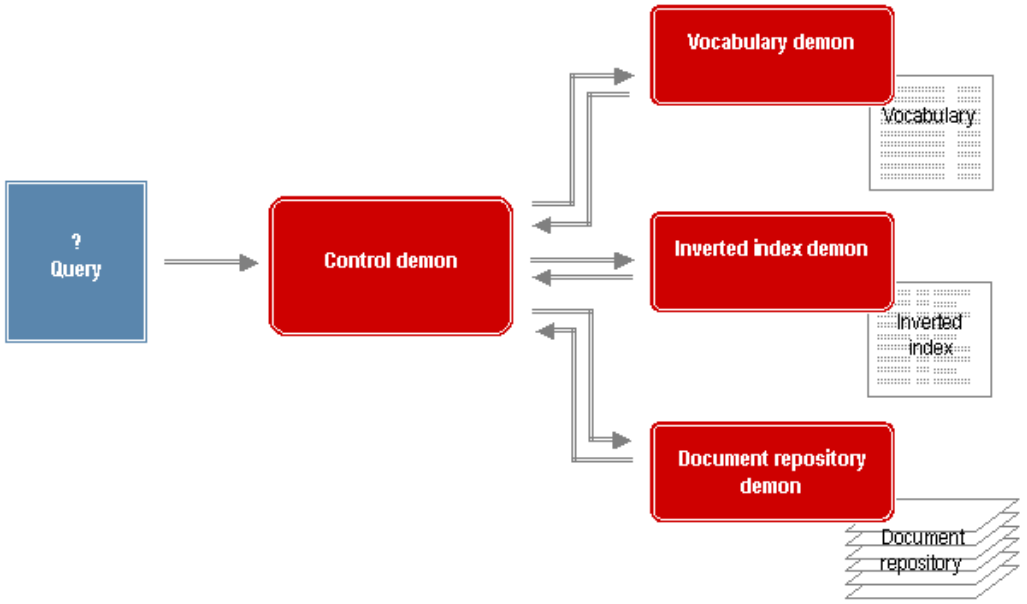


Figure 16: All querying and indexing tasks in a Siets storage are separated among different operating system processes for better workload sharing and parallel indexing while querying.



Management Benefits

- Centralized management of all Siets servers through Web-based Siets Enterprise Manager.
- User-friendly installation SETUP file and bootable ISO CD image file available for download to start a search engine without deep Linux knowledge and skills.
- Flexible licensing policy based on per-server pricing and no need to obtain client software licenses
- Fast and professional technical support from me as an IT professional with > 12 years experience in computing
- Integrated crawler utility included for scheduled indexing of data from corporate Web sites and file systems.

Figure 17: Siets Enterprise Manager is a Web server application for centralized management of all Siets server instances (searchable document collections) accross the corporate network.

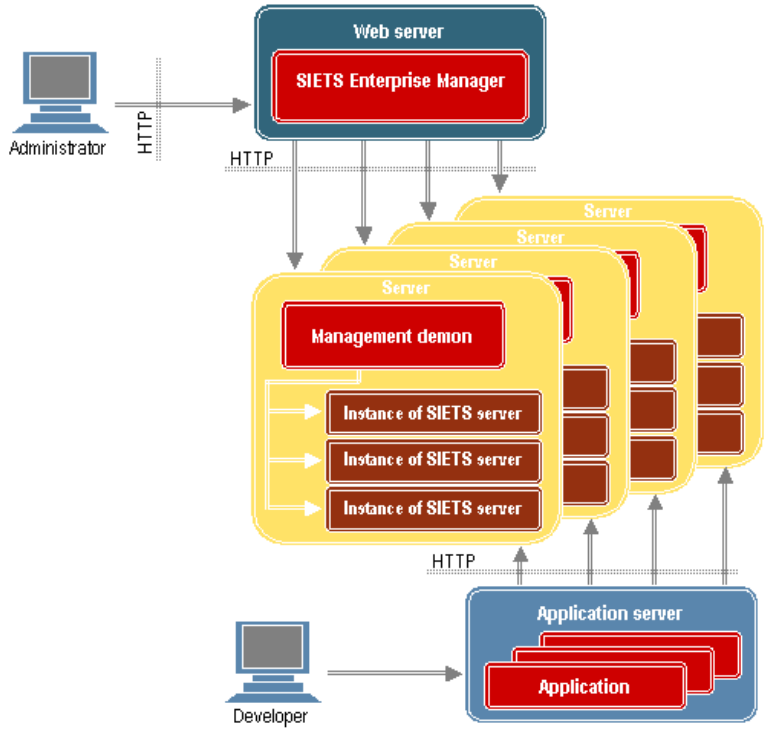
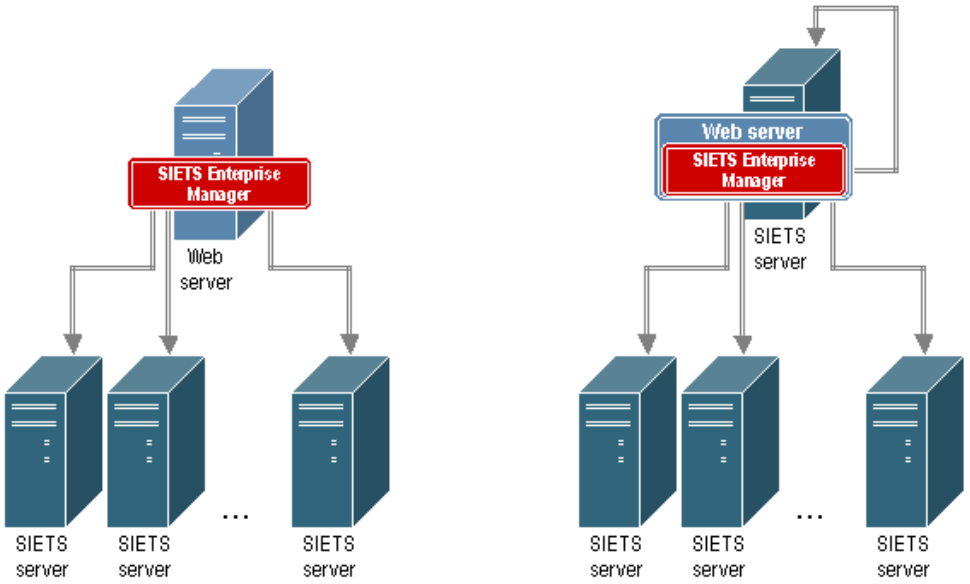


Figure 18: Siets Enterprise Manager can be installed on a separate Web server computer or on a Siets Server computer.



Siets Engine - It's All About Performance

- Siets Server engine is entirely written in C and C++ for the maximum search performance.
- Optimized for speed by fast algorithms using all available computer memory for data cache.
- Sequential disk access vs random access is used for fast index updates and reads.
- Built in data integrity controls to protect against unexpected server shutdowns.

Figure 19: Indexing of large text collections on a single server computer can take many hours.

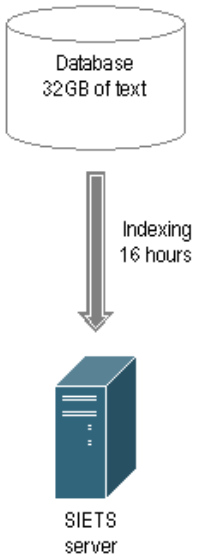


Figure 20: Siets Server indexing performance can be scaled lineary using multiple cluster computers to reduce indexing times for large document collections.

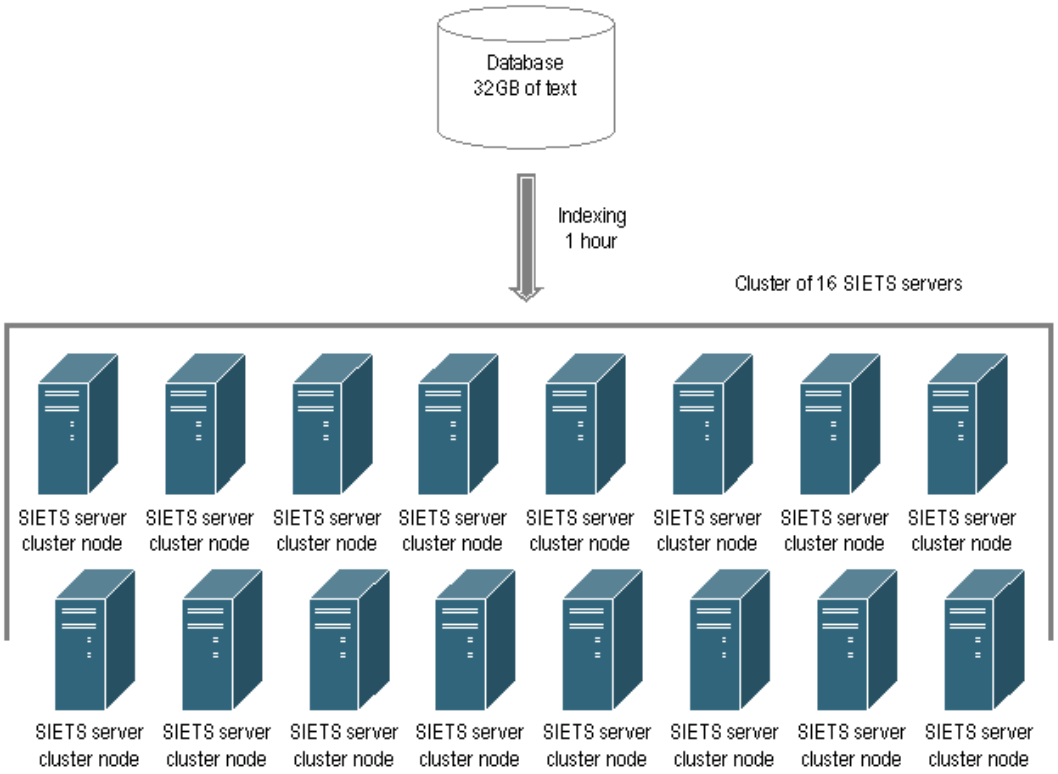


Figure 21: Test results of Siets Server search speeds on different Linux hardware (number of search queries per minute).

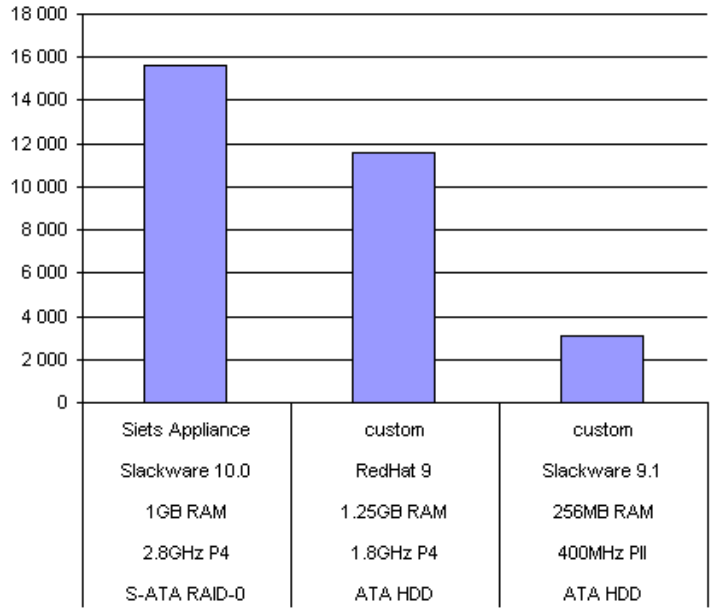


Figure 22: Tests results of Siets Server search speeds on document collections of different sizes (number of search queries per minute).

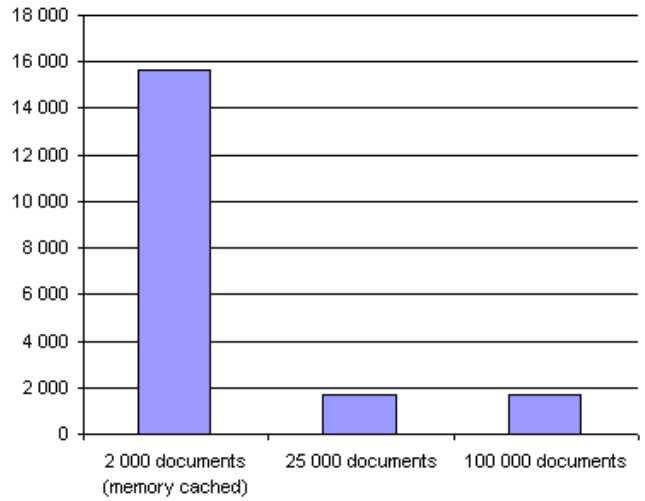


Figure 23: Test results of Siets Server search speeds on number of search terms in a query for 25,000 documents (number of search queries per minute).

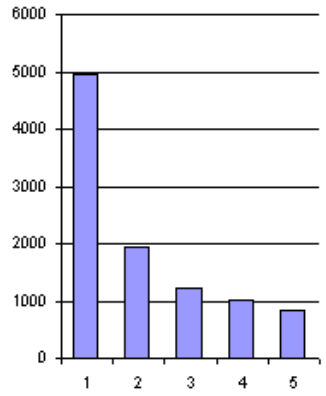


Figure 24: Test results of Siets Server search speeds on number of search terms in a query for 5,000 memory cached documents (number of search queries per minute).

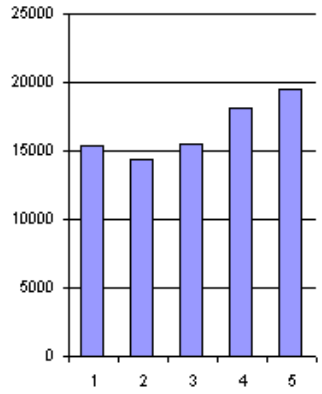
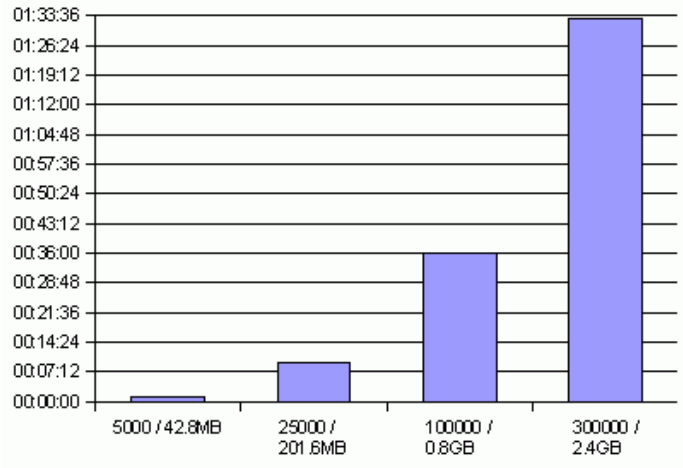


Figure 25. Test results of Siets Server indexing speed on document collections of different number of documents and different sizes (total loading and indexing time in hours:minutes).

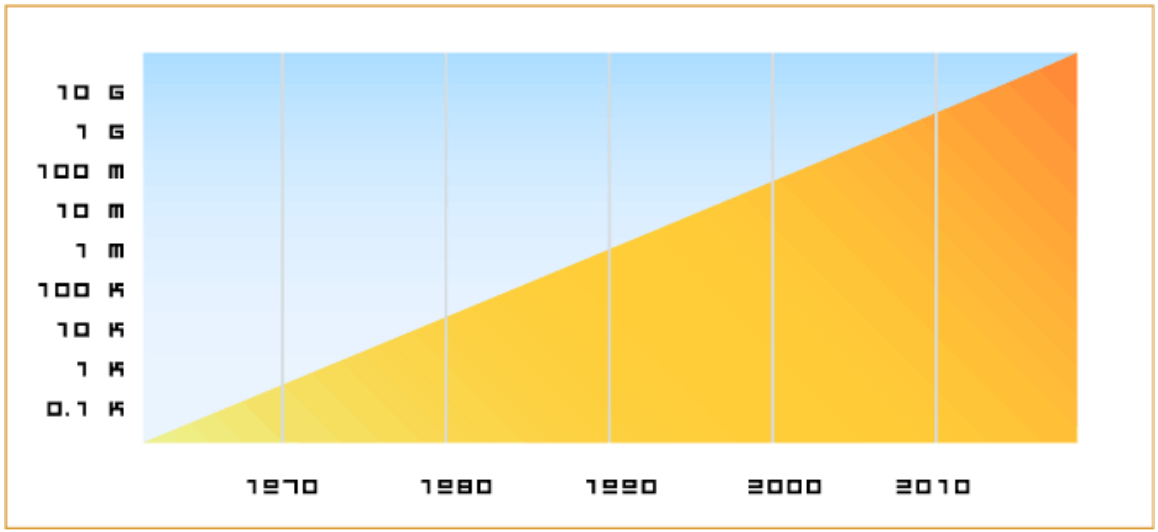


Testing Methodology

- All test documents were of different content.
- Average size of documents was 10-20 Kilobytes.
- Documents contained text with randomly selected words distributed according to their natural language frequency in normal text.
- There were no special performance tuning done at the Linux operating system level. All system configuration parameters during tests were left as default.

Figure 26: Moore's Law - Siets Server ally

- Moore's Law 60% / year Memory -- 4 x size every 3 Years



- Siets Server engine's core software is designed to accomodate future data processing applications.

Siets As a Platform for Internet Search

- **Siets Server** can be a perfect platform for building scalable Internet search engines.
- See Siets Server handling of country-wide Internet search on **www.siets.lv/en**.
- Deploy **Siets Global Internet Crawler** to collect Web data country-wide or globally.

Contact Data - Give Us a Call

Mr. Gints Ernestsons
Riga, Latvia
European Union (New Europe)
Phone: +371-2-9479905
Email: info@siets.net
Web: www.siets.net