WebSearchBench: Ein Open Source Baukasten für Internet Suchmaschinen der nächsten Generation

Christoph Lindemann

Universität Dortmund
Informatik IV
-Rechnersysteme und Leistungsbewertung-
August-Schmidt-Str. 12
44227 Dortmund
http://www4.cs.uni-dortmund.de/~Lindemann

Dieser Vortrag basiert auf gemeinsamen Arbeiten mit M. Lohmann, O. Waldhorst, Ch. Welz und M. Wintergerst

Gliederung

- State-of-the-Art in Suchmaschinentechnologie
  - Motivation für Internet Suchmaschinen
  - Komponenten einer Suchmaschine der 2. Generation
- Aufbau und Funktionsweise einer Internet Suchmaschine der zweiten Generation
  - Interaktion der Softwarekomponenten
  - Vernetzte PCs als zugrundeliegende Hardware Architektur
- Graphanalyse des World Wide Web
  - Relevanzbewertung mit PageRank und Zentren/Instanzen
  - Graphstruktur des Web
- Der Software Baukasten WebSearchBench
  - Skalierbare feinkörnig parallelisierte Software Architektur
  - Horizontale und vertikale Konfigurierung
  - PC Cluster mit Gigabit Verbindungsnetz als zugrundeliegende Hardware Architektur

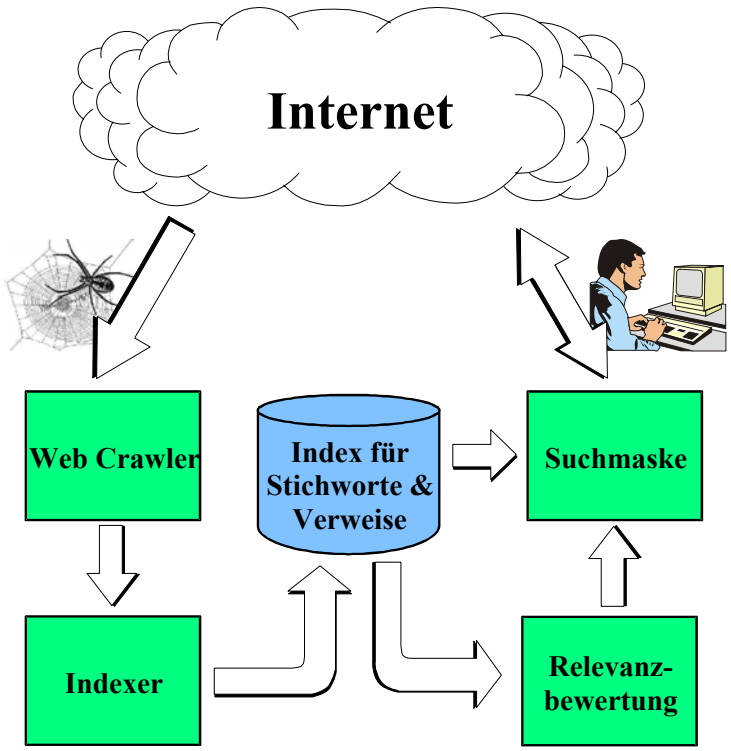
Motivation für Internet Suchmaschinen

- **World Wide Web hat sich zur Killer Applikation des Internet entwickelt**
  - Zu Beginn nur Informationsmedium für die akademische Welt
  - Vor ca. vier Jahren von der Wirtschaft als hervorragendes Marketing Instrument entdeckt (e-business, e-government,…)

- **Statische Entwicklung (Quelle: DENIC e.V.)**
  - 1992: ca. 30.000 Hosts in Domäne .de
  - 2002: 2.500.000 Hosts in Domäne .de

- **Aufgabe einer Internet Suchmaschine**
  - Software System zur Extraktion relevanter Information aus der mannigfaltigen Datenmenge des Web

Komponenten einer Internet Suchmaschine



Technische Herausforderungen

- Verwaltung sehr großer Datenmengen
  - Ca. 8 Milliarden Dokumente im öffentlichen Web
  - In 2000: 21% HTML, 72% Bilder, 5% Textapps, 1% Audio,Video
  - „Hidden Web“ umfasst mehr als 500 Milliarden Dokumente
- Herausforderung
  - Parallelisierung des Web Crawlers und des Indexers mit dem rasanten Wachstum des Web
- Relevanzbewertung der Suchergebnisse
  - Suchen in schwach strukturierten Datenmengen
  - Effektives Ordnen der Suchergebnisse
- Herausforderung
  - Extraktion von relevantem Wissen aus der Fülle verfügbarer Information

Aktuelle Forschungsarbeiten

- Hochgradig skalierbare Software Architekturen
  - PC Cluster mit Verbindungsnetz anstatt vernetzte Server
  - Feinkörnige Parallelisierung durch Ausnutzung von Programmparallelität anstatt nur Datenparallelität
- Graphbasiertes Information Retrieval im Web
  - Kartographie des Web und spezifischer Domänen z.B. .de
  - Bestimmung von Cyber Communities, Crawling Strategien,...

Vision: Going Beyond Client/Server

- Entwicklung von Informations- und Dokumentensuch-diensten für Peer-to-Peer Netze
- Entwicklung von Middleware Protokolle für Peer-to-Peer Netze mit mobilem Zugang

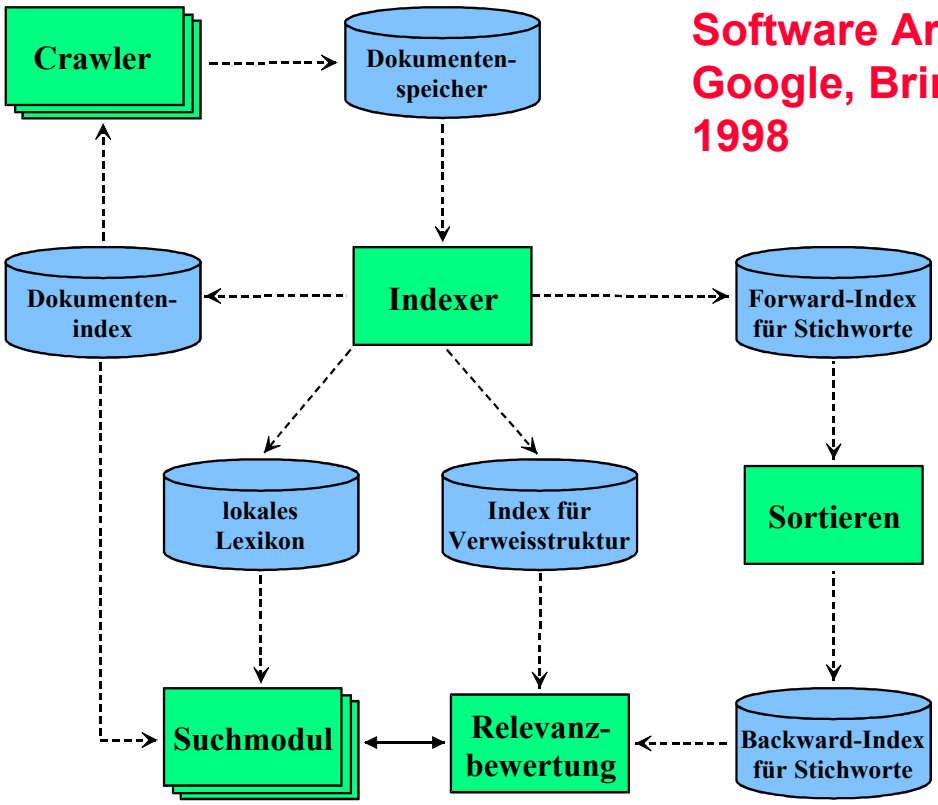
Weitere Forschungsthemen

- Semantic Web
  - Resource Description Framework (RDF)
  - Ontologien
- Maschinelle Lernverfahren für das World Wide Web
  - Automatische Kategorisierung von Web Seiten
  - Rechnergestützte Wissensextraktion aus Web Seiten
  - Probabilistische Modelle und Regeln zur Charakterisierung der Struktur des WWW
- Multimedia Information Retrieval
  - Indexierung von Bild, Audio und Video Dokumenten
  - Erweiterung derzeitiger Suchmaschinen um Spracherkennung und Mustererkennung für die multimediale Suche

State-of-the-Art nutzerorientierte Features

- Ermittlung hochrelevanter Suchergebnisse
  - Konfigurierbare Stoppwortliste und schwarze Listen
  - Finetuning durch statistische Auswertung der Suchanfragen
  - Fachspezifische Optimierung mit Thesauri
  - Betrachtung der Verweisstruktur des Web
  - Ausgabe der Relevanzmaße für die Suchergebnisse
- Differenzierte Formulierung für die Suchanfragen
  - UND/ODER Verknüpfung von Stichwörtern
  - Negationssuche (–)
  - Erzwingung der Berücksichtigung von Stoppwörtern (+)
  - Phrasensuche (“...“)
- Sehr kurze Antwortzeiten für Suchanfragen
- Einhaltung der Netiquette beim Web Crawling

Interaktion der Softwarekomponenten



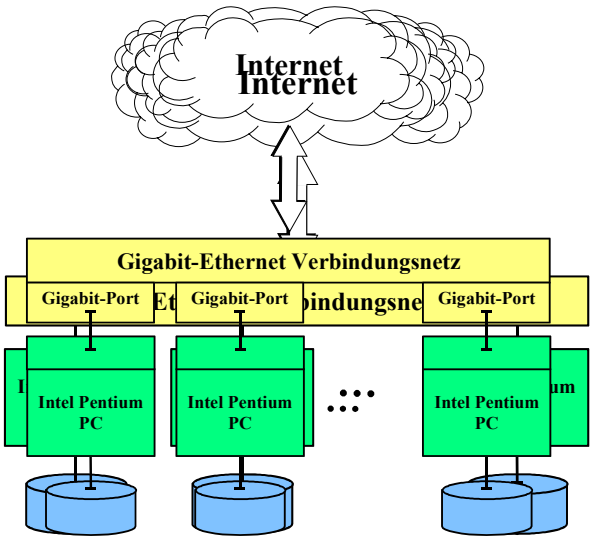
Software Architektur von Google, Brin and Page, 1998

Hardware Architektur

- **Cluster von Inktomi: 100 Knoten mit 200 CPUs, Brewer 2001**

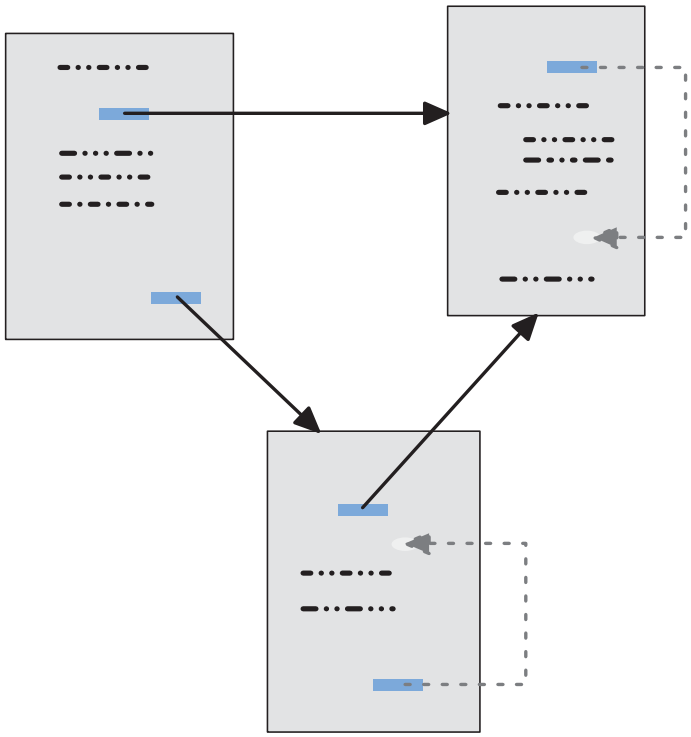
Hardware Architektur

- Google, Inktomi, FAST: PC Cluster mit Gigabit Technologie ?



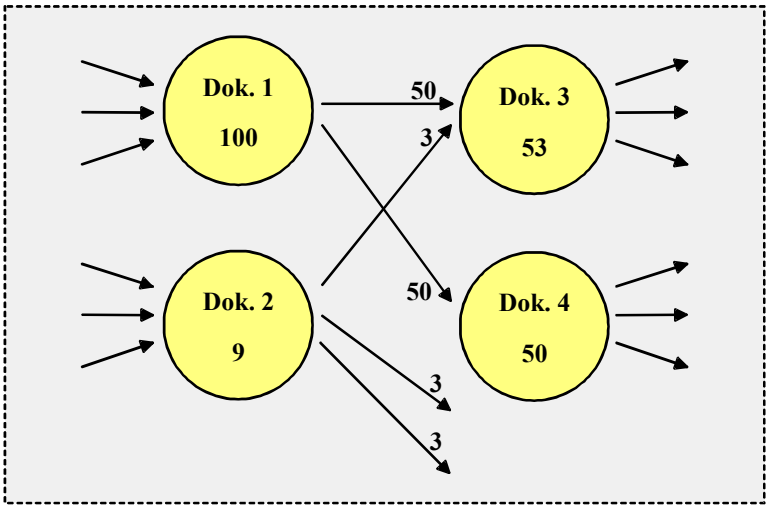
Veranschaulichung

- Modellierung des Web als gerichteter Graph



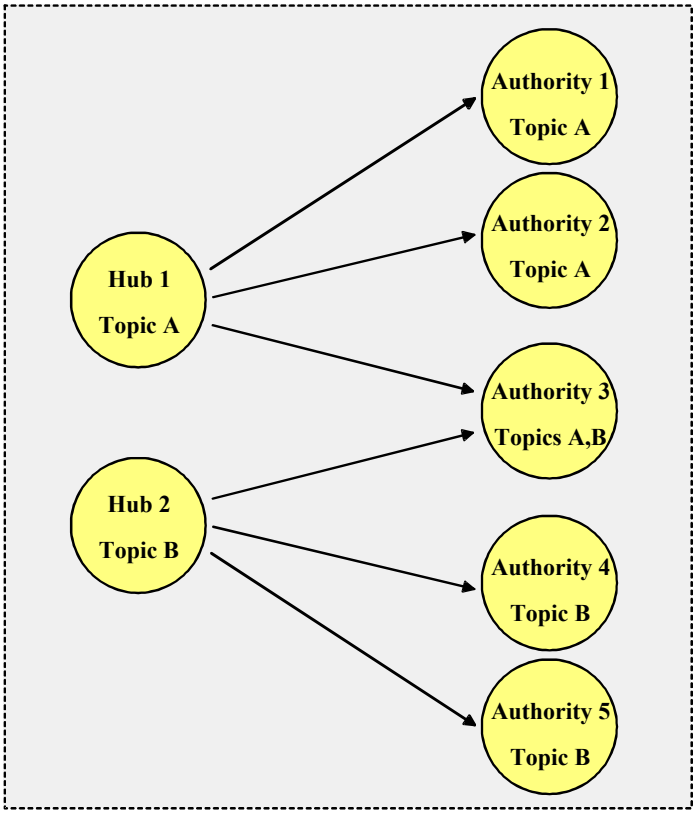
Relevanzbewertung mit PageRank

- Idee: nach allgemeiner Popularität gewichtete Bewertung der eingehenden Verweise (Page et al. 1998)
- Numerische Lösung eines sehr grossen, dünn besetzten linearen Gleichungssystems



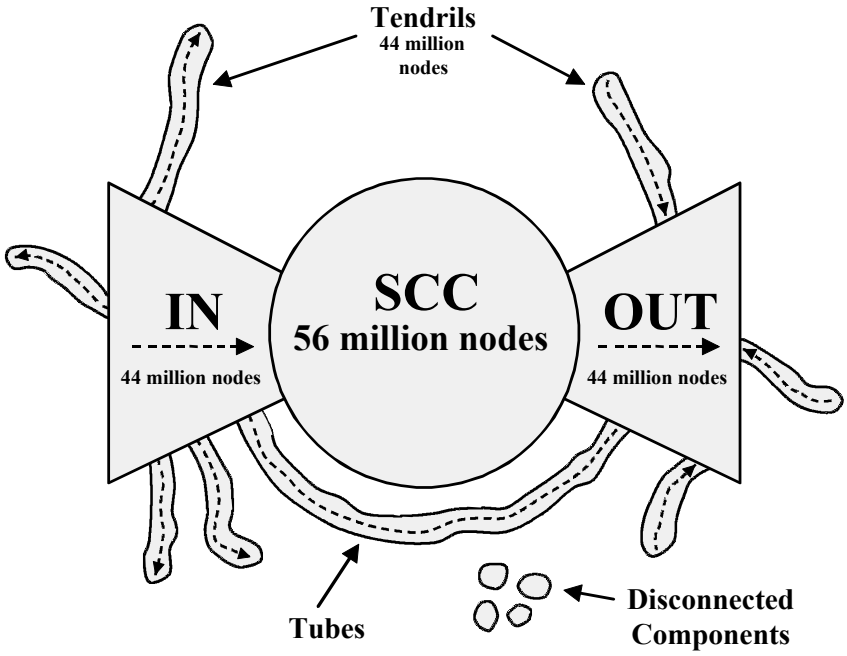
Relevanzbewertung mit Authorities/Hubs

- Idee: themenspezifisch gewichtete Bewertung der eingehenden und ausgehenden Verweise (Kleinberg, 2000)
- Numerische Lösung grosser, dünn besetzter Eigenwertprobleme



Graphstruktur des World Wide Web

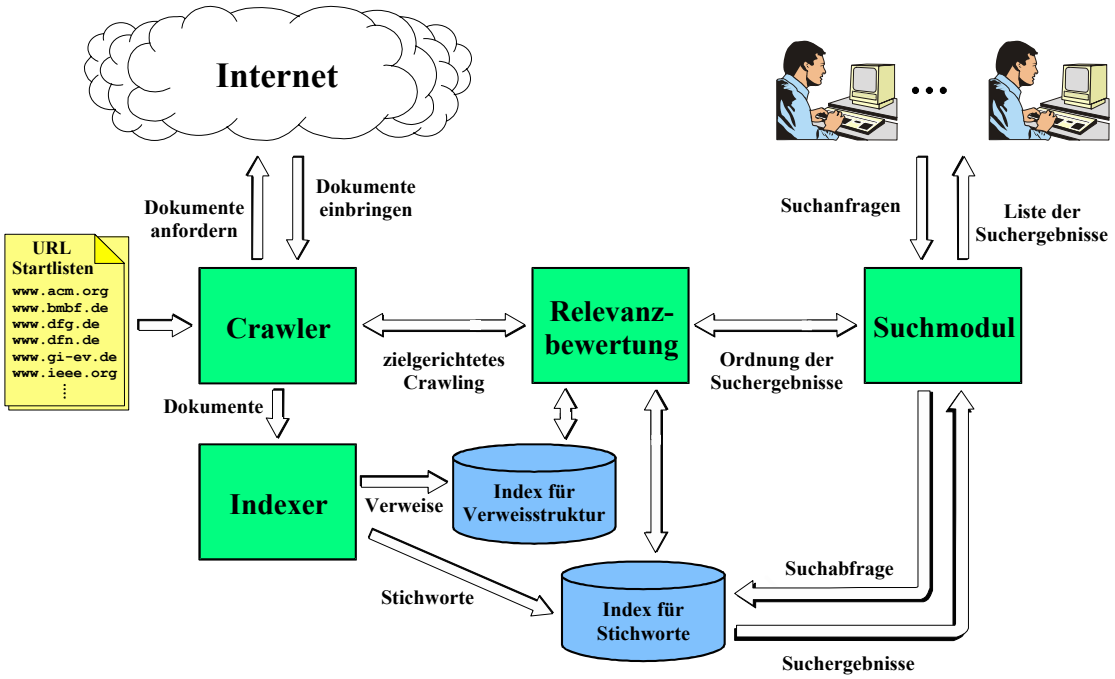
- Das WWW hat eine Bow Tie Struktur, Broder et al. 2000



Ergebnisse der Graphanalyse des WWW

- Wesentliche theoretische Ergebnisse
  - WWW ist sehr gut untereinander verzeigert
  - Trotzdem ist das WWW nicht „eine kleine Welt“
- Praktische Nutzung dieser Ergebnisse
  - Optimierung von Suchverfahren für Web Dokumenten
  - Optimierung von Crawling Strategien
  - Erkennen von Index Spamming verursacht durch kommerzielle „Suchmaschinenoptimierung“
  - Auffinden von Interessengemeinschaften im Web (Cyber Communities)
- Graphstruktur des WWW eingeschränkt auf eine Domäne, ein Themengebiet oder eine Web Site hat auch eine Bow Tie Struktur (Dill et al. 2001)

Interaktion der Softwarekomponenten

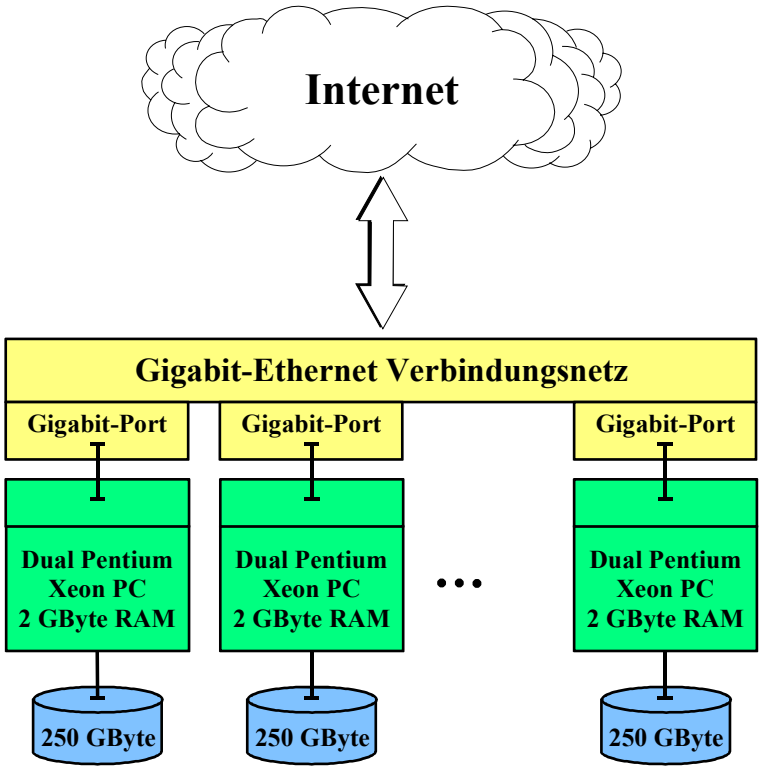


Innovative systemorientierte Features

- **High Performance Web Crawler**
  - Mindestens 200.000 Dokumente/Stunde/Rechner
  - Möglichkeit des zielgerichteten Crawling durch Ausnützen der Verweisstruktur
- **Leistungsstarker und speichereffizienter Indexer**
  - Mindestens 150.000 Dokumente/Stunde/Rechner
  - Indexierung von mehr als 40 Mio. Dokumente pro Rechnerknoten mit 100 Gigabyte Plattenspeicher
- **Semi-Memory Verfahren zur Relevanzbewertung**
  - Graphanalyse für bis zu 200 Mio. Verweise (off-page)
  - URL, Häufigkeit des Stichworts, Fett-, Kursivdruck, (in-page)

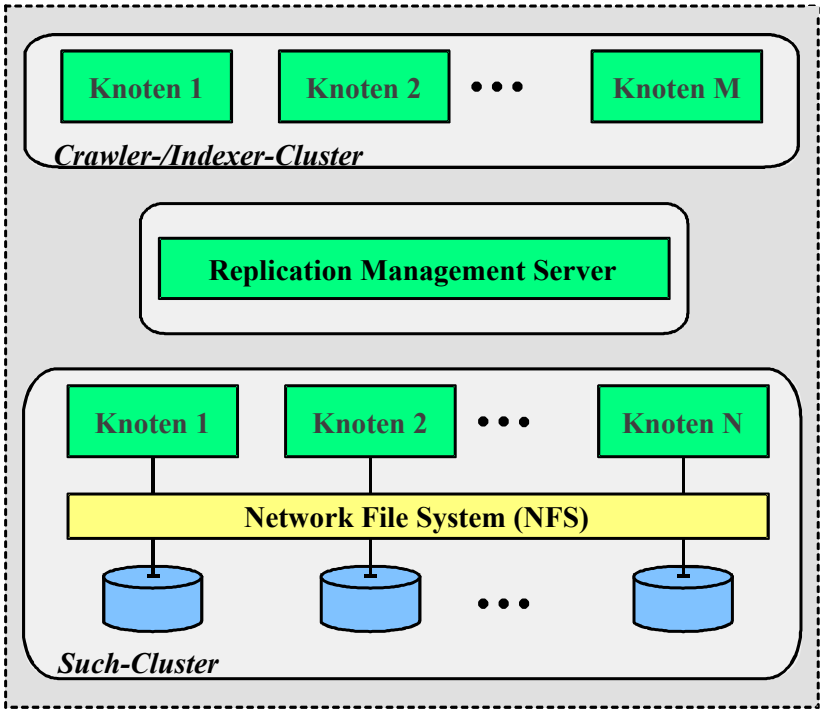
PC Cluster als Hardware Plattform

- Optimales Kosten/ Leistungsverhältnis
- High End Dual Prozessor PC als Rechnerknoten
- Skalierbare Konfigurierung
- Hohe Verfügbarkeit



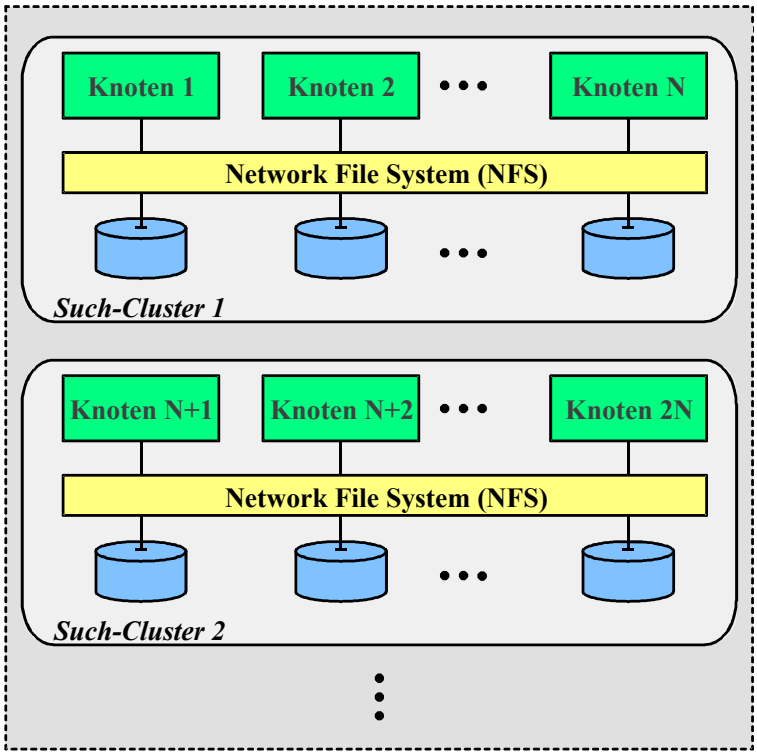
Horizontale Konfigurierung

- Knotenanzahl M bestimmt durch Bandbreite der ISP Anbindung
- Knotenanzahl N bestimmt durch Konfiguration des Gigabit Verbind-ungsnetzes

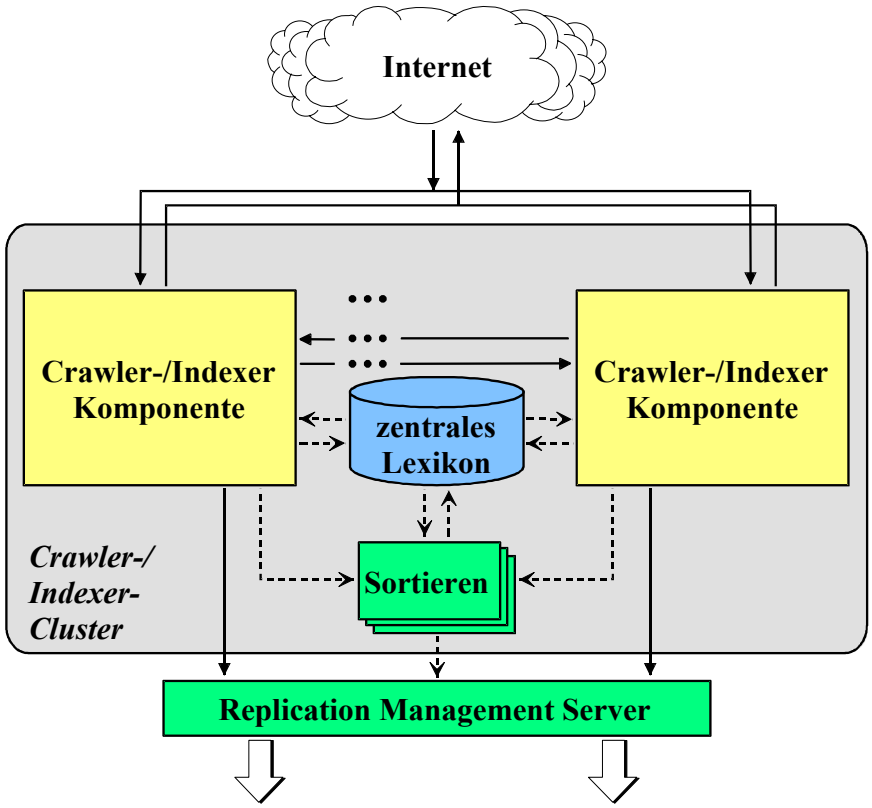


Vertikale Konfigurierung

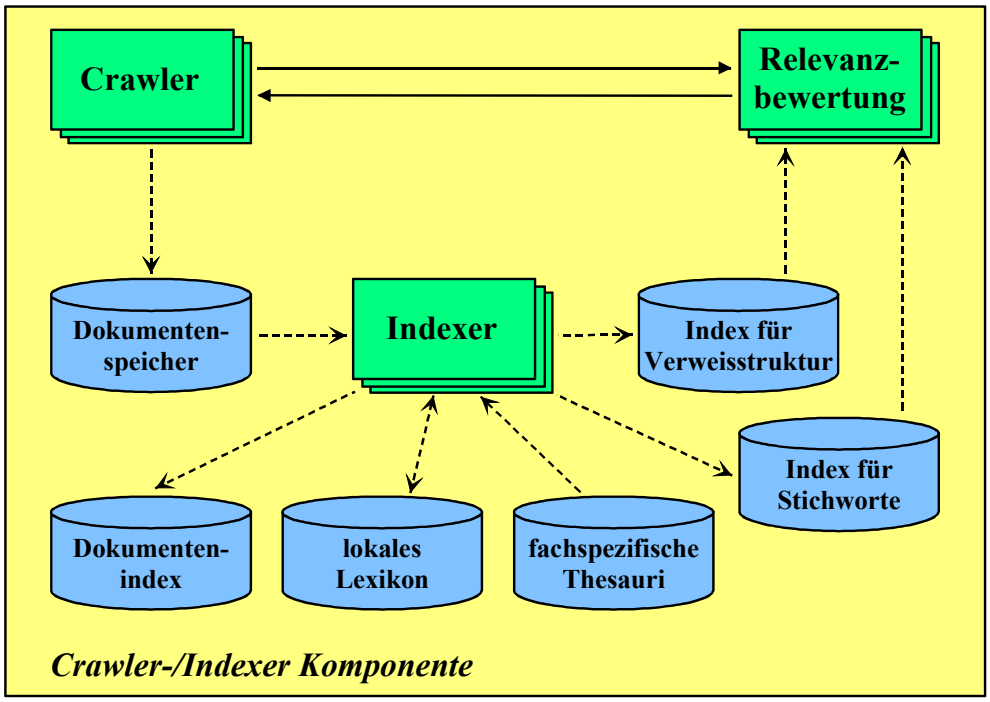
- Anzahl der replizierten Such-Cluster bestimmt durch Dienstgüte-anforderung
  - maximale Antwortzeit unter einer Lastspitze
  - Geforderte Verfügbar-keit: (MTBF – MTTR)/MTBF)
  - minimal zu erzielender Ertrag (Yield)



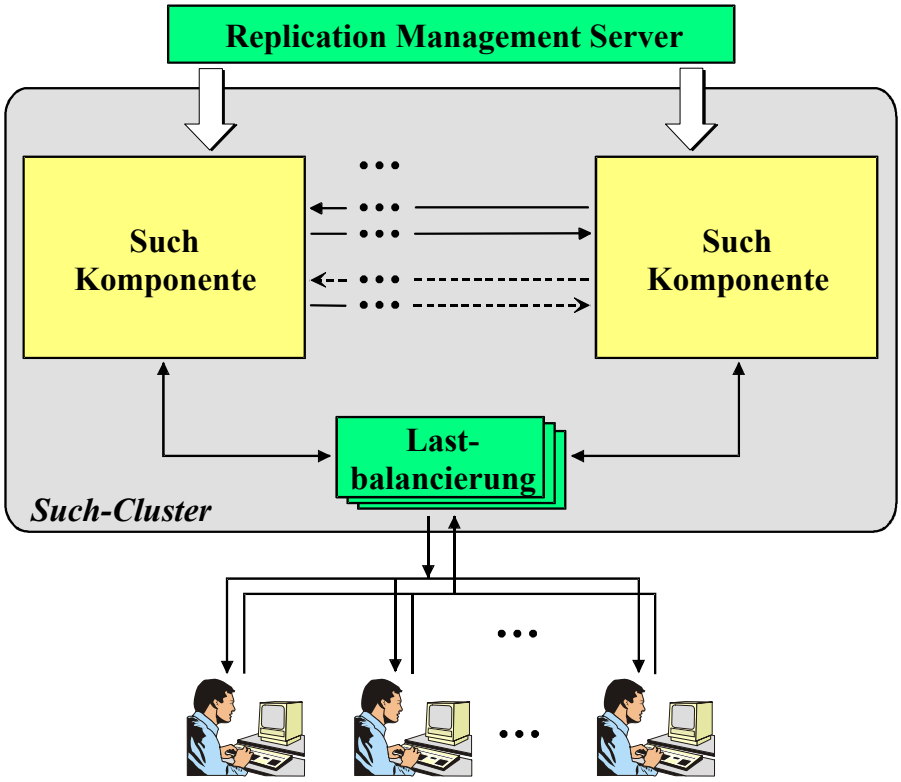
Frontend von WebSearchBench



Aufbau einer Crawler/Indexer Komponente



Backend von WebSearchBench



Aufbau einer Suchkomponente