

Invenio Technology

Selected Practical Software Development Lessons From A Large Digital Library System

Tibor Šimko

`<tibor.simko@cern.ch>`

Department of Information Technology
CERN

August 2010 / openlab talk

Outline

- 1** Introduction
 - Digital Library
 - Invenio
- 2** Case Studies
 - Episode 1: Python
 - Episode 2: Git
 - Episode 3: Test Suite
 - Episode 4: Building Efficient Indexes
 - Episode 5: Load-balancing
- 3** Conclusions

What is Digital Library?

Invenio
Technology

Tibor Šimko

Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python

Episode 2: Git

Episode 3: Test Suite

Episode 4: Building
Efficient Indexes

Episode 5:

Load-balancing

Conclusions

- *“library in which collections are stored in digital formats (as opposed to print, microform, or other media) and accessible by computers”*
- (1) institutional document repositories
- (2) world-wide subject-based information systems

Example: CERN Document Server

- managing CERN and selected non-CERN high-energy physics and related documents since ~1993
- more than 1,000,000 records
- articles, books, theses, photos, videos, and more
- powered by Invenio, free digital library software
- <http://cdsweb.cern.ch/>

CDS: Collection Tree

Invenio
Technology

Tibor Šimko

Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python

Episode 2: Git

Episode 3: Test Suite

Episode 4: Building
Efficient Indexes

Episode 5:

Load-balancing

Conclusions

Search 1,042,138 records for:

any field

Search

Browse

[Search Tips](#) :: [Advanced Search](#)

NEW Check out photos and videos of the [LHC First Physics](#).

Narrow by collection:

- Articles & Preprints** (902,678)
[Published Articles](#) (317,500)
[Preprints](#) (526,756) [Theses](#) (16,829)
[Reports](#) (5,571) [CERN Internal Documents](#) (22,545)
[Notes](#) (15,807) [Committee Documents](#) (22,545)
- Books & Proceedings** (71,769)
[Books](#) (46,326) [Proceedings](#) (16,894)
[Standards](#) (8,553)
- Presentations & Talks** (17,514)
[Conference Announcements](#) (15,065)
[Academic Training Lectures](#) (615) [Summer Student Lectures](#) (616) [General Talks](#) (1,212) [Videotapes](#) (291)
- Periodicals & Progress Reports** (2,829)
[Periodicals](#) (2,223) [Progress Reports](#) (606)
- Multimedia & Outreach** (52,208)
[Photos](#) (13,640) [Videos](#) (1,137)
[Press](#) (31,701) [Audio Archives](#) (436)
[Exhibition Objects](#) (179) [Brochures](#) (125)
[Posters](#) (400) [LHC Lectures](#) (2,000)

Focus on:

- CERN Articles & Preprints** (95,348)
[CERN Published Articles](#) (52,371) [CERN Preprints](#) (16,115)
[CERN Theses](#) (3,318) [CERN Reports](#) (1,114) [Committee Documents](#) (22,545)
- CERN Series** (15,924)
[CERN Annual Reports](#) (2) [CERN Yellow Reports](#) (1,130)
[CERN Theory](#) (12,510) [Academic Training Lectures](#) (615)
[Summer Student Lectures](#) (616) [General Talks](#) (1,212)
- CERN Departments** (75,937)
[Accelerator Technology \(AT\)](#) (5,185) [Accelerators & Technology Sector](#) (19,260) [Beams Department \(BE\)](#) (427) [Engineering Department \(EN\)](#) (147) [Finance \(FI\)](#) (1,154) [Human Resources \(HR\)](#) (170) [Information Technology \(IT\)](#) (4,337) [Physics \(PH\)](#) (38,419) [Secretariat-General \(SG\)](#) (11,028) [Technical Support \(TS\)](#) (1,386) [Technology Department \(TE\)](#) (100)
- CERN Experiments** (22,435)
[Fixed Target Experiments](#) (118) [LEP Experiments](#) (5,545) [LHC Experiments](#) (16,228) [Recognized Experiments](#) (552)
- CERN R&D Projects** (931)
[CERN Accelerator R&D Projects](#) (931)

CDS: Search for Books

Invenio
Technology

Tibor Šimko

Introduction

Digital Library

Invenio

Case Studies

Episode 1: Python

Episode 2: Git

Episode 3: Test Suite

Episode 4: Building
Efficient Indexes

Episode 5:

Load-balancing

Conclusions

Search:

[Search Tips](#) :: [Advanced Search](#)

Search collections:

Sort:

Display results:

Output format:

Books

2 records found

Search took 0.10 seconds.



Python Cookbook . - 2nd ed. / [Martelli, Alex](#)

Beijing : O'Reilly, 2005. - 807 p.

[Purchase from CERN Bookshop](#) - [CERN library copies](#)

[This book
at Amazon](#)

[Detailed record](#) - [Similar records](#)



Python Cookbook / [Martelli, Alex](#) (ed.); [Ascher, David](#) (ed.)

Beijing : O'Reilly, 2002. - 574 p.

[CERN library copies](#)

[This book
at Amazon](#)

[Detailed record](#) - [Similar records](#)

CDS: Search for Photos

Invenio
Technology

Tibor Šimko

Introduction

Digital Library

Invenio

Case Studies

Episode 1: Python

Episode 2: Git

Episode 3: Test Suite

Episode 4: Building
Efficient Indexes

Episode 5:

Load-balancing

Conclusions

Photos

Search:

[Search Tips](#) :: [Advanced Search](#)

Search collections:

Sort by:

Display results:

Output format:

Photos

178 records found

1 - 12

▶▶ jump to record:

Search took 0.23 seconds.



Photos : 178 records found

1 - 12

▶▶ jump to record:

CDS Features: Commenting

Invenio
Technology

Tibor Šimko

Introduction

Digital Library

Invenio

Case Studies

Episode 1: Python

Episode 2: Git

Episode 3: Test Suite

Episode 4: Building
Efficient Indexes

Episode 5:
Load-balancing

Conclusions

SCIENCE

Search

Submit

Personalize

Help

Home > Articles & Preprints > Articles > Detailed record #74 > Comments

Record 74

[\(Back to search results\)](#)

Quasinormal modes of Reissner-Nordstrom Anti-de Sitter Black Holes / [Wang, B.](#); [Lin, C.Y.](#)

[Abdalla, E.](#) [[hep-th/0003295](#)]

Complex frequencies associated with quasinormal modes for large Reissner-Nordstrom Anti-de Sitter black holes have been computed. [...]

<http://documents.cern.ch/cgi-bin/setlink?base=preprint&categ=hep-th&id=0003295>

[Detailed record](#) - [Similar records](#)

Comment

There is a total of 5 comments

[Write a comment](#)

[acmir](#) wrote on *09 Jan 2006, 09:48*

My comment

[Reply](#) | [Report abuse](#)

[acmir](#) wrote on *11 Jan 2006, 16:02*

admin wrote on 10 Jan 2006, 09:48:

My comment

no!

[Reply](#) | [Report abuse](#)

[acmir](#) wrote on *11 Jan 2006, 16:03*

admin wrote on 11 Jan 2006, 16:02:

My comment

no!

[Reply](#) | [Report abuse](#)

admin wrote on 10 Jan 2006, 09:48:

My comment

indeed

Invenio Features: Reviewing

Invenio
Technology

Tibor Šimko

Introduction

Digital Library

Invenio

Case Studies

Episode 1: Python

Episode 2: Git

Episode 3: Test Suite

Episode 4: Building
Efficient Indexes

Episode 5:
Load-balancing

Conclusions

People who viewed this page also viewed:

(3) [The Feynman lectures on physics](#) - [Feynman, Richard Phillips et al](#)

(3) [Learning Windows server 2003 2nd ed. :](#) - [Hassell, Jonathan](#)

(2) [With the unveiling of its new sign, the CERN Control Centre was officially inaugurated on Thursday 16 March.](#) - [T-UDS-AVC Team](#) - CERN-VIDEOCLIP-2006-08

(2) [Liability hedging and portfolio choice](#) - [Scherer, Bernd](#)

(2) [Conduite de projet Web2e éd. :](#) - [Bcrdaqe, Stephane](#)

Rate this document:

Average review score: ★★★★★ based on 1 reviews

Readers found the following reviews to be most helpful.

★★★★★ **A wonderful (and fun) guide to Common Lisp**

Reviewed by [zsj](#) on 14 Nov 2006, 17:48

0 out of 0 people found this review useful

(Test.) I've been recommending this text to people who want to start learning Common Lisp since it was first available in draft form on the author's web site. Now that it's out in print I can enthusiastically recommend that anybody who is interested in learning Common Lisp - or even curious about how the language can improve your productivity - purchase it.

Peter has a very enjoyable and easy-to-understand writing style, and he starts early with practical examples that show how Common Lisp can be used to solve problems. Chapter 3, "A Simple Database", is a great explanation of how programs are grown from pieces in Common Lisp to solve large problems. It's presented early and draws people in to the problem solving techniques used when programming in Lisp.

[Report abuse](#)

Was this review helpful? [Yes](#) / [No](#)

CDS: Create Personal Alert

Invenio
Technology

Tibor Šimko

Introduction

Digital Library

Invenio

Case Studies

Episode 1: Python

Episode 2: Git

Episode 3: Test Suite

Episode 4: Building
Efficient Indexes

Episode 5:

Load-balancing

Conclusions

Search:

neutrino mixing

any field

Search

Browse

[Search Tips](#) :: [Advanced Search](#)

Results overview: Found **4,236** records in 0.07 seconds.

[Articles & Preprints](#), **4,193** records found

[Books & Proceedings](#), **19** records found

[Presentations & Talks](#), **11** records found

[Multimedia & Outreach](#), **13** records found

1. **Constraining sterile neutrinos with a low energy beta-beam** / [Agarwalla, Sanjib Kumar](#)

Task hep-ex

We study the possibility to use a low energy beta-beam facility to search for sterile neutrinos by measuring the disappearance of electron anti-neutrinos. This channel is particularly sensitive since it allows to use inverse beta decay as detection reaction; thus it is free from hadronic uncertainties, provided the neutrino energy is below the pion production threshold. [...]

arXiv:1006.1640; VPI-IPNAS-10-10.- 2010 - Published in : Published in AIP Conf.Proc.: 1222 (2010) , pp. 169-173 [Preprint](#)

[Detailed record](#) - [Similar records](#)

Interested in being notified about new results for this query?

Set up a personal [email alert](#) or subscribe to the [RSS feed](#).

CDS: Add to Personal Basket

Invenio
Technology

Tibor Šimko

Introduction

Digital Library

Invenio

Case Studies

Episode 1: Python

Episode 2: Git

Episode 3: Test Suite

Episode 4: Building
Efficient Indexes

Episode 5:

Load-balancing

Conclusions

- 8. **Bridging flavour violation and leptogenesis in SU(3) family models** / [Calibbi, Lorenzo](#) (Max-Planck-Institut fuer Physik) ; [Chun, Eung Jin](#) (Korea Institute for Advanced Study)
We reconsider basic, in the sense of minimal field content, Pati-Salam x SU(3) family models which make use of the Type I see-saw mechanism to reproduce the observed mixing and mass spectrum in the neutrino sector. [...]
arXiv:1005.5563 ; KIAS-P10014 ; IC-2010-021 ; MPP-2010-58. - 2010.
[Preprint](#)
[Detailed record](#) - [Similar records](#)
- 9. **Rare muon and tau decays in A4 Models** / [Feruglio, Ferruccio](#) ; [Paris, Alessio](#)
We analyze the most general dimension six effective Lagrangian, invariant under the flavour symmetry $A_4 \times Z_3 \times U(1)$ proposed to reproduce the near tri-bimaximal lepton mixing observed in neutrino oscillations. [...]
arXiv:1005.5526 ; DFPD-10-TH-9. - 2010.
[Preprint](#)
[Detailed record](#) - [Similar records](#)
- 10. **Quark and lepton mixing angles with a dodeca-symmetry** / [Kim, Jihn E](#) ; [Seo, Min-Seok](#)
The discrete symmetry D_{12} at the electroweak scale is used to fix the quark and lepton mixing angles. [...]
arXiv:1005.4684. - 2010.
[Preprint](#)
[Detailed record](#) - [Similar records](#)

ADD TO BASKET

CDS: Display Personal Basket

Invenio
Technology

Tibor Šimko

Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python
Episode 2: Git
Episode 3: Test Suite
Episode 4: Building
Efficient Indexes
Episode 5:
Load-balancing

Conclusions

[Home](#) > [Your Account](#) > [Your Baskets](#) > [Personal baskets](#) > [Physics](#) > [Standard Model](#)

Display baskets

Personal baskets > **Physics**

[Back to Your Baskets](#)

[Create basket](#)

[Edit topic](#)

Standard Model (3)

Standard Model

3 items, 2 notes

last update: 11 Jun 2010, 14:21

[Edit basket](#)

[Delete basket](#)

1. **Non-Abelian Flat Directions in a Minimal Superstring Standard Model** / [Cleaver, G B](#) ; [Faraggi, A E](#) ; [Nanopoulos, Dimitri V](#) ; [Walker, J W](#) [ACT-2000-1] [CTP-TAMU-2000-2] [OUTP-2000-03-P] [TPI-MINN-2000-6] [hep-ph/0002060]

Recently, by studying exact flat directions of non-Abelian singlet fields, we demonstrated the existence of free fermionic heterotic-string models in which the $SU(3)_C \times SU(2)_L \times U(1)_Y$ -charged matter spectrum, just below the stringscale, consists solely of the MSSM spectrum. [...]

Published in **Mod. Phys. Lett. A: 15 (2000) pp. 1191-1202**

Fulltext: [PDF](#); [PS.GZ](#)

[Detailed record - Notes \(2\)](#)

[Copy item](#) [Remove item](#)

2. **Precise calculation of parity nonconservation in cesium and test of the standard model** / [Dzuba, V A](#) ; [Flambaum, V V](#) ; [Ginges, J S M](#) [hep-ph/0204134]

We have calculated the 6s-7s parity nonconserving (PNC) E1 transition amplitude, $E_{\{PNC\}}$, in cesium. [...]

Fulltext: [PDF](#); [PS.GZ](#)

[Detailed record - Add a note...](#)

[Copy item](#) [Remove item](#)

CDS: Organize and Share Your Baskets

Invenio
Technology

Tibor Šimko

Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python
Episode 2: Git
Episode 3: Test Suite
Episode 4: Building
Efficient Indexes
Episode 5:
Load-balancing

Conclusions

[Home](#) > [Your Account](#) > [Your Baskets](#) > [Personal baskets](#)

Display baskets

Personal baskets

Group baskets

Public baskets

Physics (1)

Standard Model

Programming (3)

Linux, Python, SQL

Search baskets for:

in

Search also in notes (where allowed)

CDS: Journals and Bulletins

Invenio
Technology

Tibor Šimko

Introduction

Digital Library

Invenio

Case Studies

Episode 1: Python

Episode 2: Git

Episode 3: Test Suite

Episode 4: Building
Efficient Indexes

Episode 5:

Load-balancing

Conclusions



The Bulletin

Archives | Contact us | Sign Up! | Staff Association | CERN Home

search

english | français

Issue No. 23-24/2010 - Monday 7 June 2010

News Articles

Official News

Training and Development

General Information

Staff Association

Lyn Evans decelerates!



people who built one of the most complex scientific instruments ever conceived by mankind. >>

After more than 40 years at CERN, 15 of which were dedicated to ensuring that the LHC comes to completion, Lyn Evans is retiring. The Imperial College Professor and recently-elected Fellow of the British Royal Society has set himself new challenges, but plans to keep strong links with CERN. His big thank you goes to the many hundreds of



What's New

News Articles

- o Lyn Evans decelerates!
- o Security needs you
- o New computer security campaign
- o A better beam quality
- o Uniting forces in physics and medicine
- o Neutrino oscillations make their first appearance in OPERA
- o It sounds good!
- o "Draw me a physicist" exhibition opens
- o Council Chamber exhibition
- o Irène Jacob visits CERN
- o News from the Library
- o Back to the 80s

Invenio Key Features

Invenio
Technology

Tibor Šimko

Introduction

Digital Library

Invenio

Case Studies

Episode 1: Python

Episode 2: Git

Episode 3: Test Suite

Episode 4: Building
Efficient Indexes

Episode 5:

Load-balancing

Conclusions

- **navigable collection tree** (regular, virtual)
- **powerful search engine**
 - Google-like speed for up to 5M records
 - combined metadata, reference and fulltext search
- **flexible metadata** (MARC, OA)
 - handling any kind of document (multimedia)
 - customizable input, formatting and linking
- **personalization** and **collaborative** features:
 - alerts, baskets, groups, reviews, comments
 - internationalization (26 languages)
- **open source**, GNU General Public License
 - co-developed by CERN (2002–), EPFL (2004–), DESY/FNAL/SLAC (2008–), CfA (2009–)
 - installed at ~30 institutions world-wide

Invenio Architecture: Overview

Invenio
Technology

Tibor Šimko

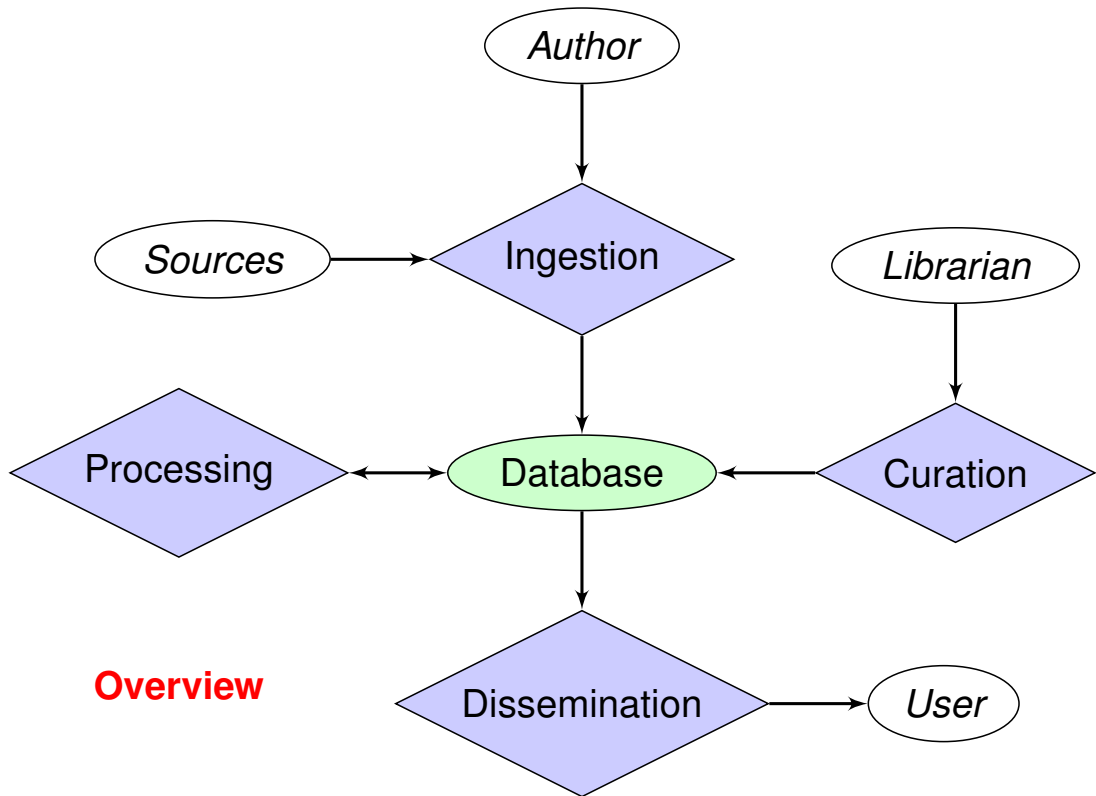
Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python
Episode 2: Git
Episode 3: Test Suite
Episode 4: Building
Efficient Indexes
Episode 5:
Load-balancing

Conclusions



Overview

Invenio Modules: Ingestion

Invenio
Technology

Tibor Šimko

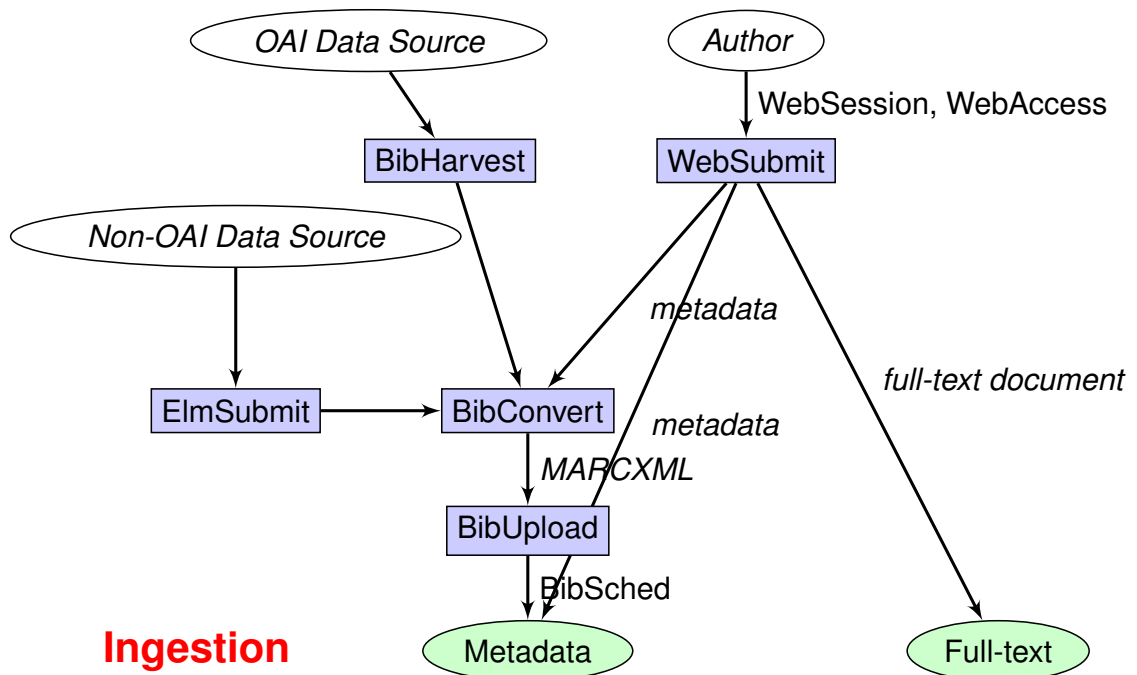
Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python
Episode 2: Git
Episode 3: Test Suite
Episode 4: Building
Efficient Indexes
Episode 5:
Load-balancing

Conclusions



Ingestion

Invenio Modules: Processing

Invenio Technology

Tibor Šimko

Introduction

Digital Library

Invenio

Case Studies

Episode 1: Python

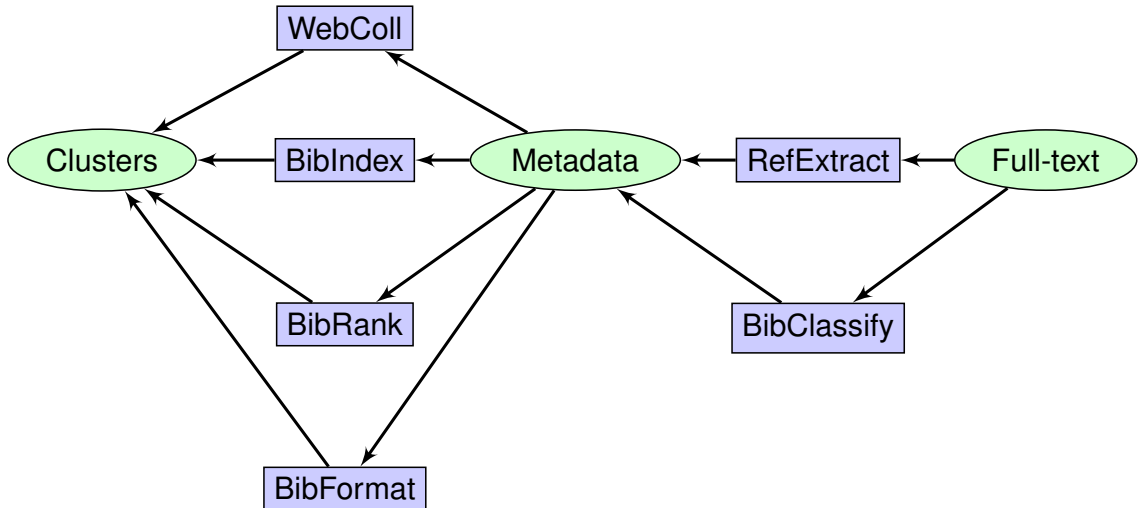
Episode 2: Git

Episode 3: Test Suite

Episode 4: Building Efficient Indexes

Episode 5: Load-balancing

Conclusions



Processing

Invenio Modules: Dissemination

Invenio Technology

Tibor Šimko

Introduction

Digital Library

Invenio

Case Studies

Episode 1: Python

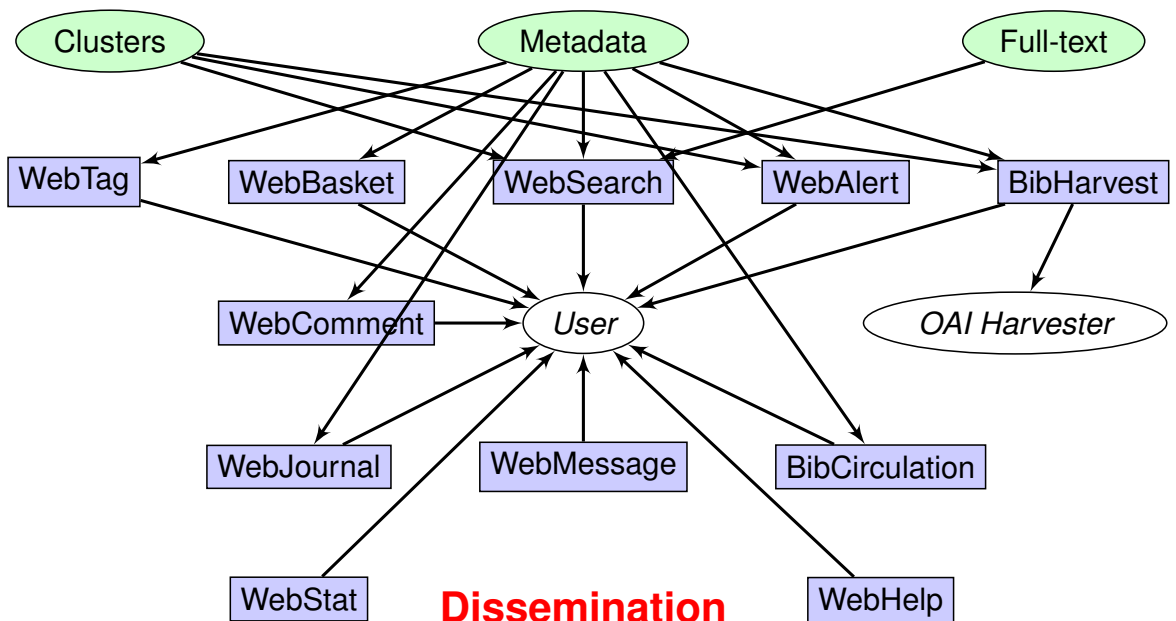
Episode 2: Git

Episode 3: Test Suite

Episode 4: Building Efficient Indexes

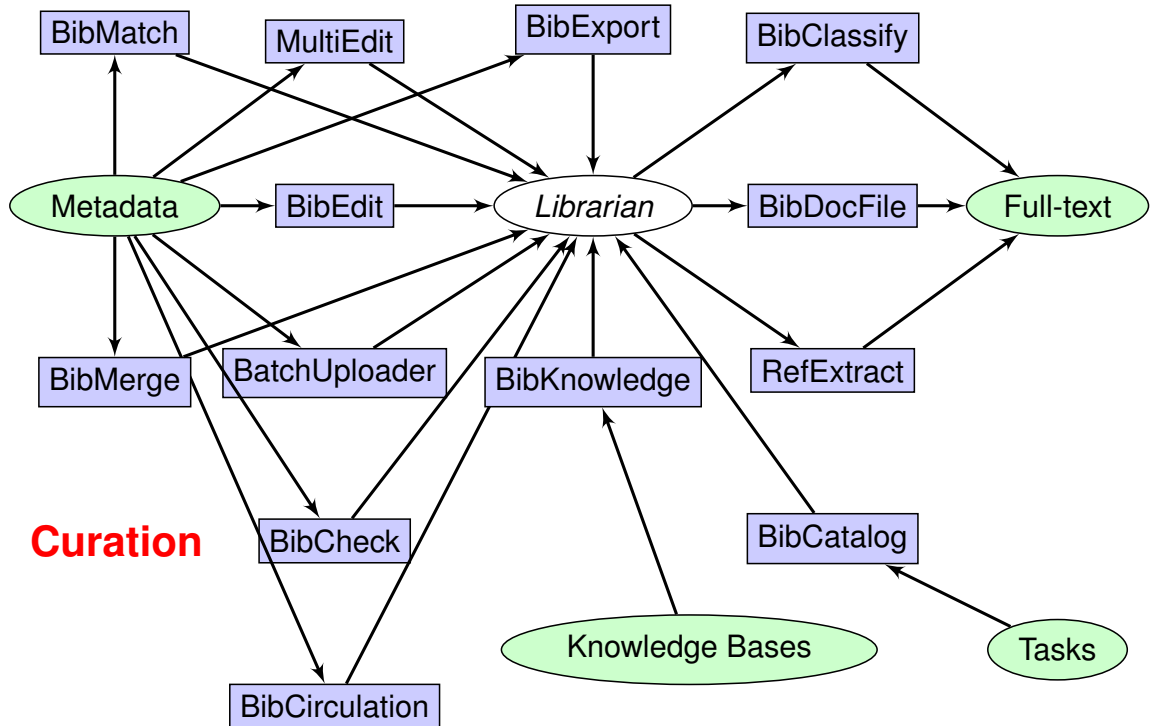
Episode 5: Load-balancing

Conclusions



Dissemination

Invenio Modules: Curation



Invenio Technology

Tibor Šimko

Introduction

Digital Library

Invenio

Case Studies

Episode 1: Python

Episode 2: Git

Episode 3: Test Suite

Episode 4: Building Efficient Indexes

Episode 5: Load-balancing

Conclusions

Invenio Modules: Summary

Invenio Technology

Tibor Šimko

Introduction

Digital Library

Invenio

Case Studies

Episode 1: Python

Episode 2: Git

Episode 3: Test Suite

Episode 4: Building Efficient Indexes

Episode 5: Load-balancing

Conclusions

- ~33 modules
- codebase
 - ~250,000 lines of Python code
 - ~10,000 lines of JavaScript code
 - ~6,000 lines of XSL code
 - ~5,000 lines of autotools code
- ~40 authors
 - many short-term students
 - importance of *informal* coding standards
- ~10 years of development
 - started at CERN, first release in 2002
 - now co-developed world-wide (EU, US)
- lego programming... but no silver bullet

Why Python?

Invenio
Technology

Tibor Šimko

Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python
Episode 2: Git
Episode 3: Test Suite
Episode 4: Building
Efficient Indexes
Episode 5:
Load-balancing

Conclusions

- easy to read and understand (good for many temporary developers)
- suitable for rapid prototyping (good for organic-growth software development model)
- write code to throw it away

Art of Ikebana

Invenio
Technology

Tibor Šimko

Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python
Episode 2: Git
Episode 3: Test Suite
Episode 4: Building
Efficient Indexes
Episode 5:
Load-balancing

Conclusions



- Japanese art of flower arrangement
- “way of flowers”
- natural shapes, graceful lines
- minimalism
- *“disciplined art form in which nature and humanity are brought together”*

Art of Ikebana Programming

Invenio
Technology

Tibor Šimko

Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python

Episode 2: Git

Episode 3: Test Suite

Episode 4: Building
Efficient Indexes

Episode 5:
Load-balancing

Conclusions

Java?

```
new Callable() {
    public Object call(Object x) {
        return x.times(k)
    }
}
```

Python!

```
lambda x: k * x
```

Speeding Up Python

Invenio
Technology

Tibor Šimko

Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python

Episode 2: Git

Episode 3: Test Suite

Episode 4: Building
Efficient Indexes

Episode 5:
Load-balancing

Conclusions

- bytecode interpreted language
- but **Cython** permits to write C extensions easily
- combining efficiency of C with high-levelness of Python

Example: intbitset.pyx

```
ctypedef unsigned long long int word_t

ctypedef struct IntBitSet:
    int size
    int allocated
    word_t trailing_bits
    int tot
    word_t *bitset
```

Why Git?

Invenio
Technology

Tibor Šimko

Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python
Episode 2: Git
Episode 3: Test Suite
Episode 4: Building
Efficient Indexes
Episode 5:
Load-balancing

Conclusions

- good for distributed teams
- offline development possible
- “pull on demand” collaboration model (as opposed to “shared push” collaboration model)
 - inherent, natural code review process
- commit early, commit often (to private repositories)
- rebase and clean (before pushing for public consumption)
- interplay with SVN

Git Branches

Invenio
Technology

Tibor Šimko

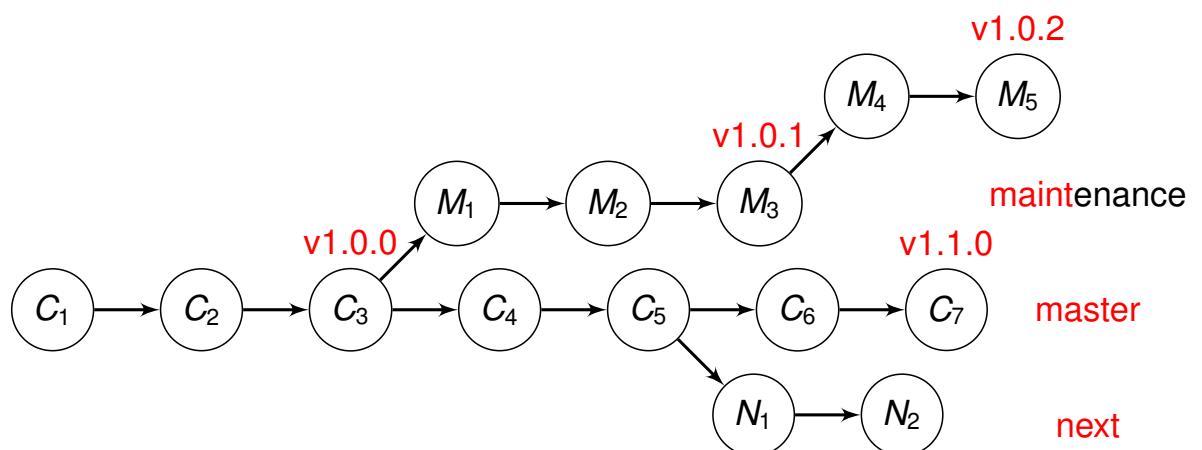
Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python
Episode 2: Git
Episode 3: Test Suite
Episode 4: Building
Efficient Indexes
Episode 5:
Load-balancing

Conclusions



- **maint** — release maintenance branch
- **master** — new feature branch
- **next** — things not yet release-ready

Git Development

Invenio Technology

Tibor Šimko

Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python

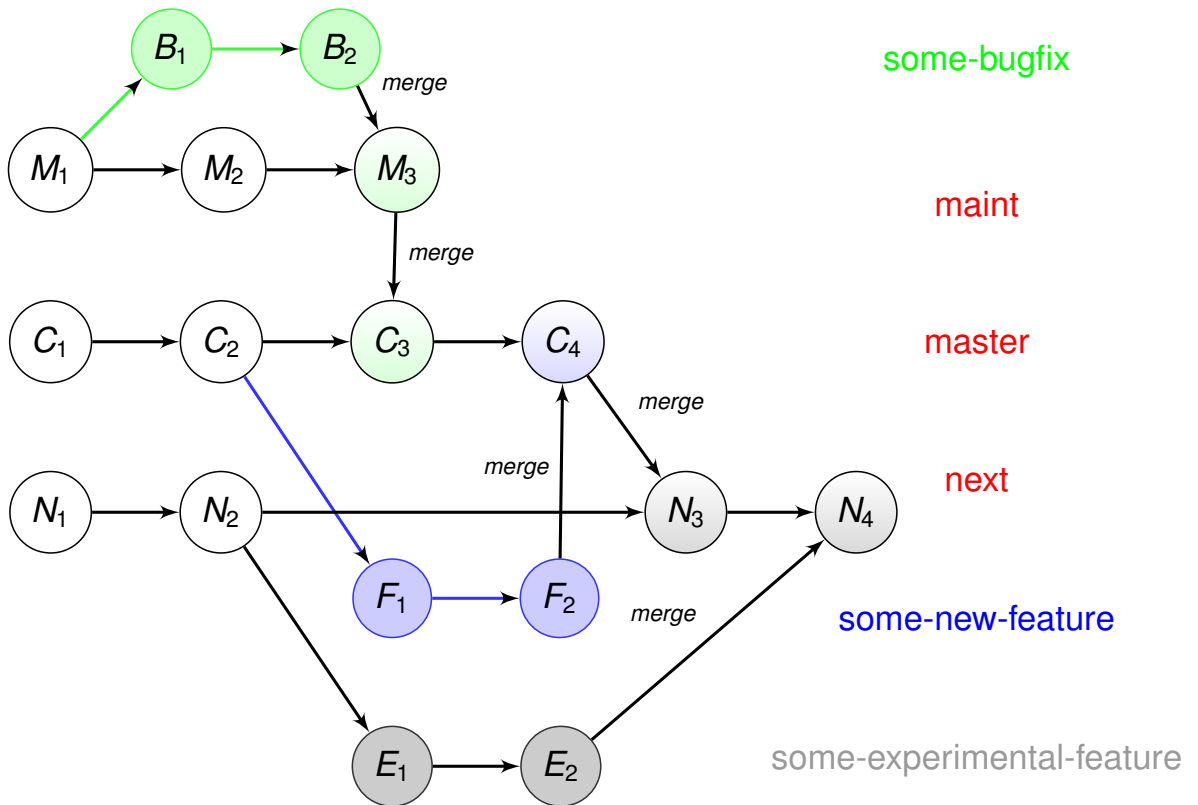
Episode 2: Git

Episode 3: Test Suite

Episode 4: Building Efficient Indexes

Episode 5: Load-balancing

Conclusions



Git collaboration model

Invenio Technology

Tibor Šimko

Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python

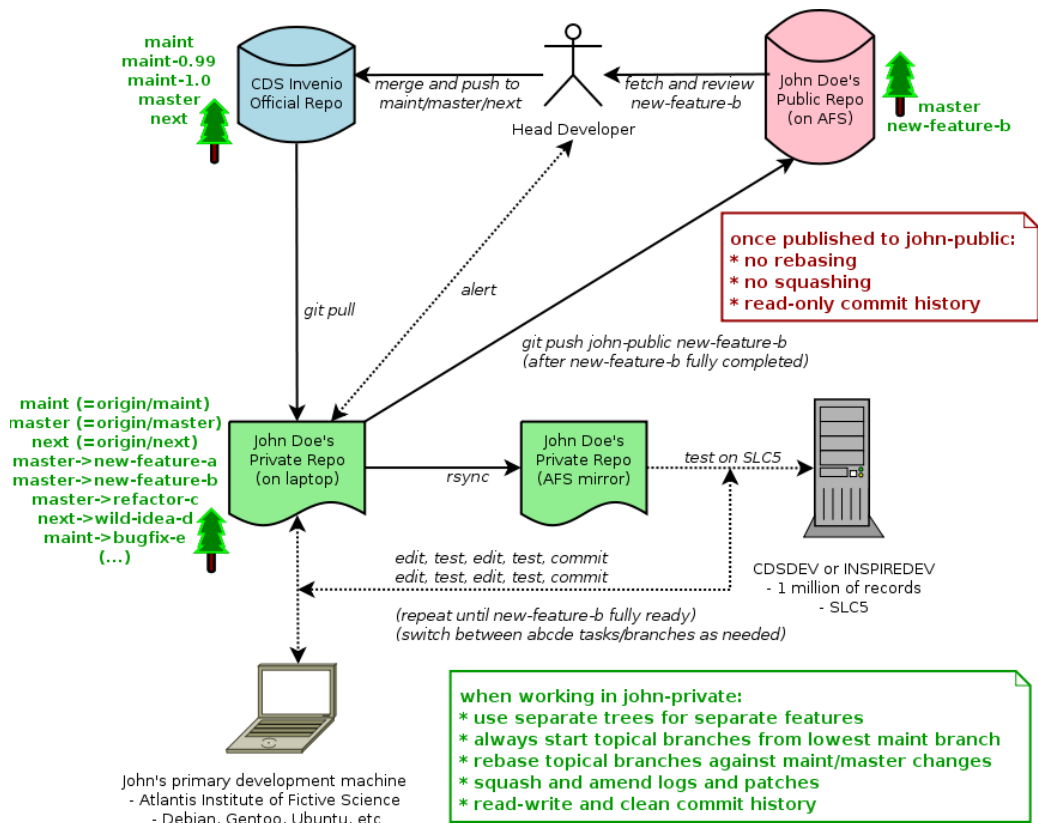
Episode 2: Git

Episode 3: Test Suite

Episode 4: Building Efficient Indexes

Episode 5: Load-balancing

Conclusions



Web testing

Invenio
Technology

Tibor Šimko

Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python
Episode 2: Git
Episode 3: Test Suite
Episode 4: Building
Efficient Indexes
Episode 5:
Load-balancing

Conclusions

- sometimes we need to run tests in real browser
 - e.g. pages with heavy JavaScript
- using **Selenium IDE** extension for Firefox
 - record and replay browser actions
 - test for text existence or non-existence on pages
 - test for link labels and targets

Example: test_search_ellis.html

```
<tr><td>open</td>
  <td>http://localhost</td>
  <td></td>                                </tr>
<tr><td>type</td>
  <td>p</td>
  <td>ellis</td>                            </tr>
<tr><td>clickAndWait</td>
  <td>action_search</td>
  <td></td>                                </tr>
<tr><td>verifyTextPresent</td>
  <td>1. Thermal conductivity of dense quark matter and cooling of stars</td>
  <td></td>                                </tr>
```

Designing A Search Engine

Invenio
Technology

Tibor Šimko

Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python
Episode 2: Git
Episode 3: Test Suite
Episode 4: Building
Efficient Indexes
Episode 5:
Load-balancing

Conclusions

- **performance-driven design** assumptions:
 - high number of selects, low number of updates
 - fast searching, slow indexation
 - cache everything cacheable
- **search functionality:**
 - search for words, phrases, regular expressions
 - search in any field, authors, titles, etc
- **index design:**
 - forward indexes: $word1 \rightarrow [rec1, rec2, \dots]$
 $word2 \rightarrow [rec2, rec7, \dots]$
 - reverse indexes: $rec1 \rightarrow [word1, word8, \dots]$
 $rec2 \rightarrow [word1, word2, \dots]$
- **Zipf's law** on word frequency:
 - few words occur very often (e.g. *the*)
 - most words are infrequent (even e.g. *boson*)

Search Engine Under Cover

Invenio
Technology

Tibor Šimko

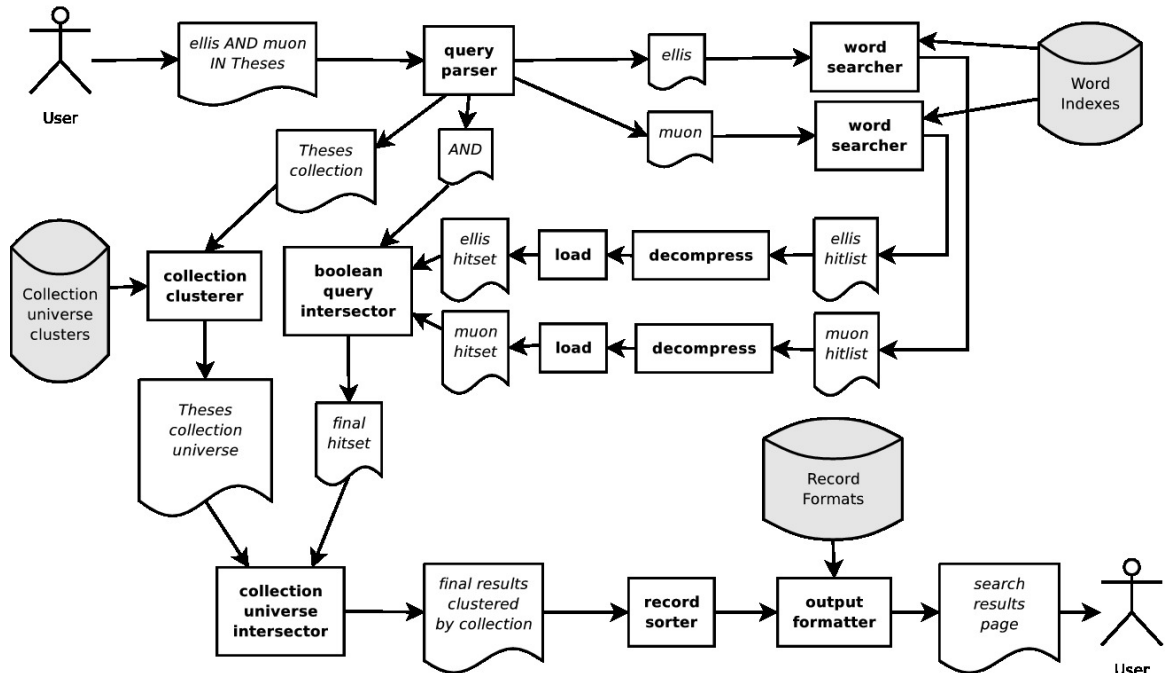
Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python
Episode 2: Git
Episode 3: Test Suite
Episode 4: Building
Efficient Indexes
Episode 5:
Load-balancing

Conclusions



Measuring the Performance

Invenio
Technology

Tibor Šimko

Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python
Episode 2: Git
Episode 3: Test Suite
Episode 4: Building
Efficient Indexes
Episode 5:
Load-balancing

Conclusions

- three important **speed factors** to consider:
 - speed of finding sets (DB Server)
 - speed of demarshaling sets (DB ↔ Web App Server)
 - speed of intersecting sets (Web App Server)

Example: speed of various parts (2002, before optimization)

action / query:	"CERN 2002"	"of the this"
fetching	0.28 sec	0.34 sec
demarshaling	0.78 sec	1.10 sec
adding colls	0.37 sec	0.63 sec
intersecting	0.64 sec	1.19 sec
total search time	2.07 sec	3.22 sec

Optimizing Data Structures

Invenio
Technology

Tibor Šimko

Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python
Episode 2: Git
Episode 3: Test Suite
Episode 4: Building
Efficient Indexes
Episode 5:
Load-balancing

Conclusions

- **data structures** tested:
 - 'sorted' (lists, Patricia trees)
 - 'unsorted' (hashed sets, binary vectors)
- **fast prototyping**: (Python, Lisp in 2002)
 - throw-away coding to test ideas

Example: lists vs dicts, 350K sets in 800K universe

```
marshaling lists ..... 532616+532571 bytes in 1.33 sec
demarshaling lists ... 350000+350000 items in 0.10 sec
merging lists ..... 546965 items in 0.34 sec
intersecting lists ... 153035 items in 0.35 sec

marshaling dicts ..... 576491+576450 bytes in 0.87 sec
demarshaling dicts ... 350000+350000 items in 0.36 sec
merging dicts ..... 546965 items in 0.09 sec
intersecting dicts ... 153035 items in 0.15 sec
```

... and the winner is:

Invenio
Technology

Tibor Šimko

Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python
Episode 2: Git
Episode 3: Test Suite
Episode 4: Building
Efficient Indexes
Episode 5:
Load-balancing

Conclusions

- **binary vectors** found the best compromise!
 - using Numeric Python module (in 2002)
 - typical search time gain: 4.0 sec → 0.2 sec (in 2002)
 - typical indexing time loss: 7 hours → 4 days (in 2002)
 - mostly spare data modelled via mostly dense data structure?
 - free your mind, think critically
- further optimization:
 - Numeric module not addressing real bits, only bytes
 - so home-made `intbitset` C extension in 2007
 - addressing real bits (factor of 8 already)
 - saving space, saving (indexing) time

Splitting Web App Server and DB Server

Invenio
Technology

Tibor Šimko

Introduction

Digital Library

Invenio

Case Studies

Episode 1: Python

Episode 2: Git

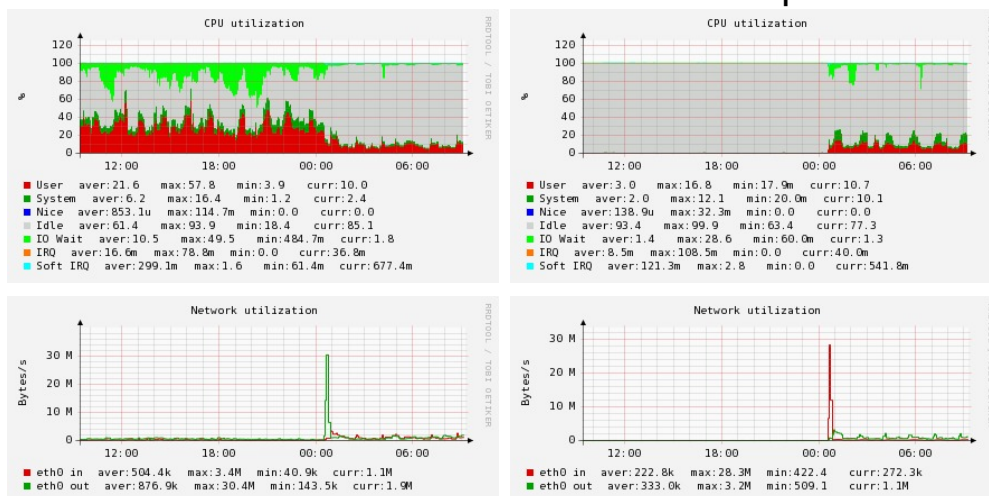
Episode 3: Test Suite

Episode 4: Building
Efficient Indexes

Episode 5:
Load-balancing

Conclusions

- load of CDS Web and DB servers at the split time:



web + db → web

idle → db

- split leads to efficient use of OS resources by lone, non-competing Web and DB daemon processes

Load-Balanced Setup

Invenio
Technology

Tibor Šimko

Introduction

Digital Library

Invenio

Case Studies

Episode 1: Python

Episode 2: Git

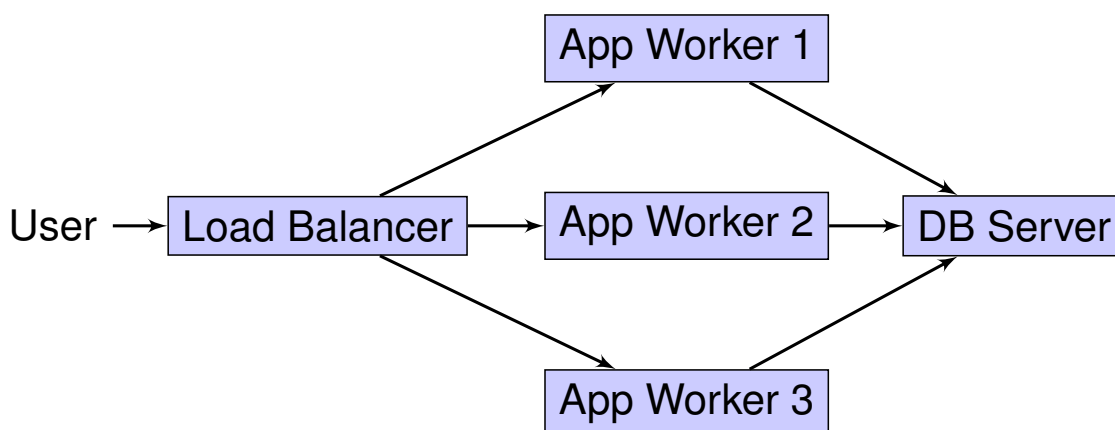
Episode 3: Test Suite

Episode 4: Building
Efficient Indexes

Episode 5:
Load-balancing

Conclusions

- useful for “LHC First Beam Day” rush situations with many concurrent visitors
- Apache `mod_proxy_balancer`



Measuring Scalability

Invenio
Technology

Tibor Šimko

Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python
Episode 2: Git
Episode 3: Test Suite
Episode 4: Building
Efficient Indexes
Episode 5:
Load-balancing

Conclusions

- using **siege** to simulate concurrent users and to measure throughput on a sample of typical URLs

Example: inspirebeta.net under gentle siege

```
$ siege -d 1 -c 20 -t 1m -f inspirebeta_urls.txt
Transactions:          1329 hits
Availability:          100.00 %
Elapsed time:           60.23 secs
Data transferred:     37.12 MB
Response time:         0.41 secs
Transaction rate:      22.07 trans/sec
Throughput:            0.62 MB/sec
Concurrency:           8.96
Successful transactions: 1329
Failed transactions:   0
Longest transaction:   3.05
Shortest transaction:  0.01
```

Conclusions

Invenio
Technology

Tibor Šimko

Introduction

Digital Library
Invenio

Case Studies

Episode 1: Python
Episode 2: Git
Episode 3: Test Suite
Episode 4: Building
Efficient Indexes
Episode 5:
Load-balancing

Conclusions

- building Invenio digital library system
 - ~250,000 LOCs from ~40 authors over ~10 years
- value of rapid prototyping
- value of organic-growth software development model
- value of coding aesthetics and minimalism
- morale from selected anecdotes?
 - *“Never Lose A Holy Curiosity”* (A. Einstein)