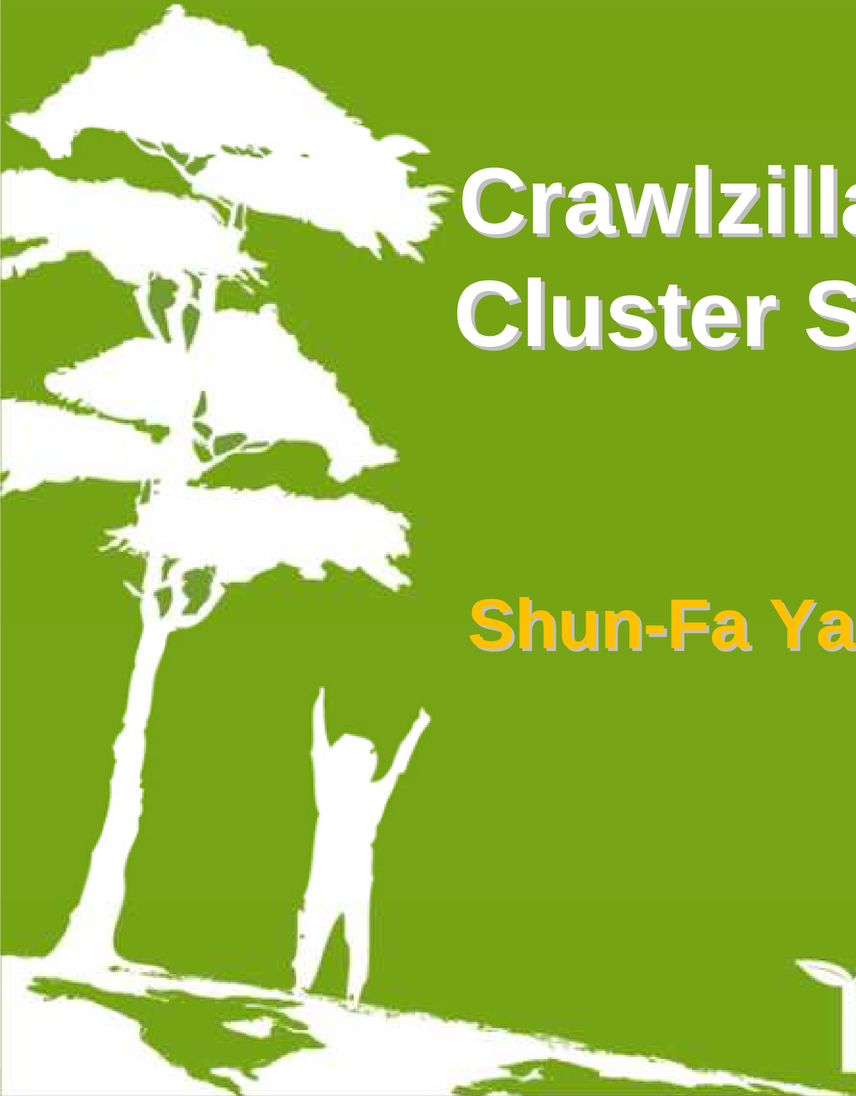
CRAWLZILLA

Crawlzilla - A Toolkit for Deploying Cluster Search Engine Quickly and Easily

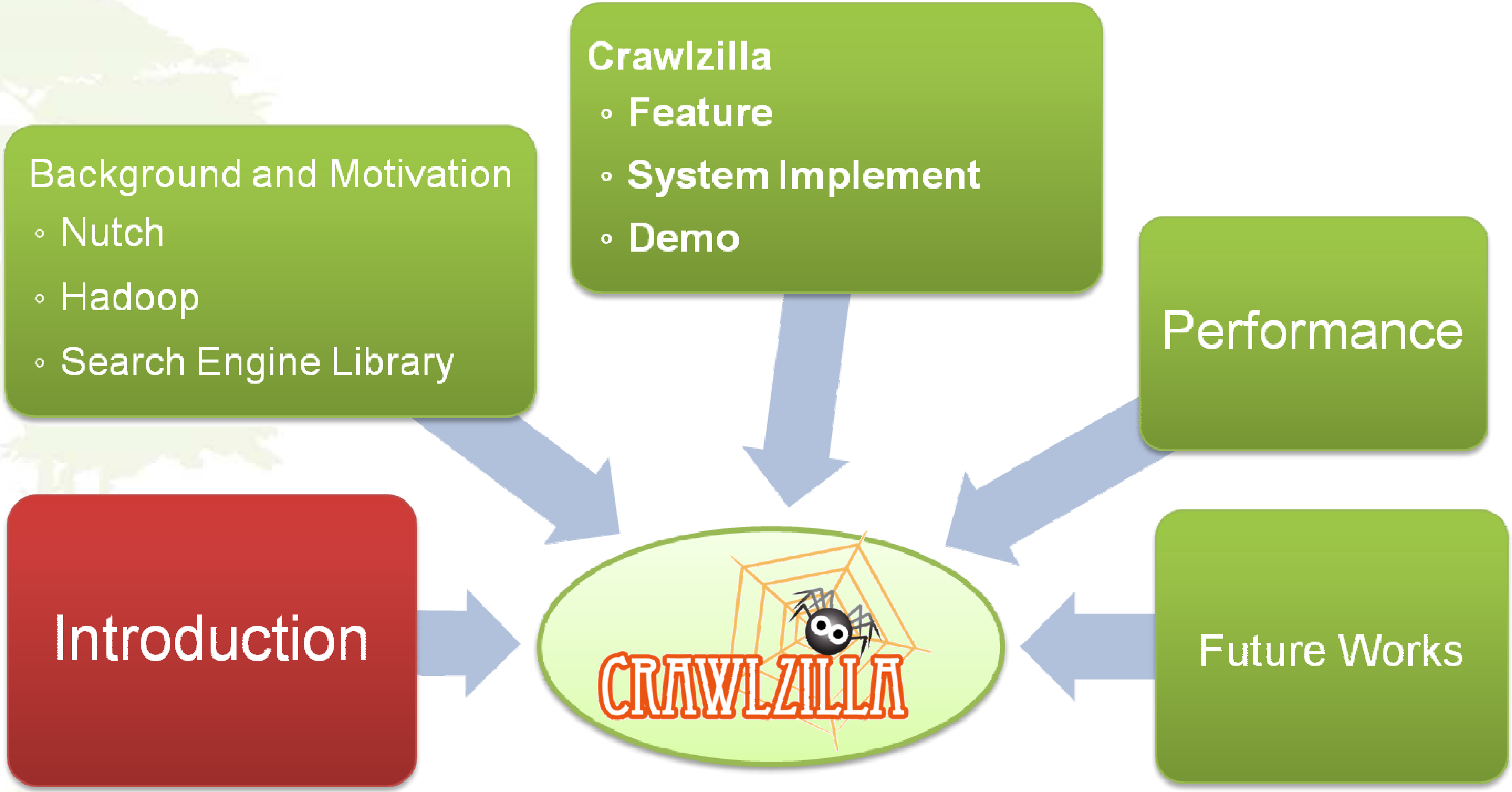
Shun-Fa Yang、Wei-Yu Chen、Wen-Chieh Kuo

Free Software Lab.@ NCHC

INVENSIVE 2011 May 23, 2011



Outline



Introduction

Background and Motivation

- Nutch
- Hadoop
- Search Engine Library

Crawlzilla

- Feature
- System Implement
- Demo

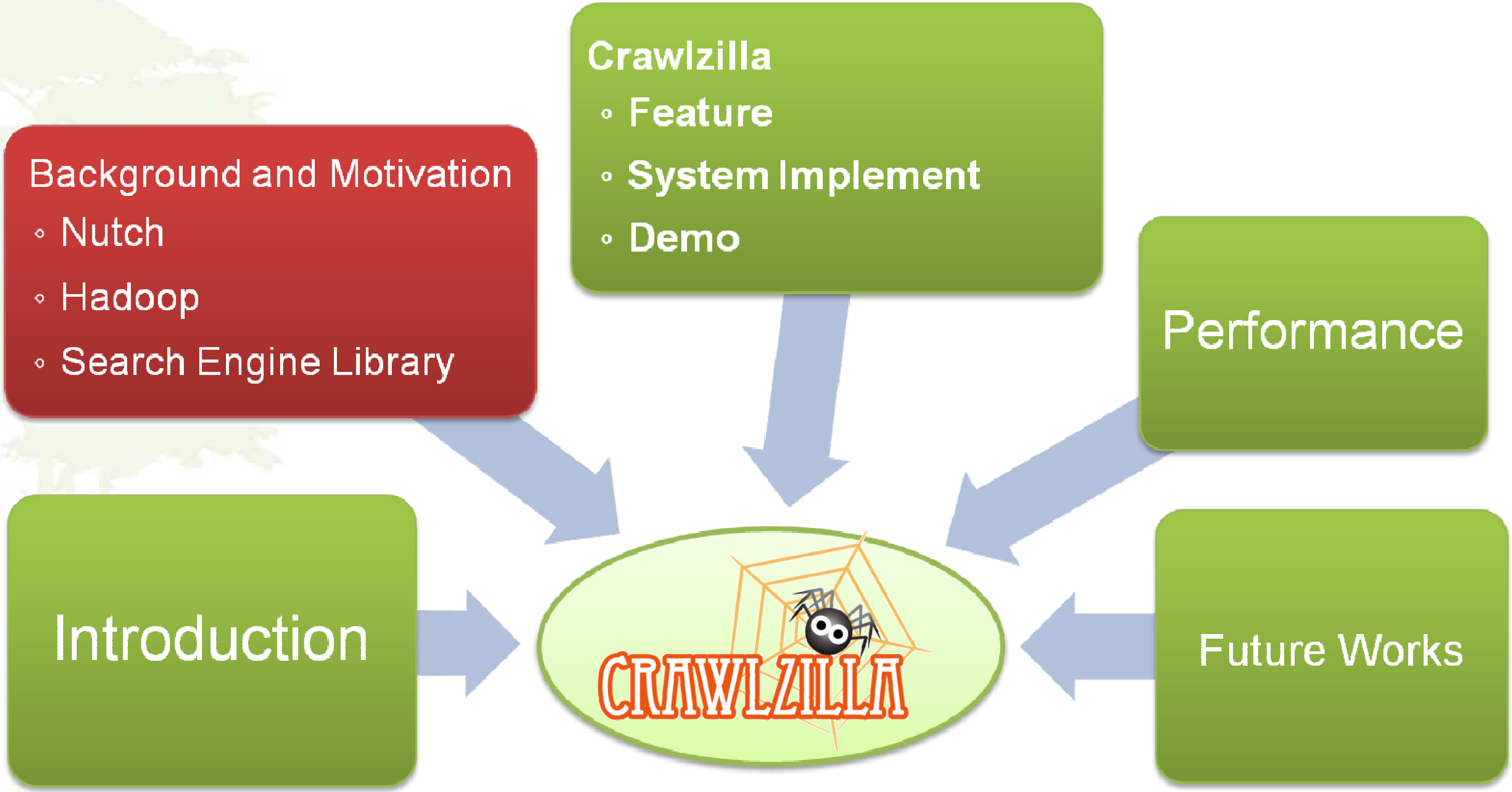
Performance

Future Works

Introduction

- The Information Explosion
- Increase Filter Efficiency by Search Engines
- Intranet also need Search Engines
- Build Search Engines isn’t very Easy
- Crawlzilla can help You!

Outline



Background and Motivation

- Nutch
- Hadoop
- Search Engine Library

Crawlzilla

- Feature
- System Implement
- Demo

Performance

Introduction

Future Works

Background and Motivation

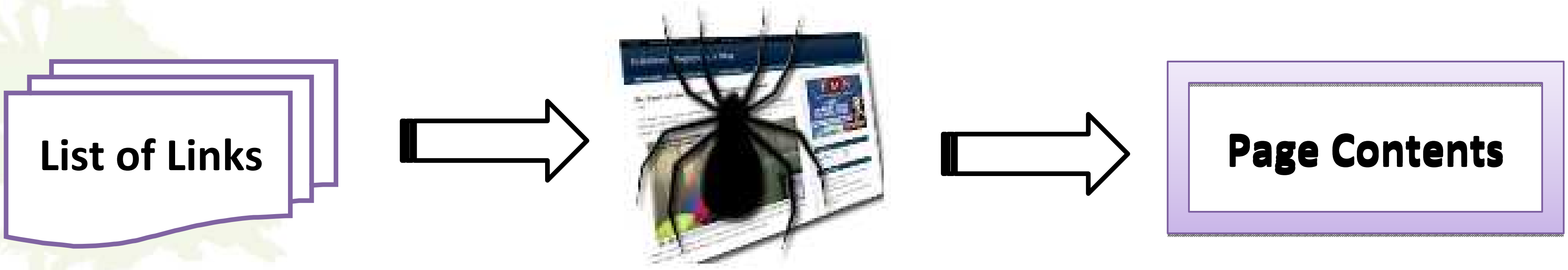
Search Engine workflow

Related Open Source Projects

Compare with Other Projects

Search Engine workflow – Phase 1

- Crawling the Web



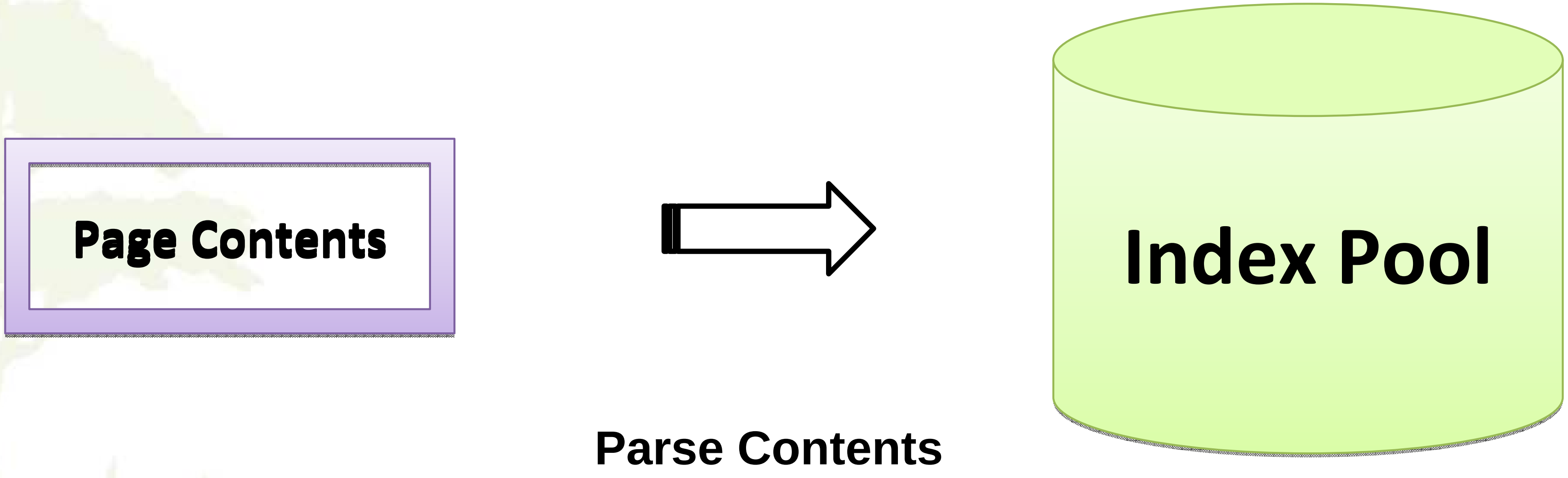
List of Links

Page Contents

Crawler visits the web pages of the links

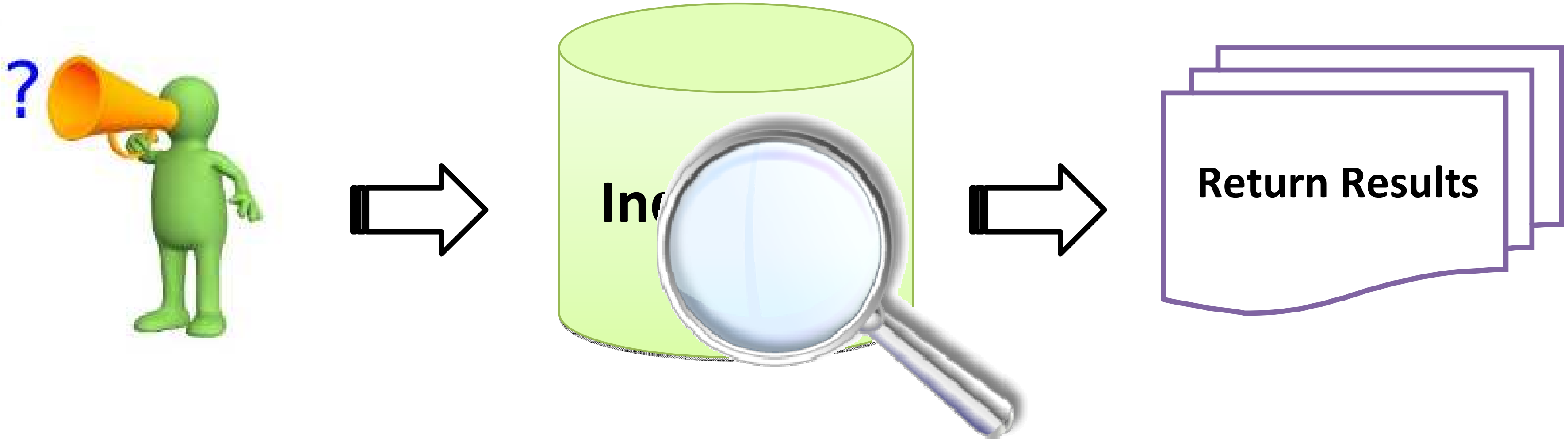
Search Engine workflow – Phase 2

- Building the Index Pool



Search Engine workflow – Phase 3

- **Serving Queries**



Return Results

User Sent a Query **Search from Index Pool**

Background and Motivation

- **Related Open Source Projects**
  - Search Engine - Ntuch
  - Distributed Computing Platform – Hadoop
  - Search Engine Library – Lucene

Background and Motivation

- **If Build Search by Yourself …**
  - Setup Hadoop
  - Deploy System Configure Files
  - Debug Errors…
  - …
  - …
  - …

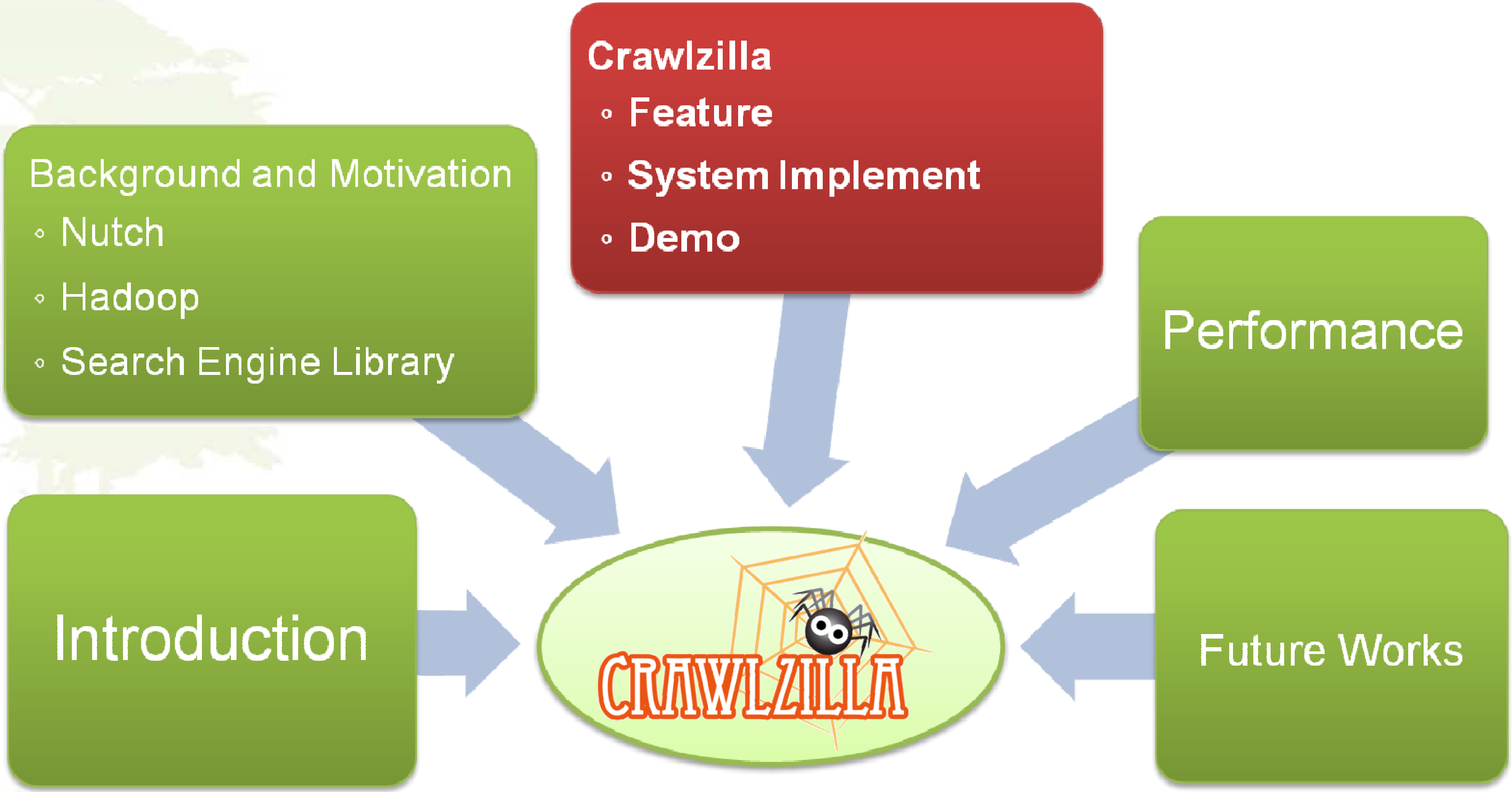
Compare with Other Projects

	Spidr	Larbin	Jcrawl	Nutch	Crawlzilla
Install	Rube Package Install	Gmake Compiler and Install	Java Compiler and Install	Deploy Configure Files	**Provide Auto Installation**
Crawl website pages	O	O	O	O	**O**
Parser Content	X	X	X	O	**O**
Cluster Computing	X	X	X	O	**O**
Interface	Command	Command	Command	Command	**Web-UI**
Support Chinese Segmentation	X	X	X	X	**O**

Goal

- To Help Users to Build Search Engines Easily!
- To Help Users to Operate System Easily!
- Crawlzilla doesn’t improve the algorithm of Nutch and Hadoop!
- Crawlzilla Provides Friendly Operating Interface and an Easy Way to Deploy Cluster Computing Environment!

Outline



Background and Motivation

- Nutch
- Hadoop
- Search Engine Library

Crawlzilla

- Feature
- System Implement
- Demo

Performance

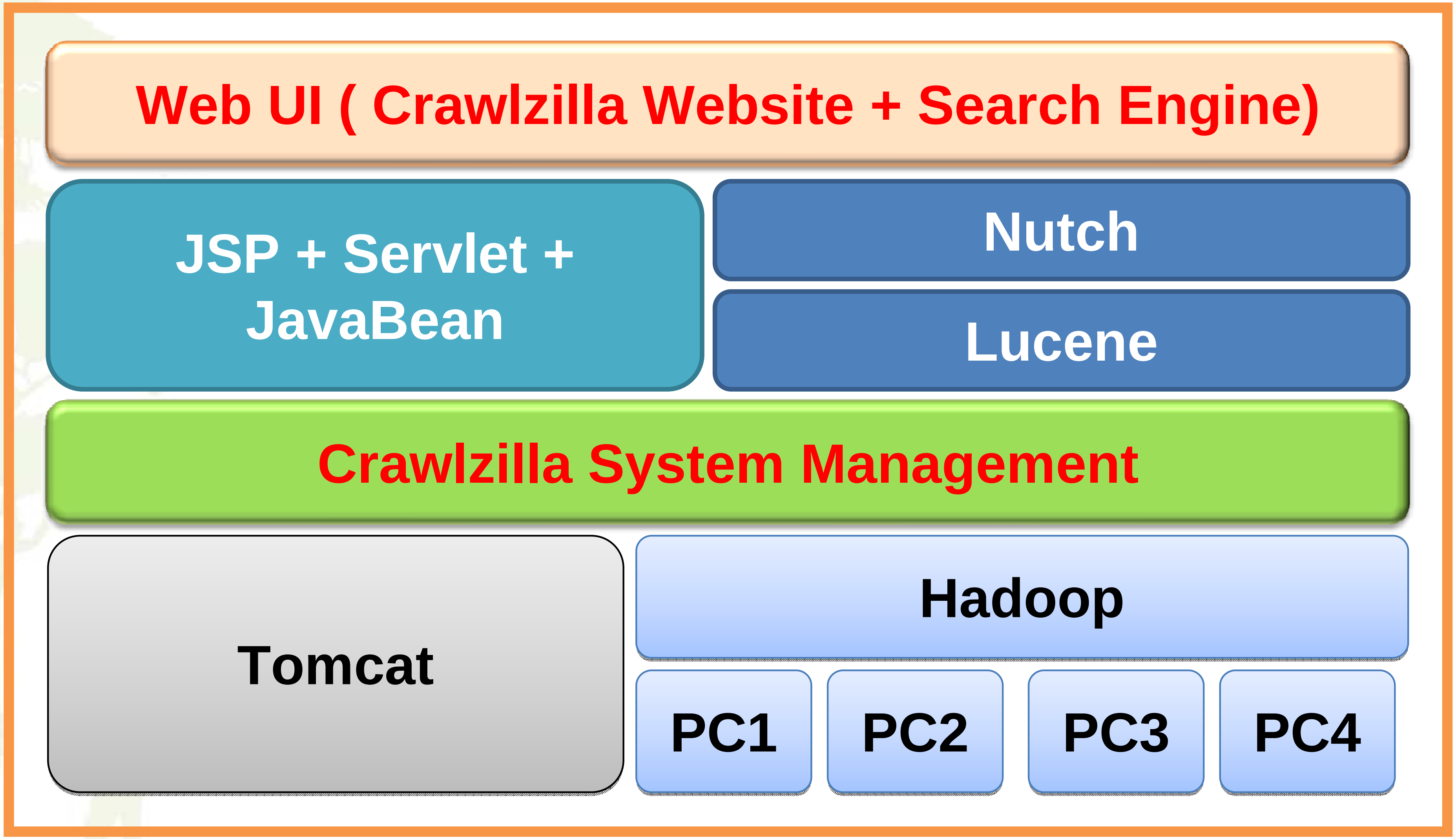
Introduction

Future Works

Crawlzilla Feature

- **Simply Install and Easy to Operate**
  - Customize user interface
- **More Powerful**
  - Support multiple search engines
- **More Search Engine Info.**
- **Developers to focus more**
  - Data mining tools

Crawlzilla Architecture


Web UI (Crawlzilla Website + Search Engine)
JSP + Servlet + JavaBean
Nutch
Lucene
Crawlzilla System Management
Tomcat
Hadoop
PC1
PC2
PC3
PC4

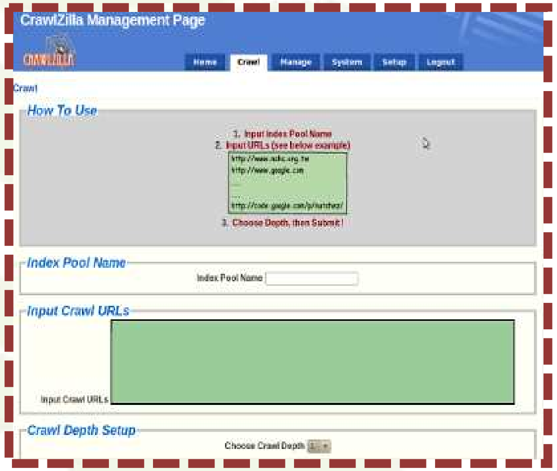
System Implement

	JSP	Shell Script
Function	User UI	Admin and MIS UI
Security	Website Session	Crawler password with RSA Keys
Environment	Browser	Terminal with SSH –Client
Architecture	MVC	Module
Multi Language	i18n	Language parameters
	Default language is depend on O.S. Env.	

Web Management

(Model 2)

Setup and Drive Crawl Procedure



CrawlZilla Management Page

Crawl

How To Use

1. Input Index Pool Name
2. Input URLs (see below example)

http://www.nchc.org.tw
http://www.google.com
...
...
http://code.google.com/p/crawlzilla/

3. Choose Depth, then Submit !

Index Pool Name

Index Pool Name

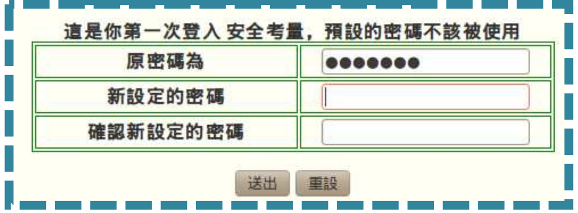
Input Crawl URLs

Input Crawl URLs

Crawl Depth Setup

Choose Crawl Depth 1

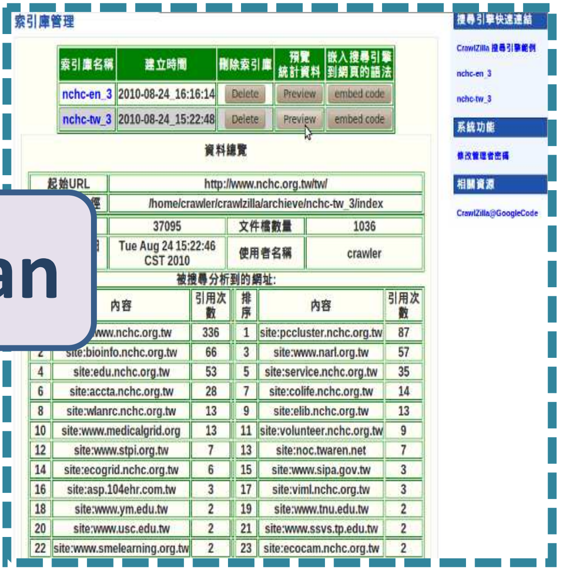
Setup PW



這是你第一次登入 安全考量，預設的密碼不該被使用

原密碼為	●●●●●●●
新設定的密碼	
確認新設定的密碼	

Capture Lucene index pool



索引庫管理

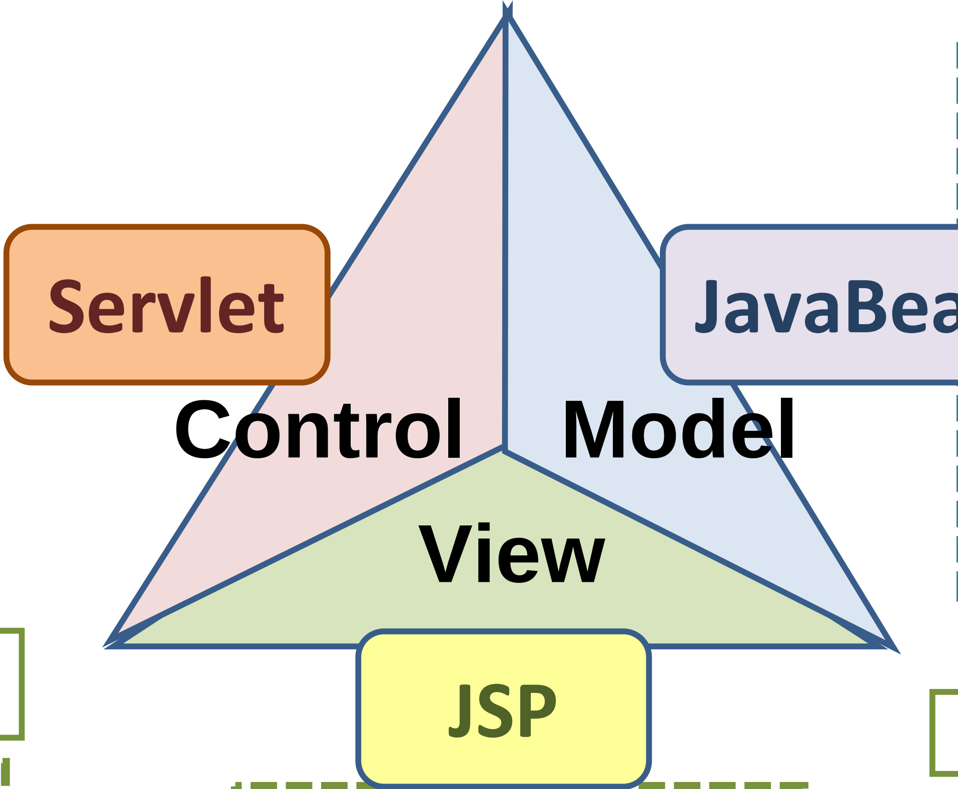
索引庫名稱	建立時間	刪除索引庫	預覽 統計資料	嵌入搜尋引擎 到網頁的語法
nchc-en_3	2010-08-24_16:16:14	Delete	Preview	embed code
nchc-tw_3	2010-08-24_15:22:48	Delete	Preview	embed code

資料總覽

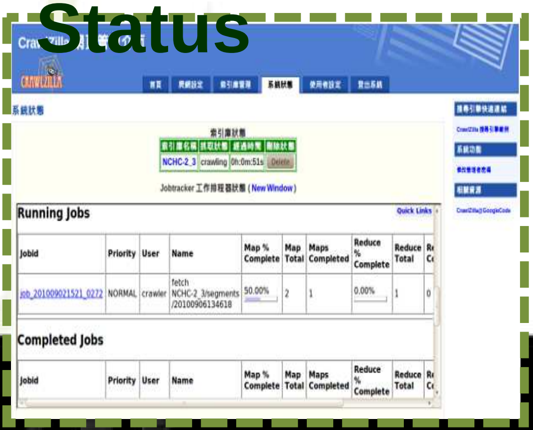
起始URL	http://www.nchc.org.tw/tw/		
徑	/home/crawler/crawlzilla/archieve/nchc-tw_3/index		
	37095	文件檔數量	1036
	Tue Aug 24 15:22:46 CST 2010	使用者名稱	crawler

被搜尋分析到的網址:

	內容	引用次數	排序	內容	引用次數
	ww.nchc.org.tw	336	1	site:pccluster.nchc.org.tw	87
2	site:bioinfo.nchc.org.tw	66	3	site:www.narl.org.tw	57
4	site:edu.nchc.org.tw	53	5	site:service.nchc.org.tw	35
6	site:accta.nchc.org.tw	28	7	site:colife.nchc.org.tw	14
8	site:wlanrc.nchc.org.tw	13	9	site:elib.nchc.org.tw	13
10	site:www.medicalgrid.org	13	11	site:volunteer.nchc.org.tw	9
12	site:www.stpi.org.tw	7	13	site:noc.twaren.net	7
14	site:ecogrid.nchc.org.tw	6	15	site:www.sipa.gov.tw	3
16	site:asp.104ehr.com.tw	3	17	site:viml.nchc.org.tw	3
18	site:www.ym.edu.tw	2	19	site:www.tnu.edu.tw	2
20	site:www.usc.edu.tw	2	21	site:www.ssvs.tp.edu.tw	2
22	site:www.smelearning.org.tw	2	23	site:ecocam.nchc.org.tw	2



System Status



系統狀態

索引庫狀態

索引庫名稱	抓取狀態	經過時間	刪除狀態
NCHC-2_3	crawling	0h:0m:51s	Delete

Jobtracker 工作排程器狀態 (New Window)

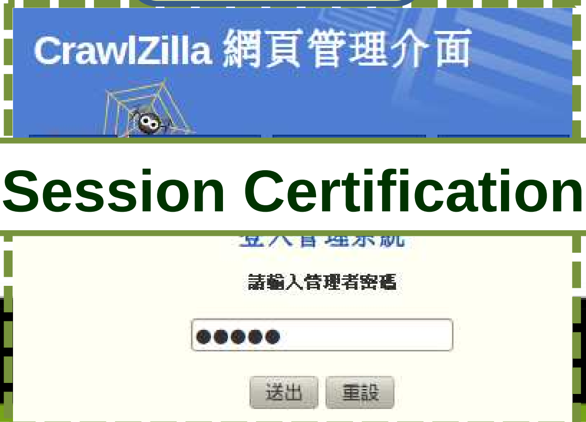
Running Jobs

Jobid	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Re Co
job_201009021521_0272	NORMAL	crawler	fetch NCHC-2_3/segments /20100906134618	50.00%	2	1	0.00%	1	0

Completed Jobs

Jobid	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Re Co

Session Certification



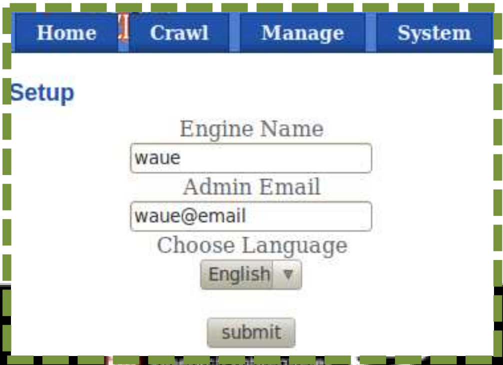
CrawlZilla 網頁管理介面

登入管理系統

請輸入管理者密碼

●●●●●

i18N language setup



Setup

Engine Name

waue

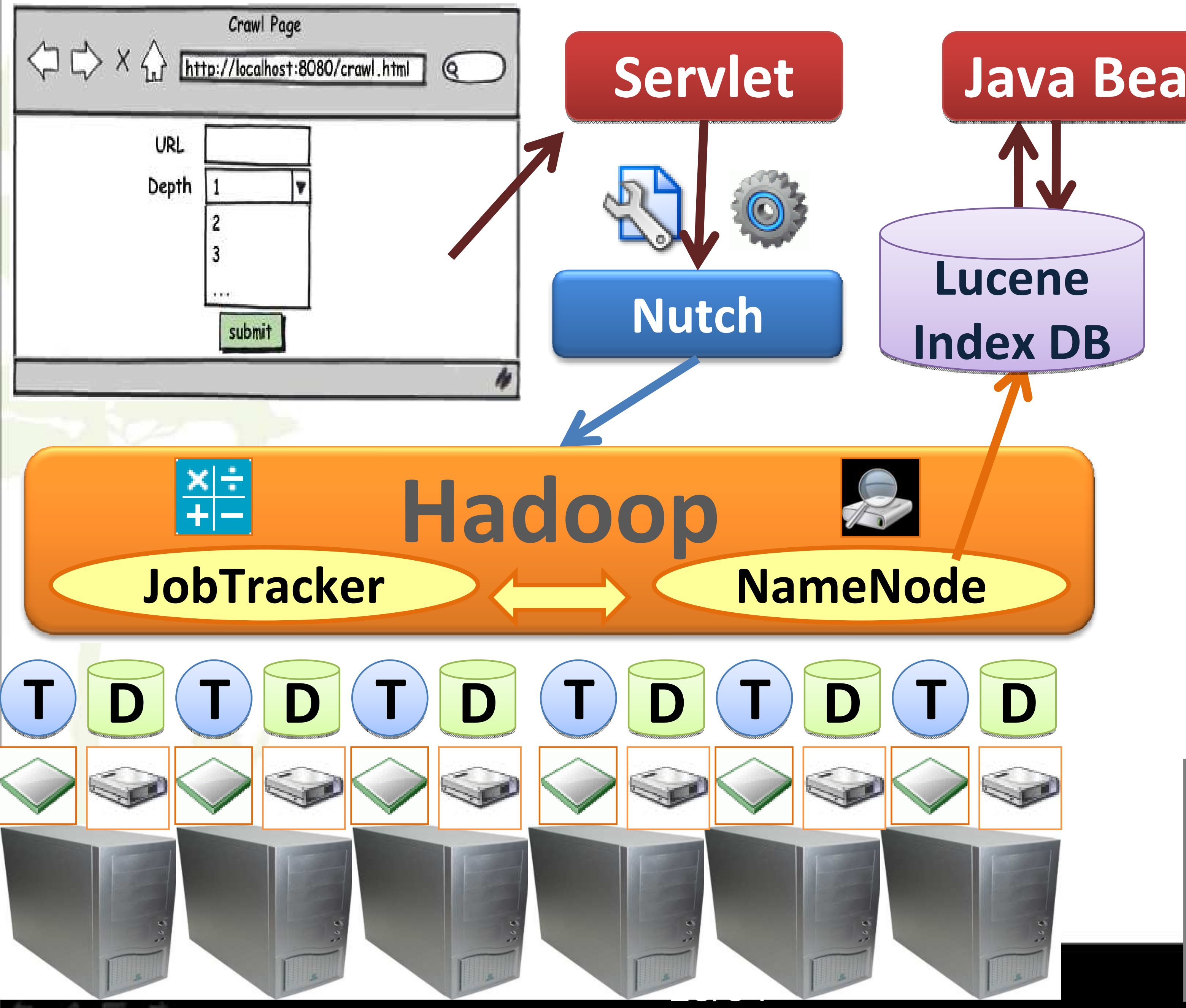
Admin Email

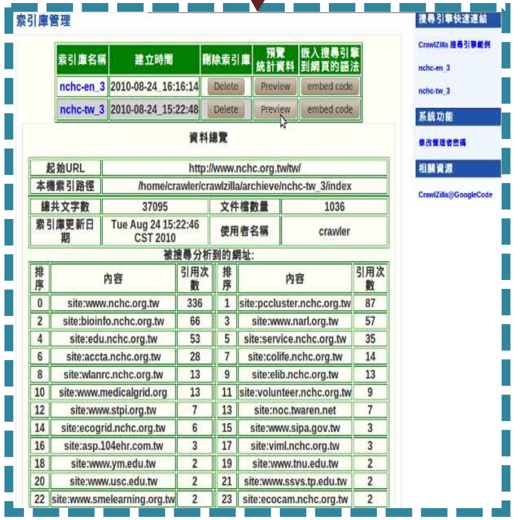
waue@email

Choose Language

English ▼

System Implement – Web Parser





索引庫管理

索引庫名稱	建立時間	刪除索引庫	預覽 統計資料	嵌入搜尋引擎到網頁的語法
nchc-en_3	2010-08-24_16:16:14	Delete	Preview	embed code
nchc-tw_3	2010-08-24_15:22:48	Delete	Preview	embed code

資料總覽

起始URL	http://www.nchc.org.tw/tw/		
本機索引路徑	/home/crawler/crawlzilla/archieve/nchc-tw_3/index		
總共文字數	37095	文件檔數量	1036
索引庫更新日期	Tue Aug 24 15:22:46 CST 2010	使用者名稱	crawler

被搜尋分析到的網址:

排序	內容	引用次數	排序	內容	引用次數
0	site:www.nchc.org.tw	336	1	site:pccluster.nchc.org.tw	87
2	site:bioinfo.nchc.org.tw	66	3	site:www.narl.org.tw	57
4	site:edu.nchc.org.tw	53	5	site:service.nchc.org.tw	35
6	site:accta.nchc.org.tw	28	7	site:colife.nchc.org.tw	14
8	site:wlanrc.nchc.org.tw	13	9	site:elib.nchc.org.tw	13
10	site:www.medicalgrid.org	13	11	site:volunteer.nchc.org.tw	9
12	site:www.stpi.org.tw	7	13	site:noc.twaren.net	7
14	site:ecogrid.nchc.org.tw	6	15	site:www.sipa.gov.tw	3
16	site:asp.104ehr.com.tw	3	17	site:viml.nchc.org.tw	3
18	site:www.ym.edu.tw	2	19	site:www.tnu.edu.tw	2
20	site:www.usc.edu.tw	2	21	site:www.ssvs.tp.edu.tw	2
22	site:www.smelearning.org.tw	2	23	site:ecocam.nchc.org.tw	2

搜尋引擎快速連結

CrawlZilla 搜尋引擎範例

nchc-en_3

nchc-tw_3

系統功能

修改管理者密碼

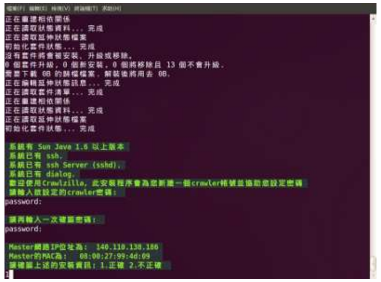
相關資源

CrawlZilla@GoogleCode

T TaskTracker: Job Executor

D DataNode: Data Storage Node

Friendly Interface!

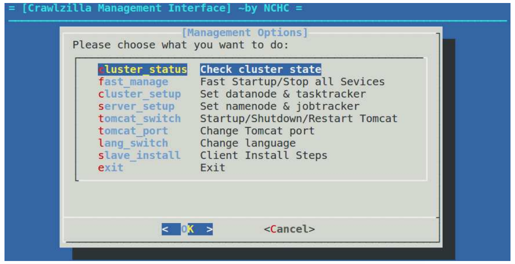


```
正在重建相依關係
正在讀取狀態資料... 完成
正在讀取延伸狀態檔案
初始化套件狀態... 完成
沒有套件將會被安裝、升級或移除。
0 個套件升級，0 個新安裝，0 個將移除且 13 個不會升級。
需要下載 0B 的歸檔檔案，解裝後將用去 0B。
正在編輯延伸狀態訊息... 完成
正在讀取套件清單... 完成
正在重建相依關係
正在讀取狀態資料... 完成
正在讀取延伸狀態檔案
初始化套件狀態... 完成

 系統有 Sun Java 1.6 以上版本
 系統已有 ssh.
 系統已有 ssh Server (sshd).
 系統已有 dialog.
 歡迎使用Crawlzilla, 此安裝程序會為您新建一個crawler帳號並協助您設定密碼
 請輸入欲設定的crawler密碼:
password:

 請再輸入一次確認密碼:
password:

 Master網路IP位址為:  140.110.138.186
 Master的MAC為:   08:00:27:99:4d:09
 請確認上述的安裝資訊: 1.正確 2.不正確
1
```



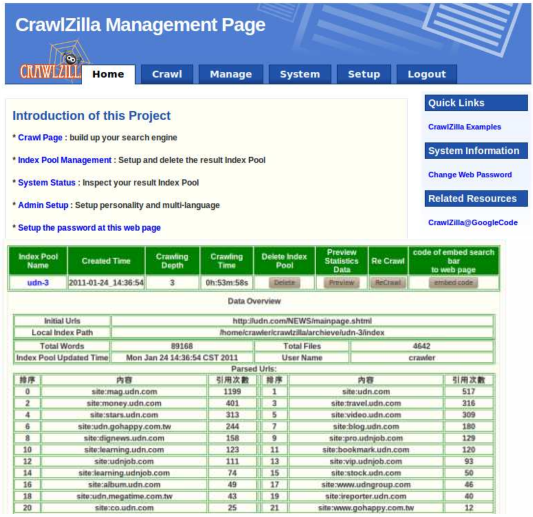
```
= [Crawlzilla Management Interface] ~by NCHC =

            [Management Options]
 Please choose what you want to do:

    cluster_status   Check cluster state
    fast_manage      Fast Startup/Stop all Sevices
    cluster_setup    Set datanode & tasktracker
    server_setup     Set namenode & jobtracker
    tomcat_switch    Startup/Shutdown/Restart Tomcat
    tomcat_port      Change Tomcat port
    lang_switch      Change language
    slave_install    Client Install Steps
    exit             Exit

          <  OK  >          <Cancel>
```

Admin



Introduction of this Project

- * Crawl Page : build up your search engine
- * Index Pool Management : Setup and delete the result Index Pool
- * System Status : Inspect your result Index Pool
- * Admin Setup : Setup personality and multi-language
- * Setup the password at this web page

Index Pool Name	Created Time	Crawling Depth	Crawling Time	Delete Index Pool	Preview Statistics Data	Re Crawl	code of embed search bar to web page
udn-3	2011-01-24_14:36:54	3	0h:53m:58s	Delete	Preview	ReCrawl	embed code

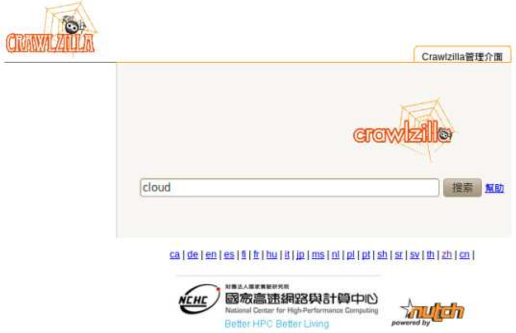
Data Overview

Initial Urls	http://udn.com/NEWS/mainpage.shtml		
Local Index Path	/home/crawler/crawlzilla/archieve/udn-3/index		
Total Words	89168	Total Files	4642
Index Pool Updated Time	Mon Jan 24 14:36:54 CST 2011	User Name	crawler

Parsed Urls:

排序	內容	引用次數	排序	內容	引用次數
0	site:mag.udn.com	1199	1	site:udn.com	517
2	site:money.udn.com	401	3	site:travel.udn.com	316
4	site:stars.udn.com	313	5	site:video.udn.com	309
6	site:udn.gohappy.com.tw	244	7	site:blog.udn.com	180
8	site:dignews.udn.com	158	9	site:pro.udnjob.com	129
10	site:learning.udn.com	123	11	site:bookmark.udn.com	120
12	site:udnjob.com	111	13	site:vip.udnjob.com	93
14	site:learning.udnjob.com	74	15	site:stock.udn.com	50
16	site:album.udn.com	49	17	site:www.udngroup.com	46
18	site:udn.megatime.com.tw	43	19	site:ireporter.udn.com	40
20	site:co.udn.com	25	21	site:www.gohappy.com.tw	12

MIS



CRAWLZILLA
Crawlzilla管理介面
crawlzilla
cloud
搜索
幫助
ca | de | en | es | fi | fr | hu | it | jp | ms | nl | pl | pt | sh | sr | sv | th | zh | cn |
NCHC
財團法人國家實驗研究院
國家高速網路與計算中心
National Center for High-Performance Computing
Better HPC Better Living
powered by nutch

USER

Live Demo I

Crawlzilla Install

(1) Master Install

(2) Cluster Slave Install

~~Live~~ Video Demo

- **Master Install (Demo Video also @ YouTube)**

~~Live~~ Video Demo

- **Slave Install (Demo Video also @ YouTube)**

Live Demo II

Dialog Management

Live Demo III

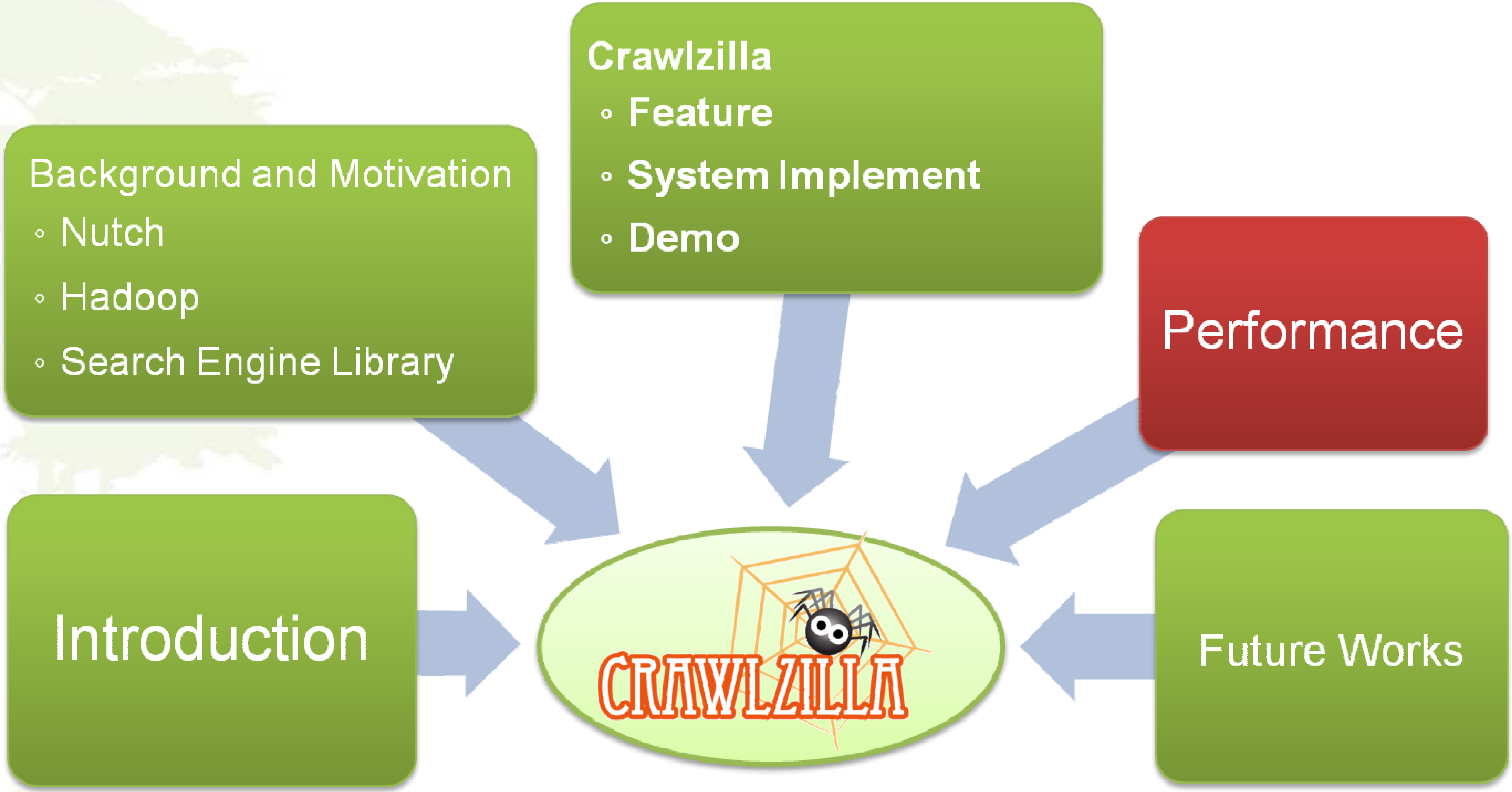
Web Management

(1) Crawl Setup

(2)Search Engine Index Pool

(3)Search it!

Outline



Background and Motivation
- Nutch
- Hadoop
- Search Engine Library

Crawlzilla
- Feature
- System Implement
- Demo

Performance

Introduction

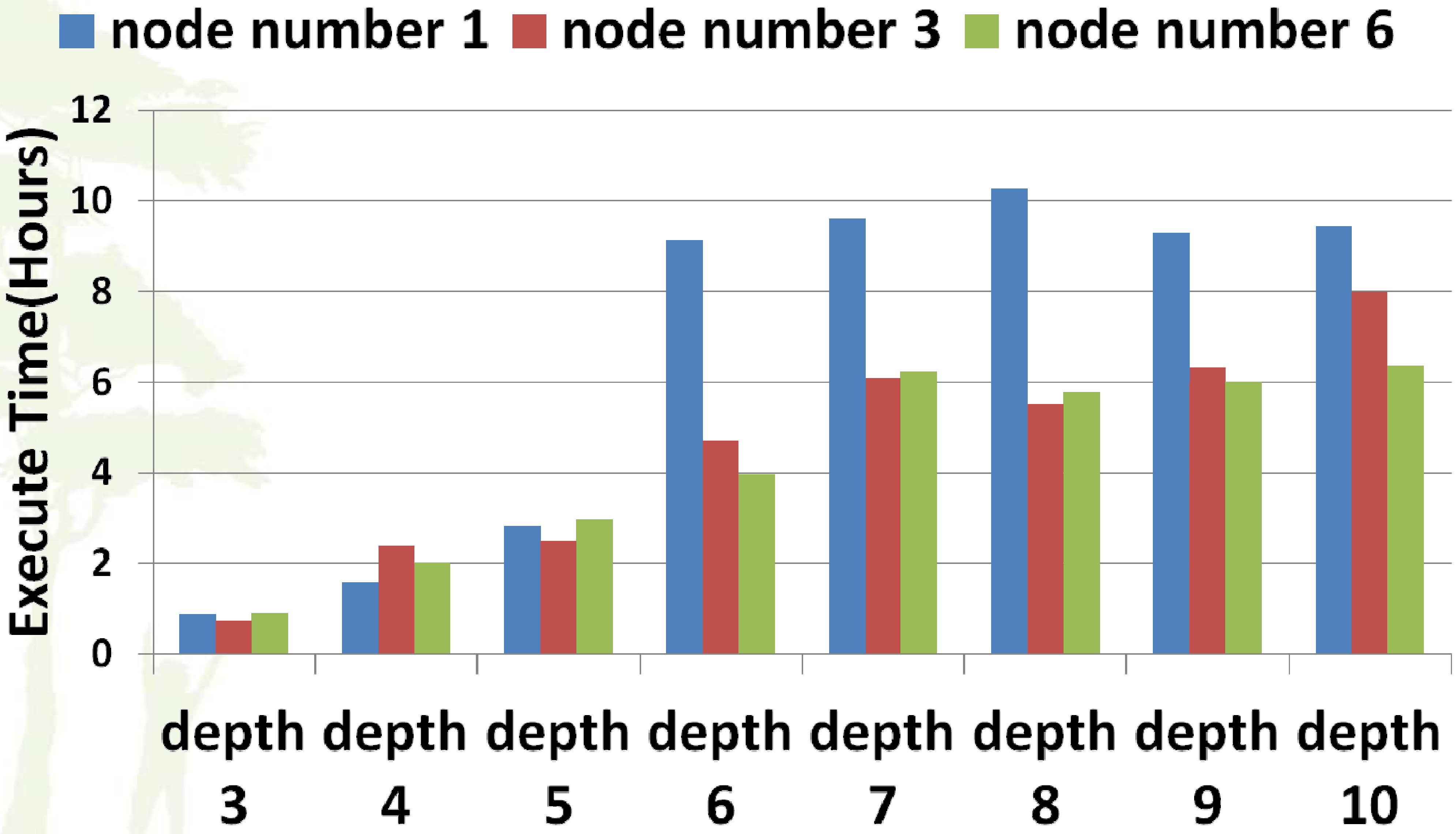
Future Works

Performance

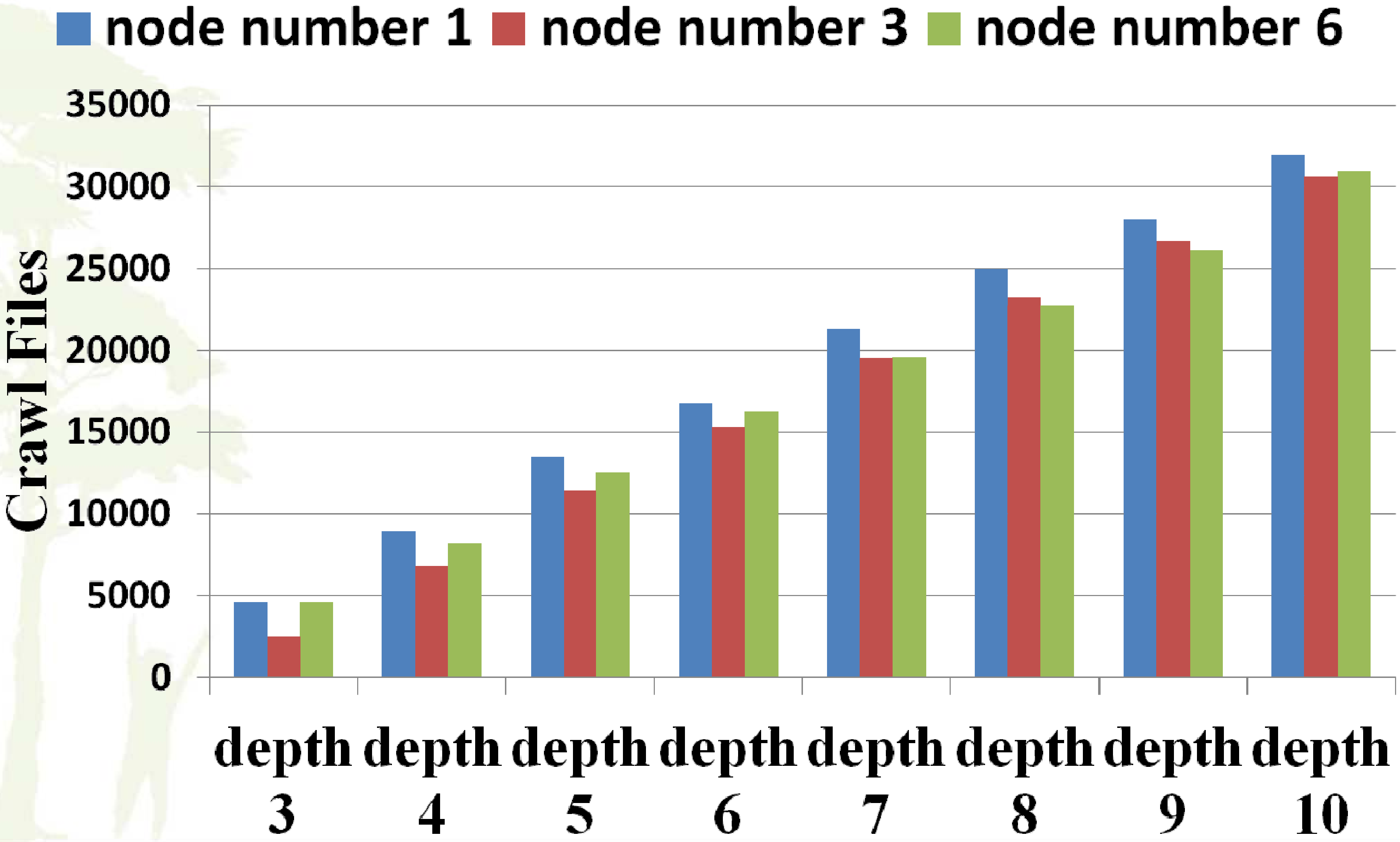
Experiment Environment

- CPU
  - Intel(R) Core(TM)2 Quad CPU Q9550 2.83GHz
- Memroy
  - 8 GigaBytes
- Operation System
  - Ubuntu 10.04 Lucid(x86)
- Crawlzilla Version
  - 0.3.0-101116

Execute Time

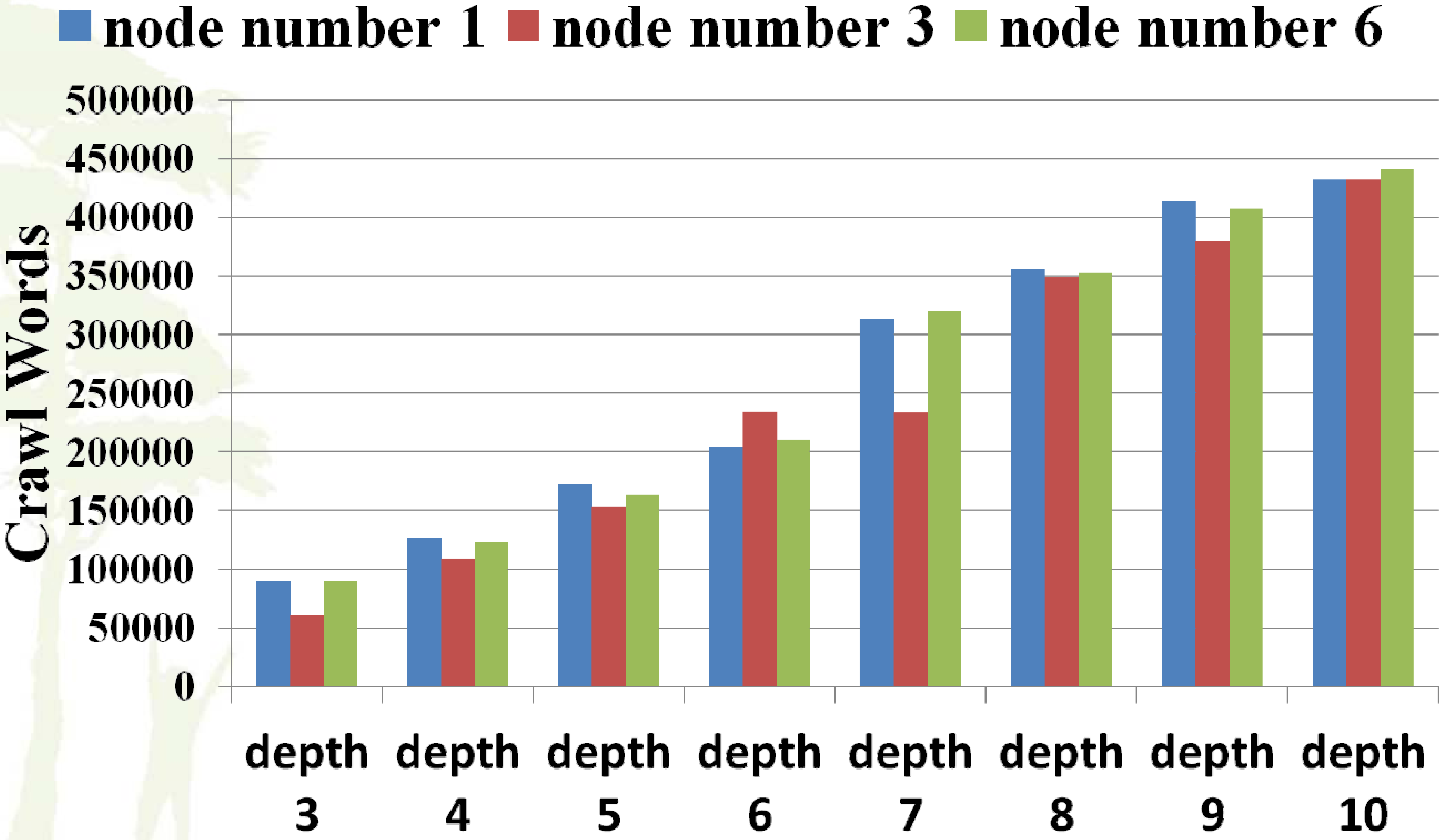


Crawl Files



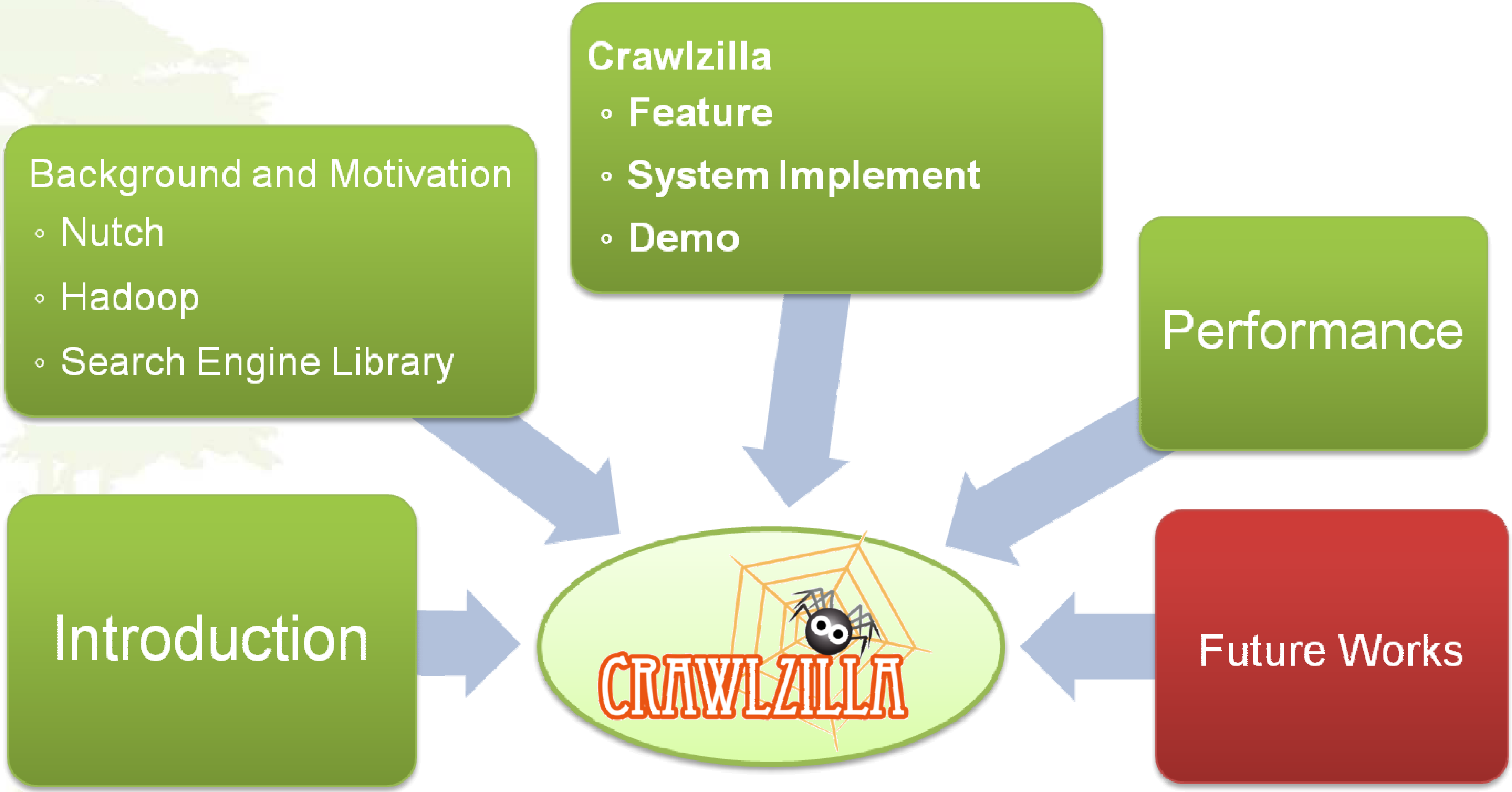
🟦 node number 1
🟥 node number 3
🟩 node number 6
Crawl Files
35000
30000
25000
20000
15000
10000
5000
0
depth 3
depth 4
depth 5
depth 6
depth 7
depth 8
depth 9
depth 10

Crawl Words



🟦 node number 1 🟥 node number 3 🟩 node number 6
Crawl Words
500000
450000
400000
350000
300000
250000
200000
150000
100000
50000
0
depth 3
depth 4
depth 5
depth 6
depth 7
depth 8
depth 9
depth 10

Outline



Introduction

Background and Motivation
- Nutch
- Hadoop
- Search Engine Library

Crawlzilla
- Feature
- System Implement
- Demo

Performance

Future Works

Future Works

- **New Version**
  - Support Multi User
  - Support Schedule
  - Update the Kernel
  - More Easily to deploy Slave Computing Nodes
  - Now is testing!
  - Release Day See http://crawlzilla.info

Reference

- J. Dean and S. Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, In Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation - Volume 6, San Francisco, CA, December 06 - 08, 2004.
- S. Ghemawat, H. Gobioff and S. T. Leung, The Google File System, 19th ACM Symposium on Operating Systems Principles, Lake George, NY, October, 2003.
- The Apache Software Foundation, Nutch, available at: http://nutch.apache.org/ , accessed 5 June 2010.
- The Apache Software Foundation, Hadoop, available at: http://hadoop.apache.org/ , accessed 5 June 2010.
- The Apache Software Foundation, Lucene, available at: http://lucene.apache.org/ , accessed 5 June 2010.
- Crawlzilla @ Google Code Project Hosting, available at: http://code.google.com/p/crawlzilla/, accessed 15 Sep 2010.

Enjoy your search engines!!!
Start from Here!

- Crawlzilla @ Google Code Project Hosting (Tutorials in Chinese)
  - http://code.google.com/p/crawlzilla/
- Crawlzilla @ Source Forge (Tutorials in English)
  - http://sourceforge.net/p/crawlzilla/home/
- Crawlzilla User Group @ Google
  - http://groups.google.com/group/crawlzilla-user
- NCHC Cloud Computing Research Group
  - http://trac.nchc.org.tw/cloud

Thank You!
Q & A