

## Performance Guide

Splunk 2.1 is the highest performance technology for indexing, searching and managing logs and IT data. It delivers higher indexing throughput, faster search speeds and denser storage than previous Splunk releases and 3-5 times the performance of other log management technologies and appliances.

If you are a Splunk customer, partner, service provider or developer you'll want to understand indexing performance and storage requirements in the early stages of planning a Splunk deployment, integration or new module development. This paper summarizes recent Splunk performance test results and explains the factors that impact index throughput and storage utilization. It also unravels the confusing performance information given for other types of log management technologies.

### Performance Test Design

#### Data Set

- **Source:** 33GB from the InteropNet network at the Interop trade show in Las Vegas May 2006.
- **Source types:** Standard single-line syslog from a mix of Extreme routers, Juniper Netscreen firewalls, Aruba wireless infrastructure and other common network data sources.
- **Record Size:** Average record size 347 bytes.

Splunk's performance tests use syslog data to facilitate the most direct possible comparison with other log technologies that often only support syslog and a few other network or security log formats. Our reference syslog sample is 33 GB of data captured at the Interop show network in May 2006.

This dataset has a much larger average byte size than is utilized in the log management industry when quoting throughput in events per second. The industry assumption is 150 bytes per record and our data set has an average of 347 bytes per record. As the throughput delivered by Splunk (and other log management technology) is actually more constrained by data size than number of records, we use megabytes per second (mbps) as a more relevant primary performance metric.

#### Platform

- **CPU:** 2 Dual Core Intel Xeon 3.0 GHz Processors
- **RAM:** 4 GB
- **OS:** Red Hat Enterprise Linux 4

Our reference platform is a two dual-core CPU, 4 GB RAM server. Splunk is able to use one or all four virtual CPU's (cores) for data processing and indexing.

#### Baseline Configuration

Our baseline test uses Splunk's universal (default) processing and indexing. This configuration provides advanced automated data processing and indexing – including automated timestamp recognition, source and event typing and source host recognition. Most important, Splunk delivers real-time search performance on anything in the original data.

Other log technologies and log management appliances either don't index anything in the raw log events and rely on slow full text scanning or index only a few specific fields as keys in a relational database.

## Alternate Configurations

Splunk's unique level of universal processing and indexing makes our baseline test results hard to compare to other log management technologies that do far less with the data. In addition, customers sometimes want to meet log data retention requirements for data that they don't need to search as regularly. Splunk supports alternate processing and indexing configurations to tune down index density and turn off certain functions. Customers may want to evaluate mixing lower and higher density indexing for different classes of log data with different policies and users.

### Indexing Density Configurations

Configuration	Full Density <sup>1</sup>	High Density <sup>1</sup>	Medium Density <sup>1</sup>	Low Density <sup>2</sup>	Minimal Density <sup>3</sup>
Meta Data (time, source, host source type, event type, event relationships)	•	•	•	•	•
Event Typing	•				
Automatic Timestamp Recognition	•	•			
Major Segments	•	•	•	•	•
Minor Segments	•	•	•		
Automatic Source Host Recognition	•	•	•	•	

<sup>1</sup> This scenario is not comparable to any other log data technology or log appliance.

<sup>2</sup> Recommended for deployments with low frequency ad hoc investigation requirements. This scenario is most directly comparable to log appliances that use commodity document indexing technologies to provide basic text indexing.

<sup>3</sup> Recommended for deployments with very infrequent ad hoc data retrieval. This scenario is most directly comparable to log storage appliances, but Splunk still delivers faster search and better performance.

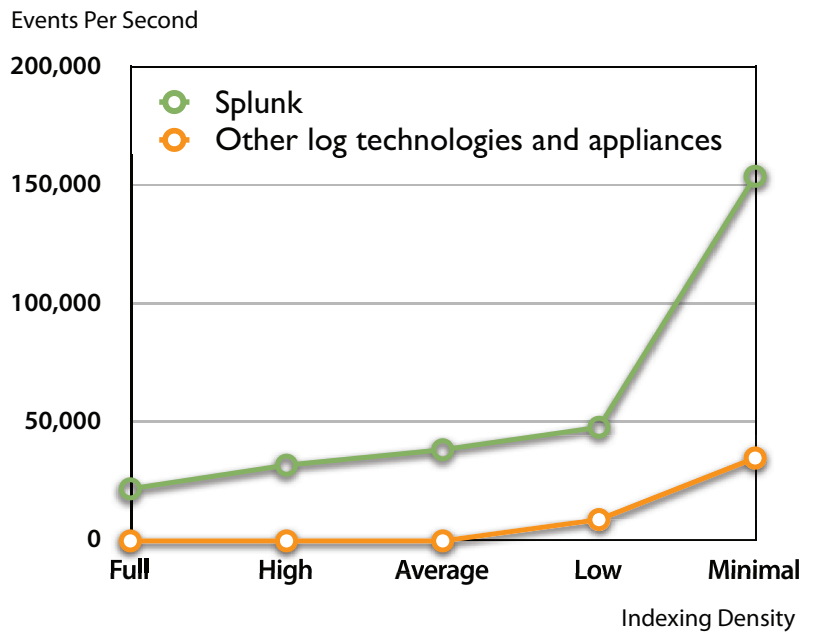
## Test Results

Test results show that Splunk achieves 3.3 mbps throughput and consumes only 40% of the raw data size with all of its advanced processing and high density indexing enabled.

Turning off some of Splunk's advanced features like automatic timestamp recognition and event typing has a moderate impact on performance and a negligible impact on storage requirements.

Lowering the density of indexing to just indexing metadata like timestamp, host, source and source type delivers a performance boost to a stunning 154,000 events per second, while squeezing storage requirements down to just 12% of the raw data size.

This makes Splunk the highest performance choice for simple log retention when compared to log appliances that just store the data organized by time with no indexing at all, yet typically deliver only 20-50,000 events per second throughput.



### Performance Comparisons

Configuration	Throughput	Storage as a % of raw data <sup>1</sup>	Events per second @ 150 bytes/event	Other log technologies and appliances
Full Density	3.3 mbps	40%	22,000 eps	Not possible.
High Density	4.6 mbps	40%	32,000 eps	Not possible.
Medium Density	5.5 mbps	40%	38,500 eps	Not possible.
Low Density	6.9 mbps	30%	48,000 eps	8,000-10,000 eps
Minimal Density	22.0 mbps	12%	154,000 eps	20,000-50,000 eps

<sup>1</sup> Network syslog data tends to get better compression than application logs, email delivery agent logs, and many other more entropic data sources. Therefore caution should be used when projecting storage requirements for other classes of data. Splunk recommends allocating more storage for a more diverse data set.