Universidad Rey Juan Carlos

GSyC LibreSoft WE STUDY LIBRE SOFTWARE

Measuring Etch:

The Size of Debian 4.0

Debian Conference Debconf7
Edinburgh, June 17th 2007

Juan José Amor, Gregorio Robles,
Jesús M. González-Barahona and Javier Fernández-Sanguino Peña
Grupo GSyC/Libresoft – Universidad Rey Juan Carlos
{jjamor,grex,jgb,jfs}@gsyc.escet.urjc.es

Universidad Rey Juan Carlos

GSyC LibreSoft WE STUDY LIBRE SOFTWARE

- GSyC/LibreSOFT @ URJC: Research group in «Libre Software Engineering».
- About 20 researchers involved in more than 10 projects.

Universidad Rey Juan Carlos

GSyC LibreSoft WE STUDY LIBRE SOFTWARE

What is this talk about?

- Lies, damn lies, statistics
  - Some nice graphs too!
- Rehash of Debconf5 talk, updated for Etch
- Give an overview of Debian's ¿sustainable? growth

Summary

- Debian 4.0 (etch)
  - Counting source lines of code (SLOC)
  - Counting packages, files, programming languages
- Effort estimations (Basic COCOMO)
- Comparison with other OS
  - And previous releases
- Future lines of work

About Debian

- A Libre Software Operating System
- Based in libre kernels (Linux, HURD, *BSD) and libre applications (GNU tools, Mozilla, etc)
- Created and maintained by volunteers
- Debian Social Contract
- Debian policies
- Several contemporary Debian versions (stable, testing, unstable, experimental)

Universidad Rey Juan Carlos

GSyC LibreSoft WE STUDY LIBRE SOFTWARE

Studying Debian: Methodology

- Methodology: download, unpack, count, sum
- Results: physical source lines of code (SLOCs)
- Identifying programming languages
- Identifying identical files (using md5)
- Estimations using classical software engineering techniques (Basic COCOMO)

Issues

- Binary code in packages
  - Solved by uncompressing ALL contents within a source package
- Duplicated code: across packages, different releases, forked versions...
  - E.g. gcc-{2.95,3.3,3.4,4.1,4.2}
- Patch within Patch
  - Eg: yada, cdbs...

Size of Debian Etch

- Count of upstream packages:
  - 260,000,000 SLOC
- Count of Debian source packages:
  - 283,000,000 SLOC
- Debian src packages without *debian* directory:
  - 277,000,000 SLOC

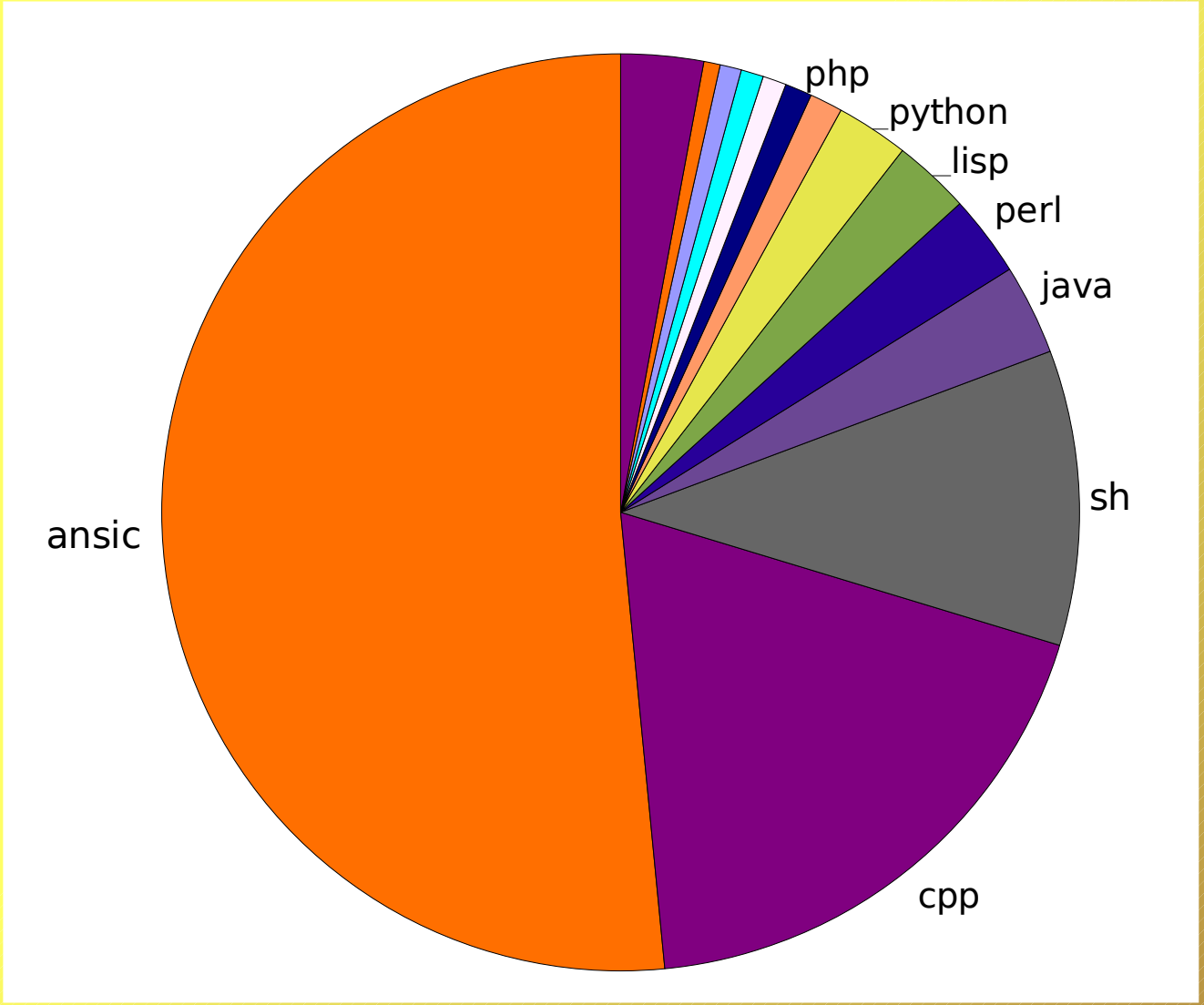
Question #1

Name the 5 most used programming languages in the Debian OS.

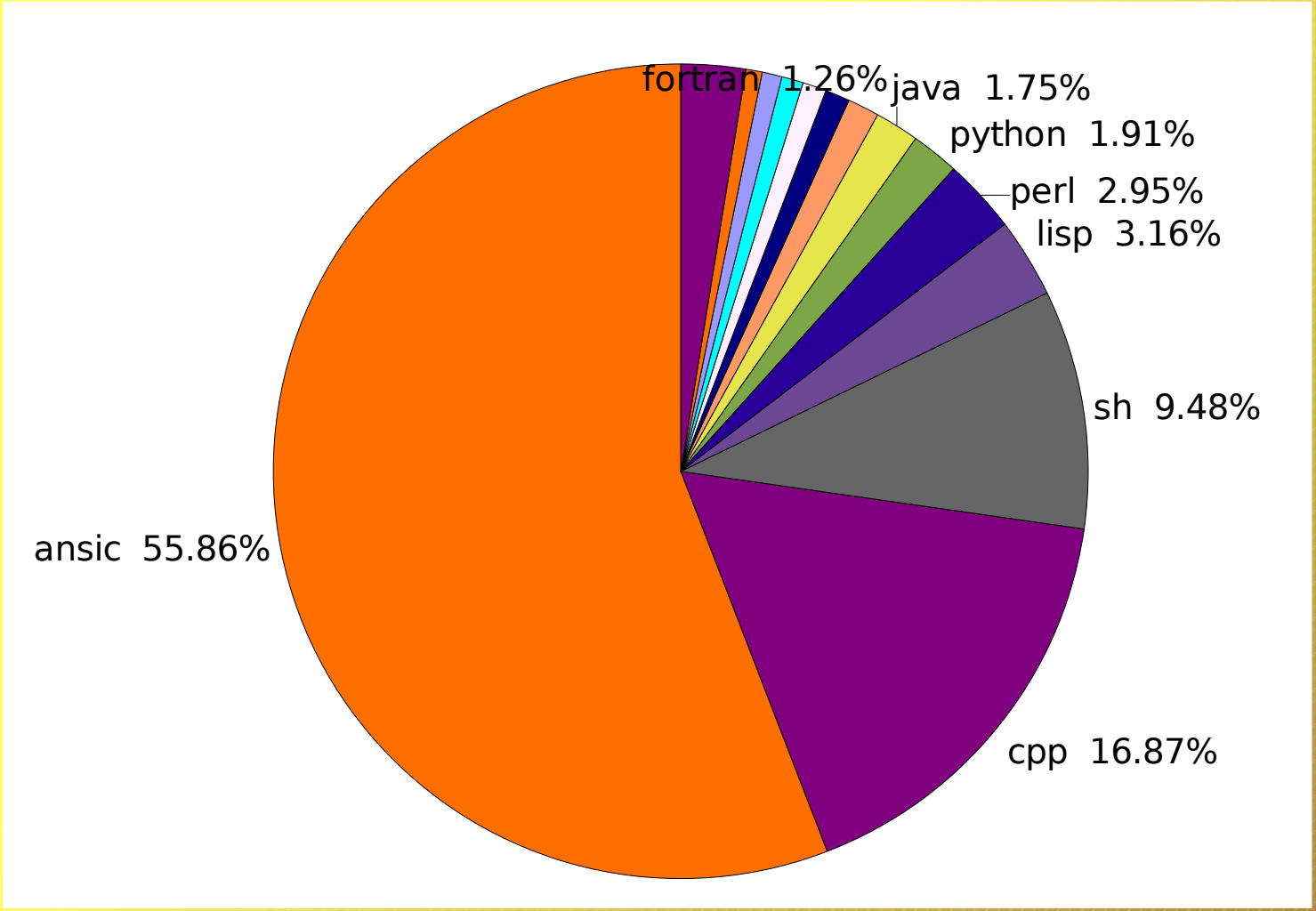
Programming Languages

LANGUAGE	SLOCs	PERCENT
C	145,278,000	51,00%
C++	52,983,000	18,70%
Shell	29,327,000	10,40%
Java	8,969,000	3,17%
PERL	8,074,000	2,85%
LISP	7,659,000	2,70%
Python	7,219,000	2,55%
Assembler	4,121,000	1,46%
PHP	3,270,000	1,15%
FORTRAN	2,678,000	0,95%
C#	2,336,000	0,83%
Pascal	2,240,000	0,79%
TCL	1,635,000	0,58%

Programming Languages



Programming Languages (Sarge)



Programming languages (thoughts)

- ANSI C usage going down consistently
- C++ usage increasing
- Java:
  - Was new in Sarge
  - Explosive growth (x2,4), replaces Lisp in rank
  - More SLOCs that even Perl!

Question #2

Name the 2 largest packages in the Debian OS.

Universidad Rey Juan Carlos

GSyC LibreSoft WE STUDY LIBRE SOFTWARE

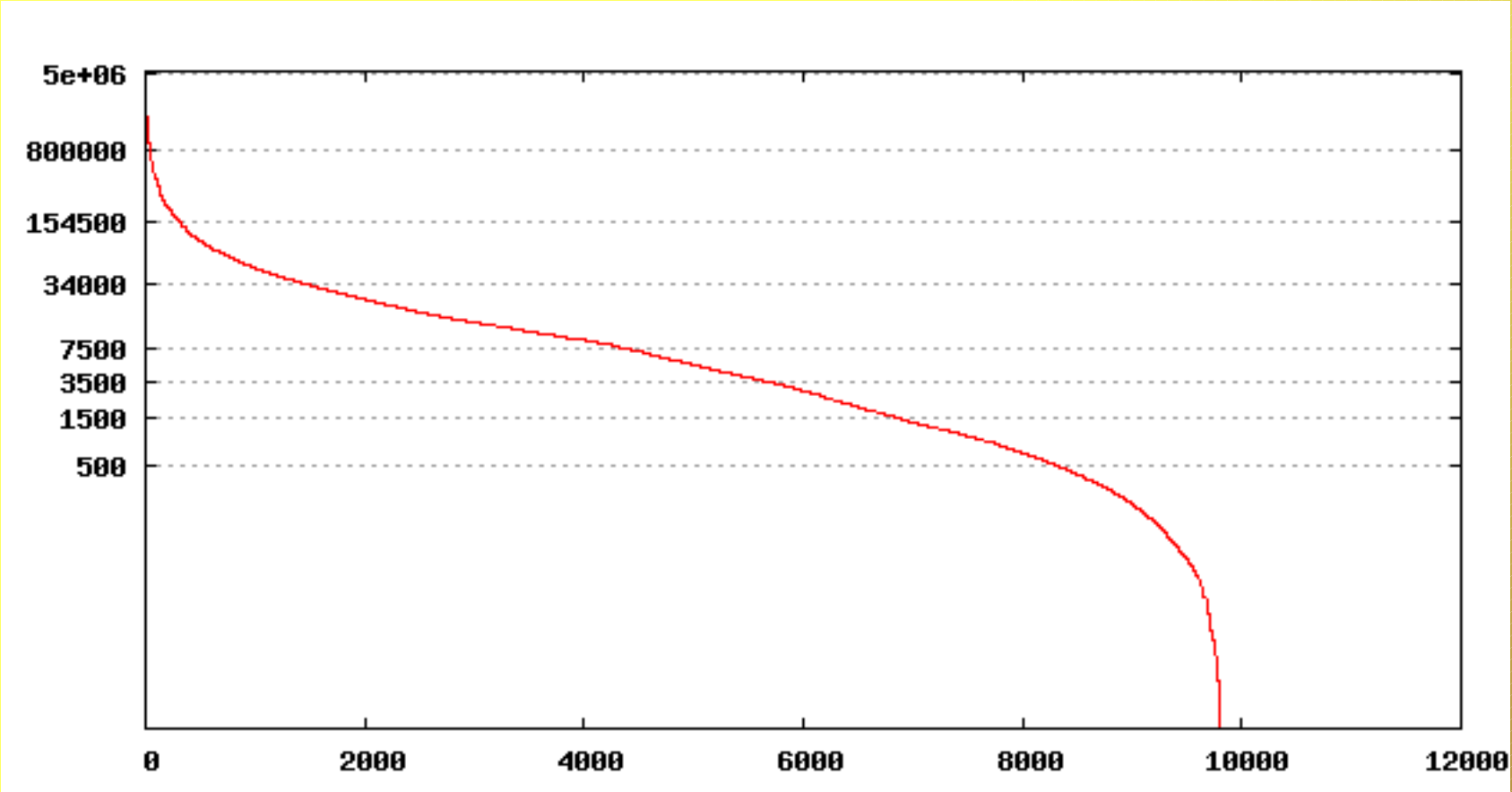
Top 10 packages in Debian 4.0

Rank	Package	Version	Size	Files
#1	Openoffice.org	2.0.4.dfsg.2	5,215,000	16,503
#2	Linux2.6	2.6.18.dfsg.1	4,921,000	17,215
#3	ia32-libs	1.19	4,006,000	13,108
#4	gcc-4.1	4.4.1ds2	3,630,000	26,584
#5	iceweasel	2.0.0.3	2,776,923	11,570
#6	icedove	1.5.0.10.dfsg.1	2,709,000	11,477
#7	vnc4	4.4.1+X4.3.0	2,357,000	7,312
#8	eclipse	3.2.1	2,214,000	15,807
#9	stalin	0.11	1,885,000	284
#10	mono	1.2.2.1	1,766,000	13,569

Top 10 packages (thoughts)

- End-user software, development tools and kernels make up the Top 10 list
- Similar to Woody's
  - Watch out for KfreeBSD! (#11)
- Contribution of Top 100 packages to overall size keeps going down
  - 65% in 2.0 to 34% in 4.0

Package sizes



Average size: 28,000 SLOC for more than 10,000 packages

Question #3

How much it would take (cost, time...) to write the Debian OS from scratch?

COCOMO

- COCOMO is an effort estimation technique used in 'classical' software engineering [Boehm81]
- Effort in Debian 4.0:
  - Estimated effort: 73,400 person-years
  - Scheduled time: nearly 9 years
  - Estimated cost: more than 5,300 Million € (6.7 **b**illion USD)

Universidad Rey Juan Carlos

GSyC LibreSoft WE STUDY LIBRE SOFTWARE

Debian and other OS

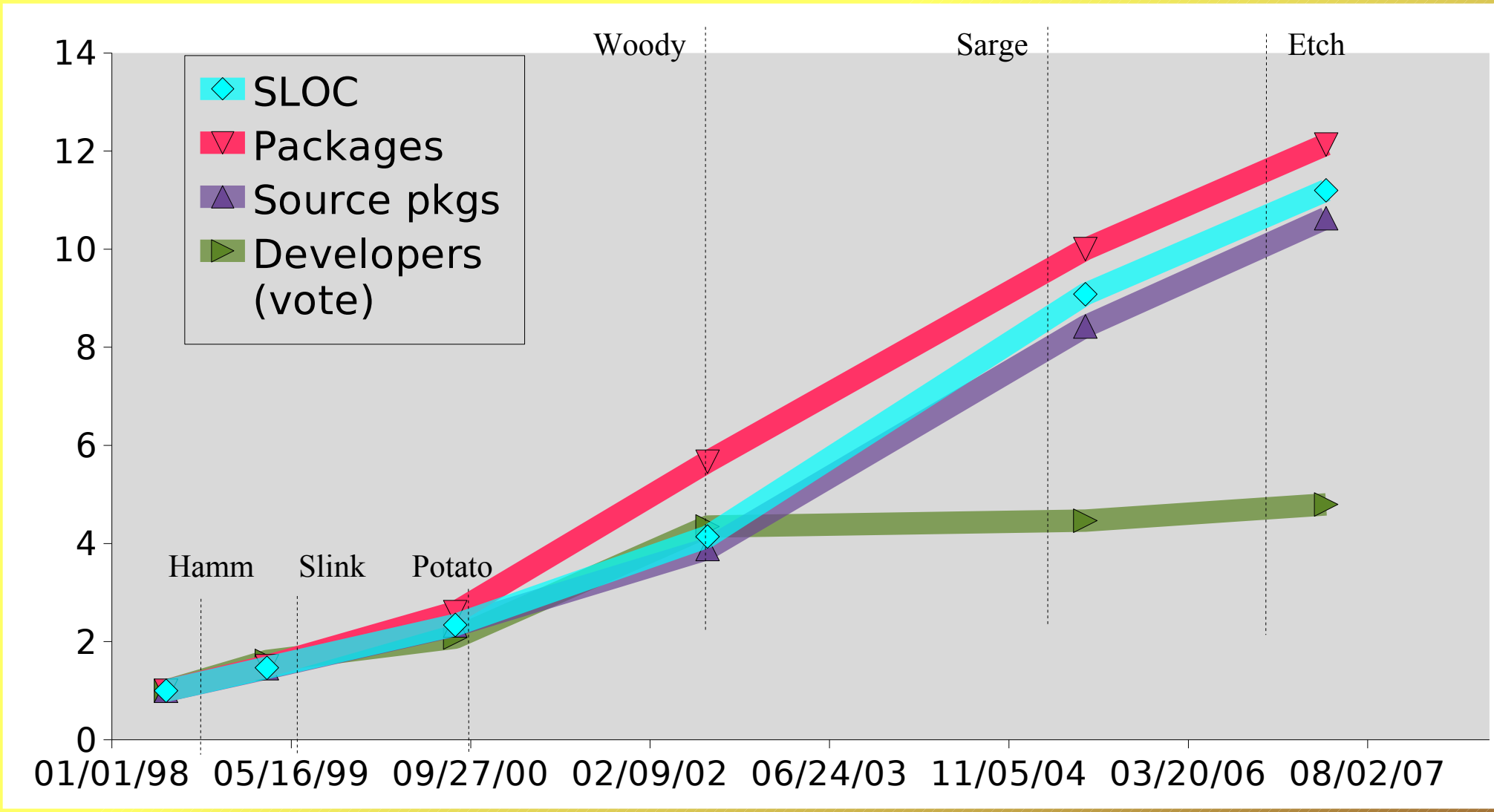
OS/Apps	Date	MSLOCs (*)
Windows 3.1	April 92	3
Windows 95	August 95	15
Windows NT 4.0	July 96	16
Debian 2.0	July 98	25
Solaris 7	Oct 98	7,5
Debian 2.1	March 99	37
Windows 2000	Feb 00	29
Windows XP	2002	40
Debian 2.2	August 00	55
Debian 3.0	July 02	105
Fedora Core 4	May 05	76
Debian 3.1	June 05	229
OpenSolaris	June 05	4,6
Debian 4.0	April 07	283

(*) MSLOCs for non-libre-software are uncertain

Etch vs. older releases

- Debian 4.0 'etch' SLOCs is:
  - 11 x hamm (~9 years ago)
  - 7.63 x slink (~8 years ago)
  - 4.79 x potato (~6,5 years ago)
  - 2.7 x woody(~5 years ago)
  - 1.23 x sarge (~2 years ago)
- But number of DDs does not grow as fast

Debian growth (compared)



Future work

- Analysis of authorship / licenses
  - Was done for Sarge already
- More in-depth analysis comparing releases
  - Understand how Debian is growing
- Analise volunteer activity
  - Mailing lists, package uploads, CVS / SVN usage ...

Universidad Rey Juan Carlos

GSyC LibreSoft WE STUDY LIBRE SOFTWARE

Demo!

http://debian-counting.libresoft.es/

Conclusions

- Debian 4.0 (etch), released in April 2007 has
  - 283,000,000 SLOC.
  - Estimated cost of 5,400 Million € and needs a estimated work of nearly 9 years.
  - C, C++, Shell scripts, Java, PERL and LISP are the most used languages
- Debian 4.0 (probably) still represents the biggest FLOSS compilation.

Tools

- Main source code analysis:
  - Done with Sloccount.
    - http://www.dwheeler.com/sloccount/
    - aptitude install sloccount
- Copyright analysis:
  - (will be) done with pyTernity.
    - http://forja.linex.org/projects/pyternity (available soon)
- More tools:
  - http://libresoft.urjc.es/

References

- Libre Software Engineering website at URJC
  - http://libresoft.urjc.es
- Debian count results (from Debian 2.0 to 4.0):
  - http://libresoft.urjc.es/debian-counting/
- COCOMO:
  - [Boehm81]: Book 'Software Engineering Economics', Prentice Hall