OpenSourceWatershed

"watching software flow downstream"

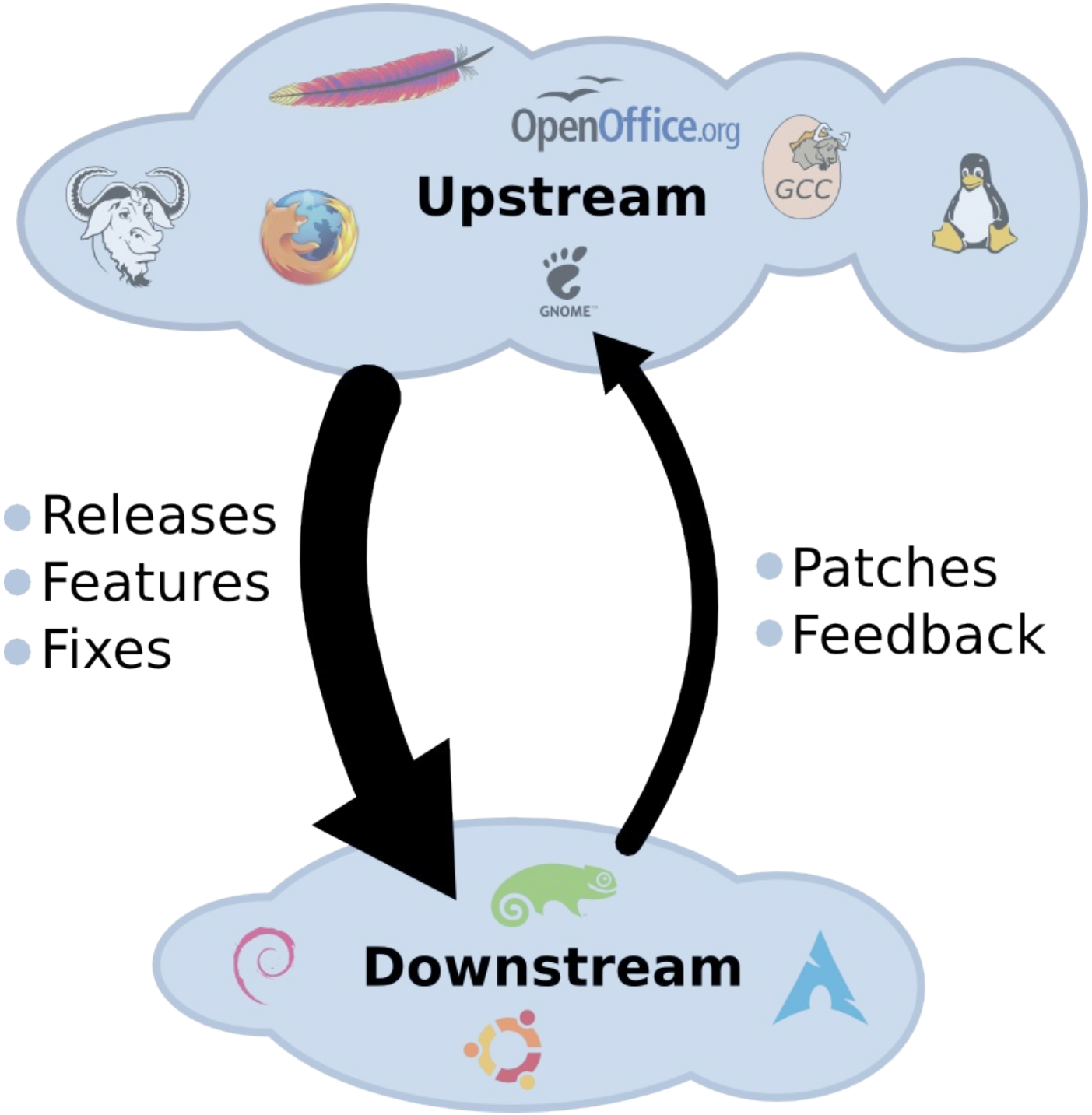
http://oswatershed.org

Scott Shawcroft
July 22, 2009

Scott Shawcroft

- Class of 2009 ~ University of Washington
- Computer Engineer
- Creative Commons, Google and more Google
- OS Projects: touchd, Menzies, Annoamp, denu
- Linux since spring '04.
- LFS → Gentoo → Ubuntu → Gentoo

Watershed


OpenOffice.org
Upstream
GCC
GNOME
Releases
Features
Fixes
Patches
Feedback
Downstream

Study of

distrology dis·trol·o·gy dĭ-strŏl-ə-jē

The formal study of open source software distributions.

Data Gathered

Gather release information from upstream and downstream.

- **Name**
- **Version**
- **Date**
- Revision

Data Sources

Upstream

- **Directory Listings**
- **Sourceforge**

Distributions/Repositories

- **Name**
- **Codename**
- **Component**
- **Architecture**

Branches

- **Experimental**
- **Future**
- **Current**
- **LTS**
- **Past**

Results

Upstream/Downstream relationship metrics:

- **% Obsolete**
- **# Obsolete**
- **Lag**

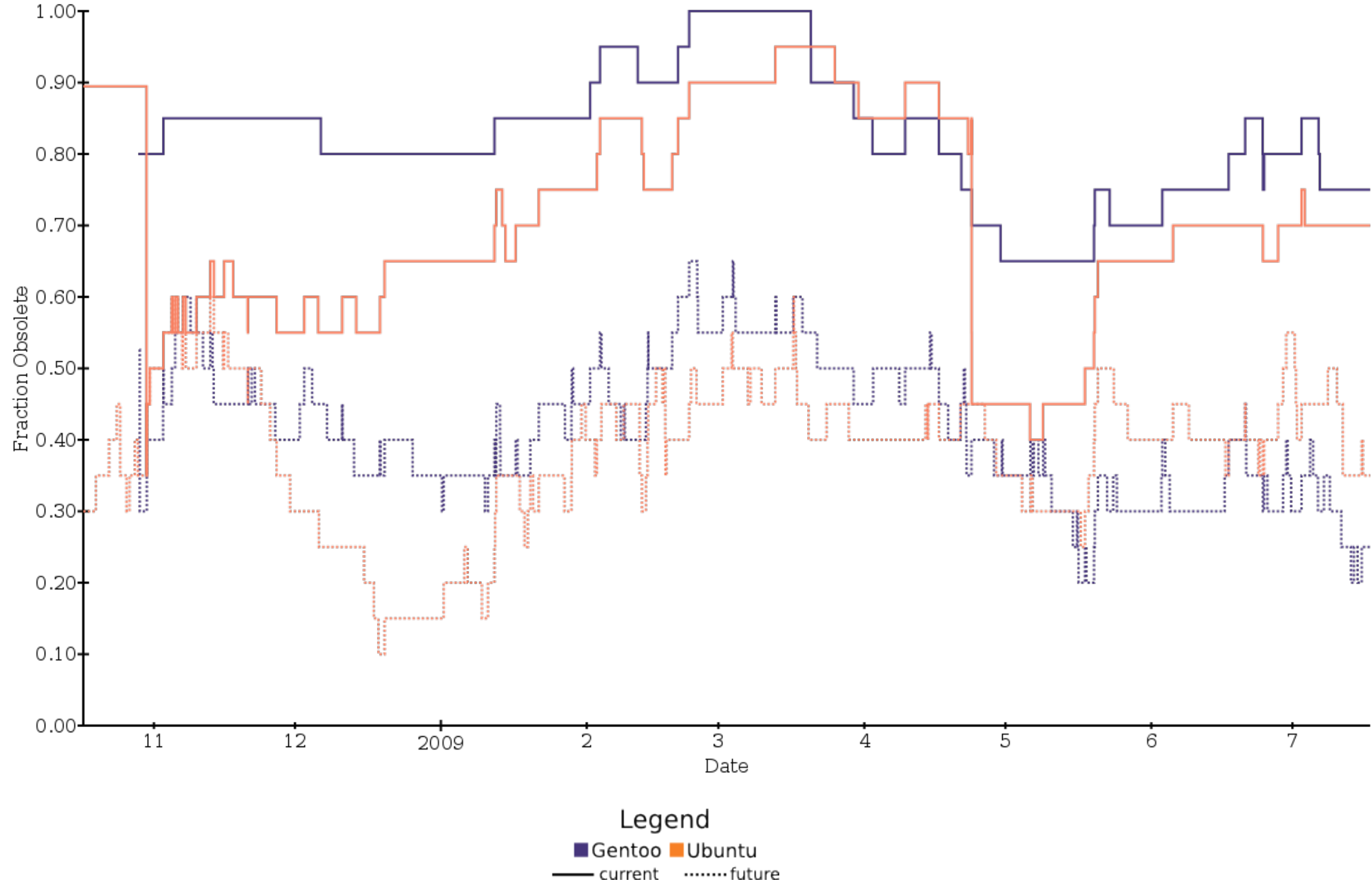
oswatershed.org

- **Per Package Data (badges)**
- **Per Distro Data**
- **Different Group Analysis**
- **Data Quality Tools**

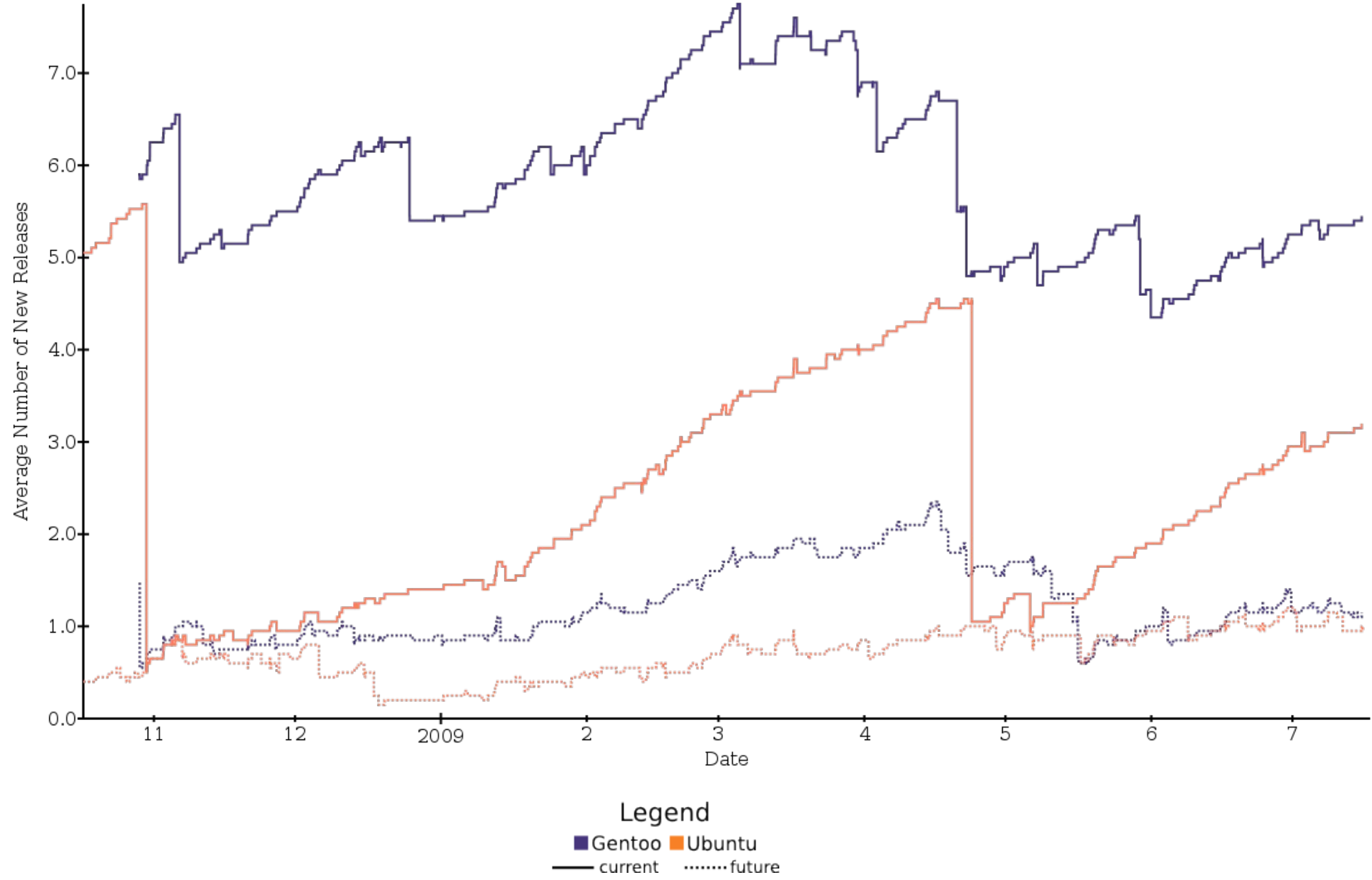
Scott's Chosen 20

alsa-utils
cups
emacs
firefox
gcc
ghostscript-gpl
gimp
glibc
gnome-desktop
gnupg
httpd (apache)
kdebase
linux
NetworkManager
openssh
pidgin
postgresql
python
ruby
xorg-server

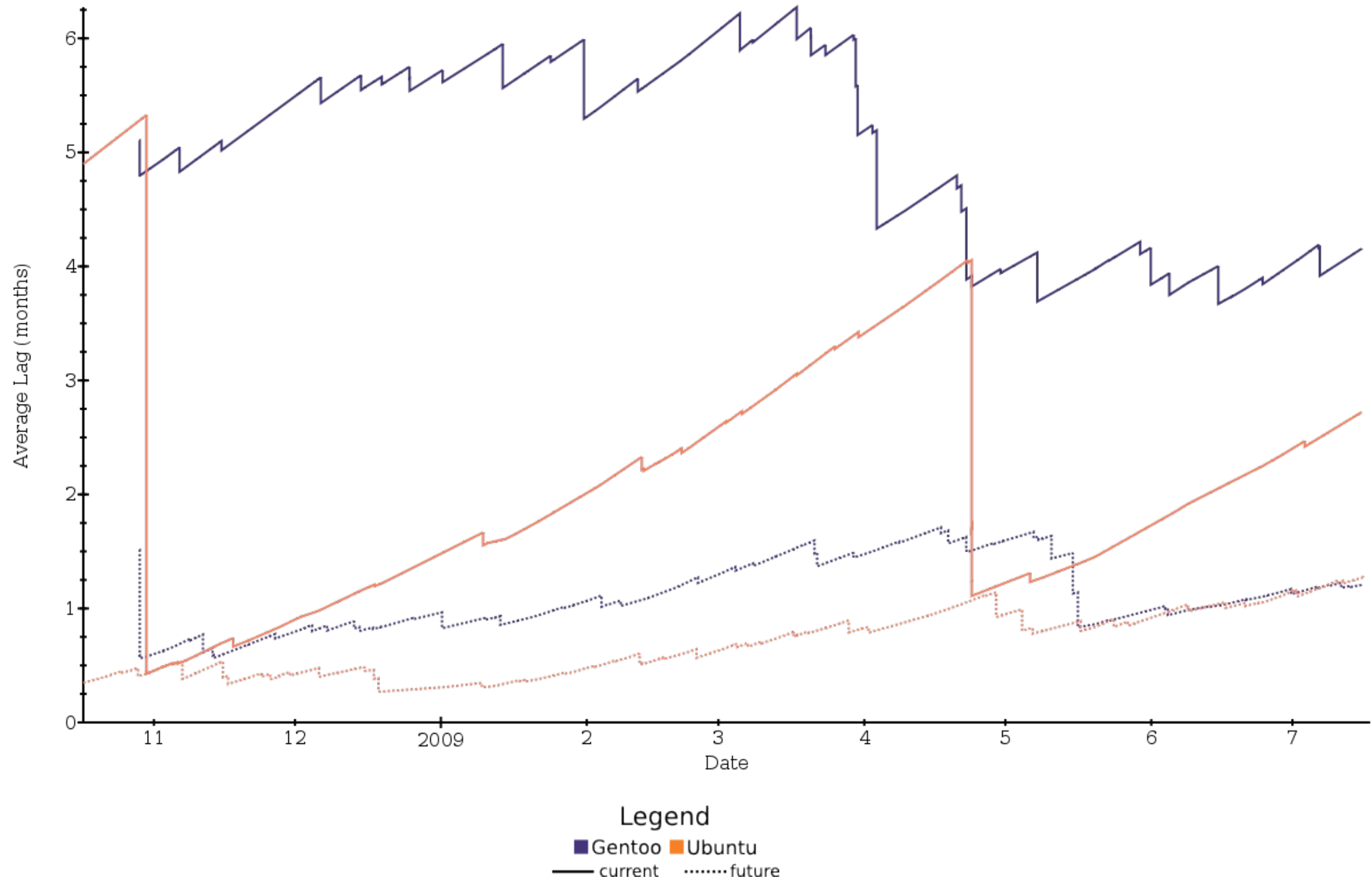
Ubuntu/Gentoo (% obsolete)



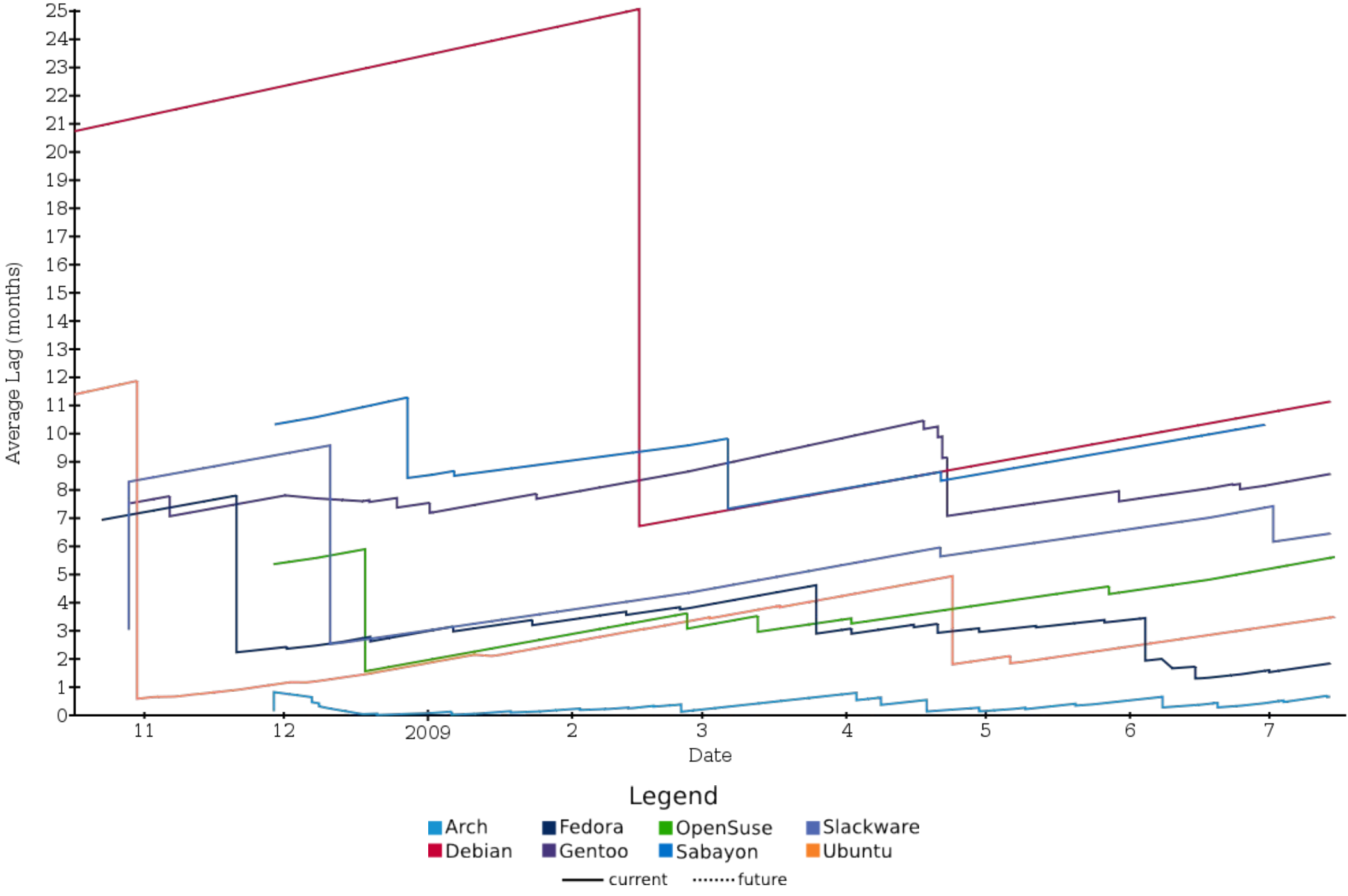
Ubuntu/Gentoo (# obsolete)



Ubuntu/Gentoo (lag)



LAMPPP (lag)


Average Lag (months)
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
11
12
2009
2
3
4
5
6
7
Date
Legend
Arch
Fedora
OpenSuse
Slackware
Debian
Gentoo
Sabayon
Ubuntu
current
future

LAMPPP (lag)

OSW linux **2.6.31-rc3**	2.6.30.1	2.6.26	2.6.30	2.6.30	2.6.30	2.6.27.23	--	2.6.27.7	2.6.28.14.18
OSW httpd **2.2.11**	**2.2.11**	2.2.9	**2.2.11**	**2.2.11**	**2.2.11**	2.2.10	**2.2.11**	**2.2.11**	**2.2.11**
OSW mysql **5.4.1-beta**	5.1.36	5.0.51a	5.1.35	5.0.70	5.0.72	5.0.67	5.0.76	5.0.67	5.1.31
OSW python **3.1**	2.6.2	2.5.2	2.6	2.5.4	2.5.4	3.0b3	2.5.4	2.5.2	3.0~rc3
OSW php **5.3.0**	5.2.10	5.2.6	5.2.10	5.2.10	5.2.10	5.2.9	5.2.8	5.2.10	5.2.6
OSW perl **5.8.9**	**5.10.0**	**5.10.0**	**5.10.0**	**5.10.0**	5.8.8	**5.10.0**	5.8.8	**5.10.0**	**5.10.0**

Challenges

- Lots of data.

- Comparing it all.
  - Normalizing names.
  - Determining obsoletion.
    (aka understanding versions)

Lots of Data

- **9** Distributions
  - Each has its own custom crawl script.

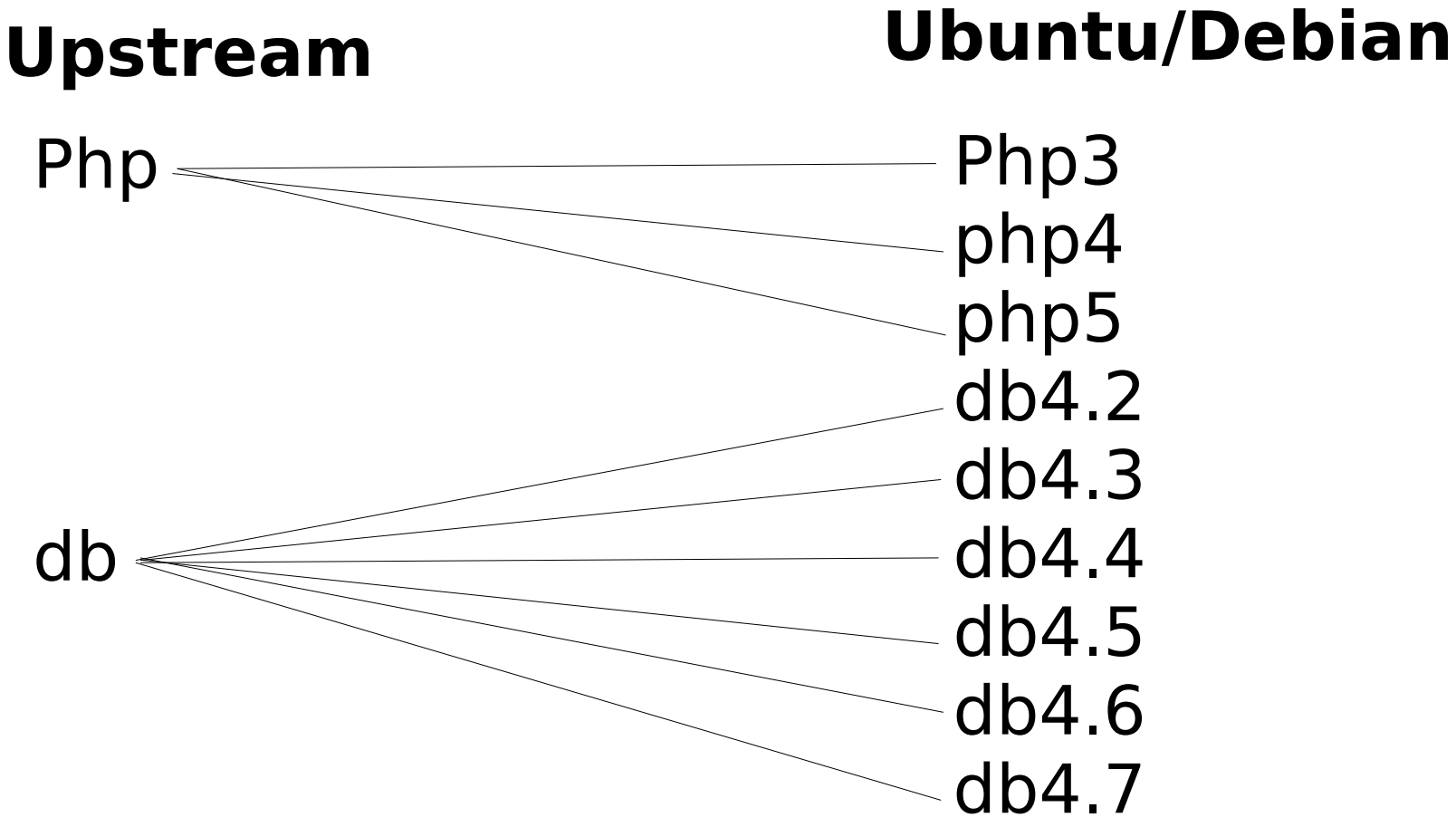
- **78,476** Total Packages
  - Mostly inflated by custom distro names. 10K – 15k estimated distinct.

- **735,859** Releases
  - Distinct package name and version combinations. Skewed by different naming.

- **2,463** Upstream Packages
  - **78** Sourceforge Sources
  - **106** Directory Sources
  - **3** Custom Scripts

Normalizing Names

Distros must also deal with package branches. Gentoo uses 'slotting', most use new package names.



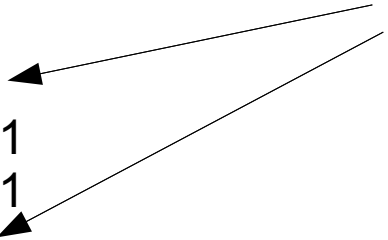
Ordering Versions

Original ordering based only on release date.

Problems:

- All new releases obsolete old ones.
- Any new downstream release that doesn't match an upstream release is completely fresh.

2008-10-02 20:24:00	2.6
2008-11-07 04:30:00	3.0rc2
2008-11-21 02:50:00	3.0rc3
2008-12-03 20:37:00	3.0
2008-12-05 05:57:00	2.6.1
2008-12-13 14:43:00	2.4.6c1
2008-12-13 16:47:00	2.5.3c1
2008-12-19 16:14:00	2.4.6
2008-12-19 16:15:00	2.5.3
2008-12-23 14:28:00	2.5.4
2009-02-14 01:10:00	3.0.1

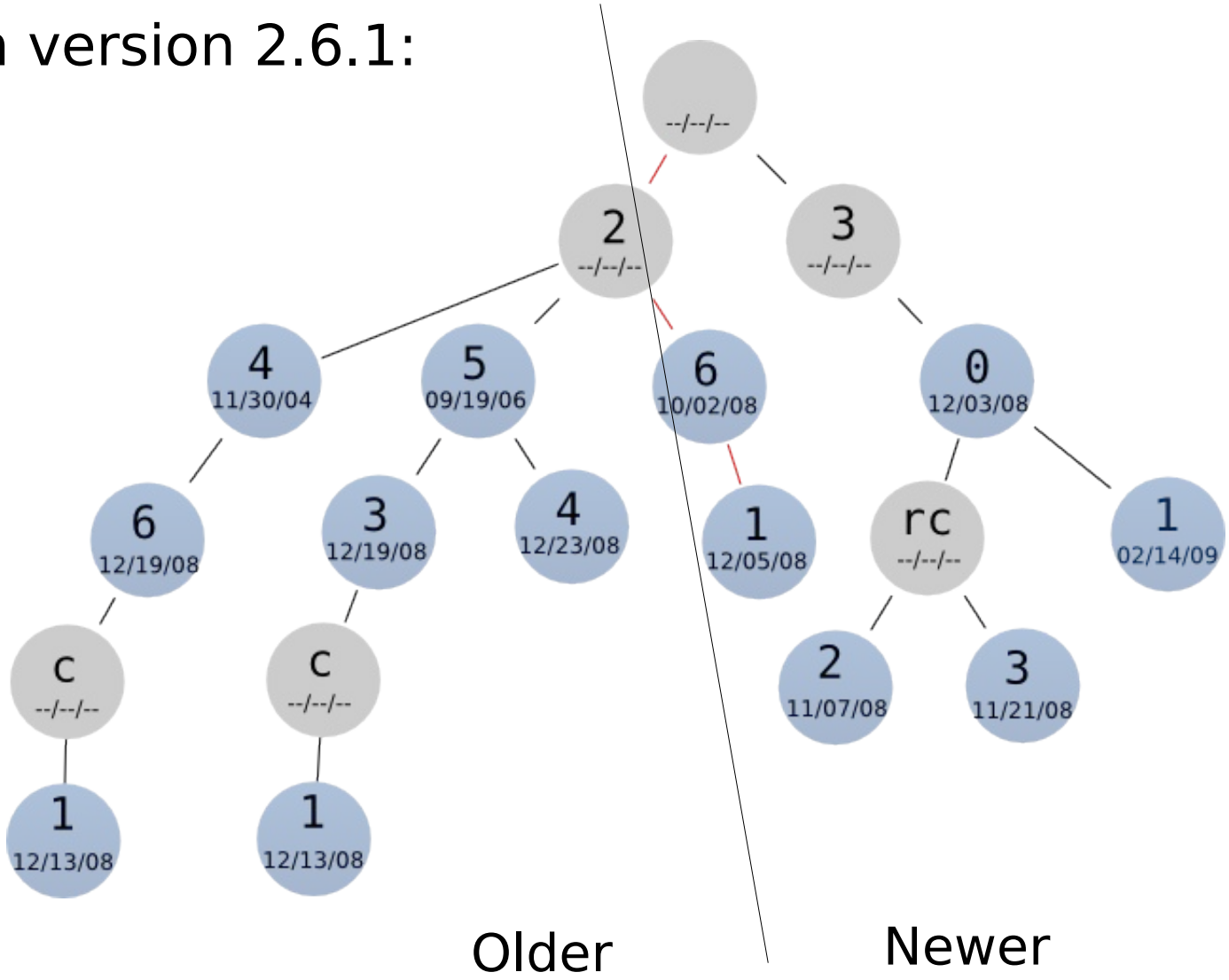


Should we obsolete 2.6.1 with 2.4.6?

Ordering Versions

1) Split the version.
2) Build a tree with children sorted by release date.

For Python version 2.6.1:



Conclusions

Release cycle does not effect overall freshness.

Package management includes many hacks.

- Many downstream to one upstream.
  - Libpng → libpng, libpng-dev
- Mangling package names. (Slotting)
  - Php → php3, php4, php5
- Mangling version numbers.
  - Mysql 5.1.30really5.0.83

OSW Future

Need volunteers and supporters!

- User centric features brainstormed.
  - Add sources.
  - Verify sources.
  - Link packages.
  - Custom package groups.
- More eyes on all of the data and code.
- Publicity! Online articles, blogs and badges.

Links

- scott.shawcroft @ gmail.com
- oswatershed.org
- github.com/tannewt/open-source-watershed

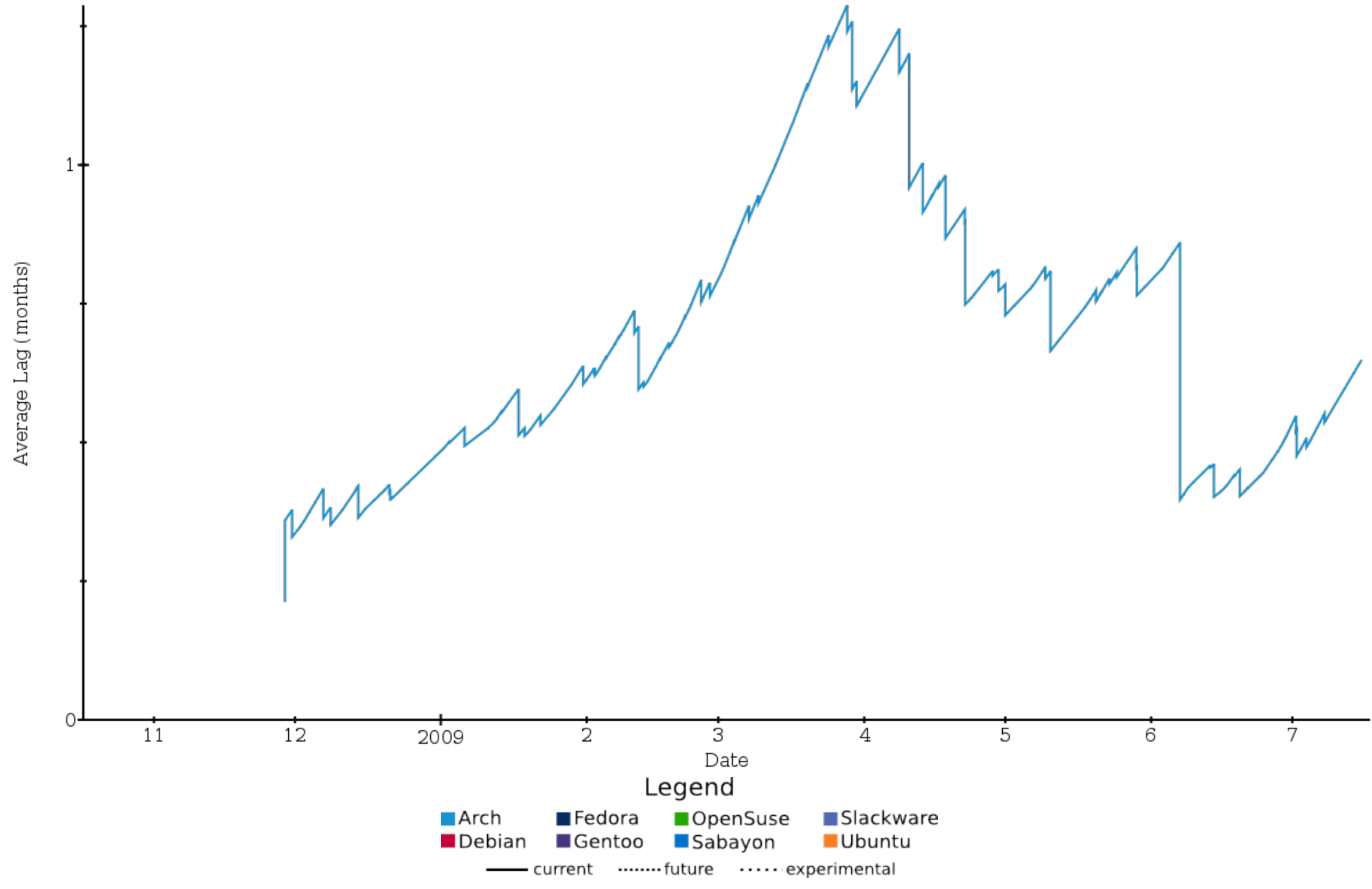
OpenSourceWatershed

"watching software flow downstream"

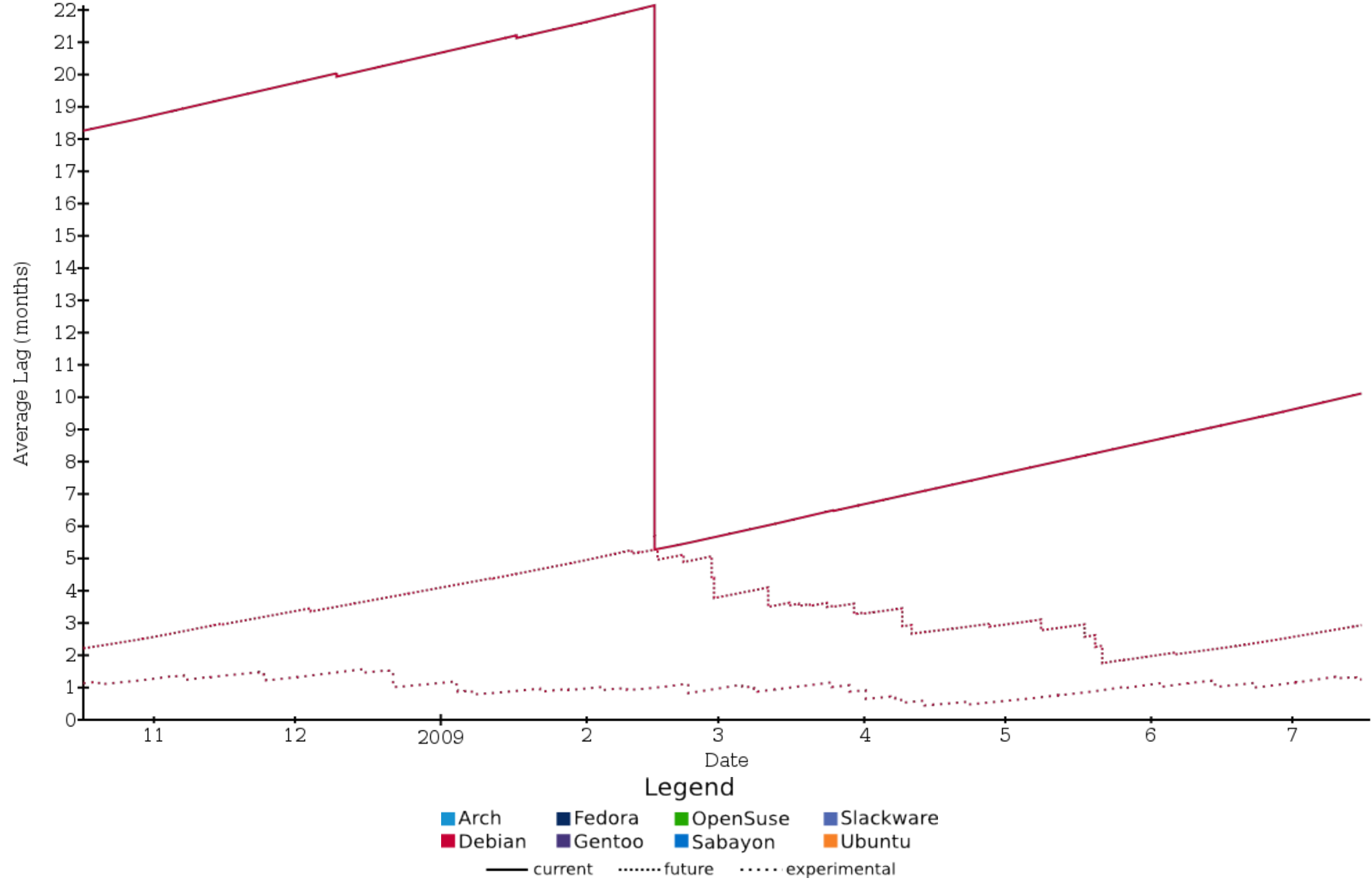
Appendix

- Arch
- Debian
- Fedora
- Gentoo
- OpenSUSE
- Sabayon
- Slackware
- Ubuntu

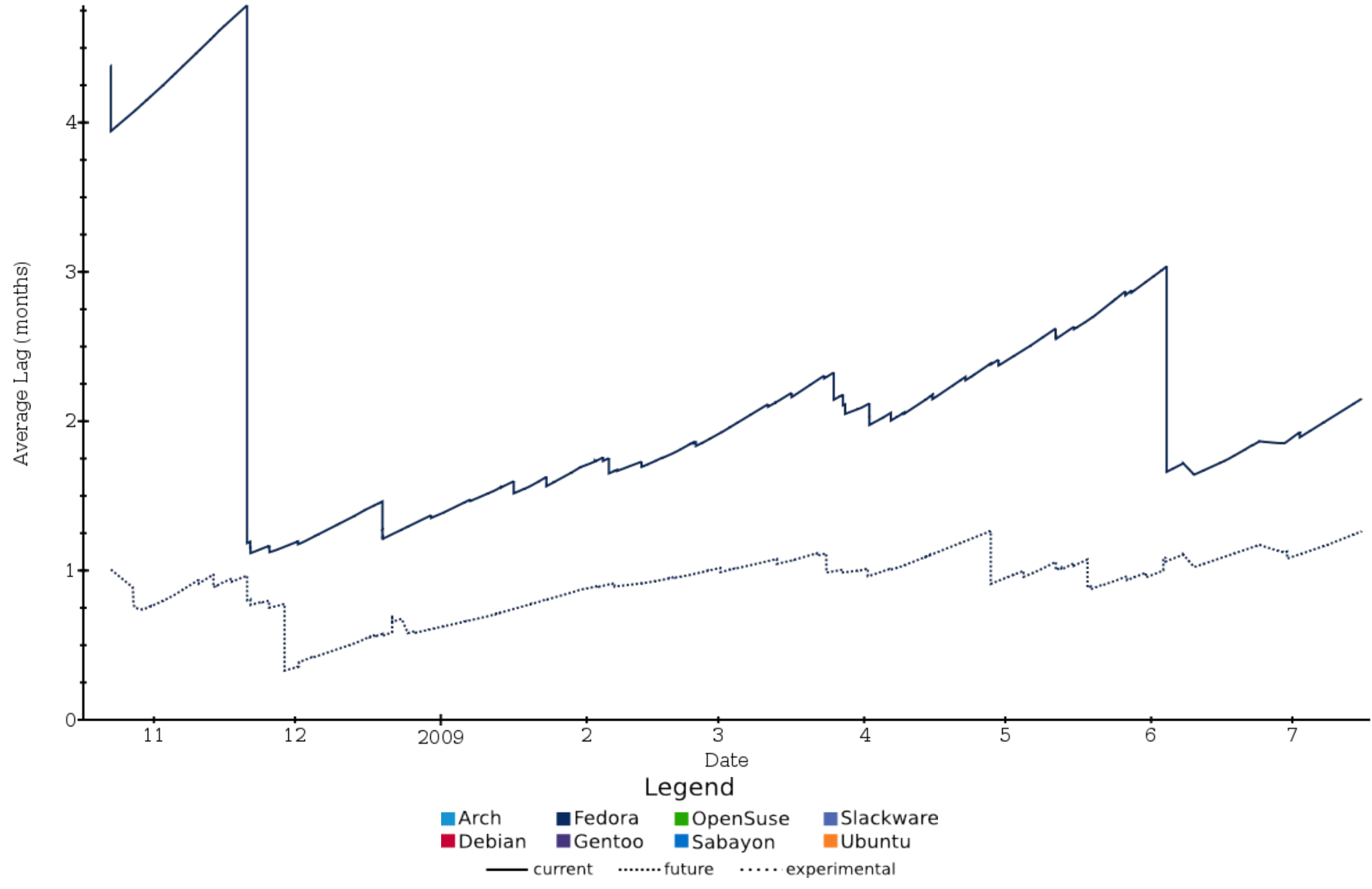
Arch


Average Lag (months)
1
0
11
12
2009
2
3
4
5
6
7
Date
Legend
Arch
Fedora
OpenSuse
Slackware
Debian
Gentoo
Sabayon
Ubuntu
current
future
experimental

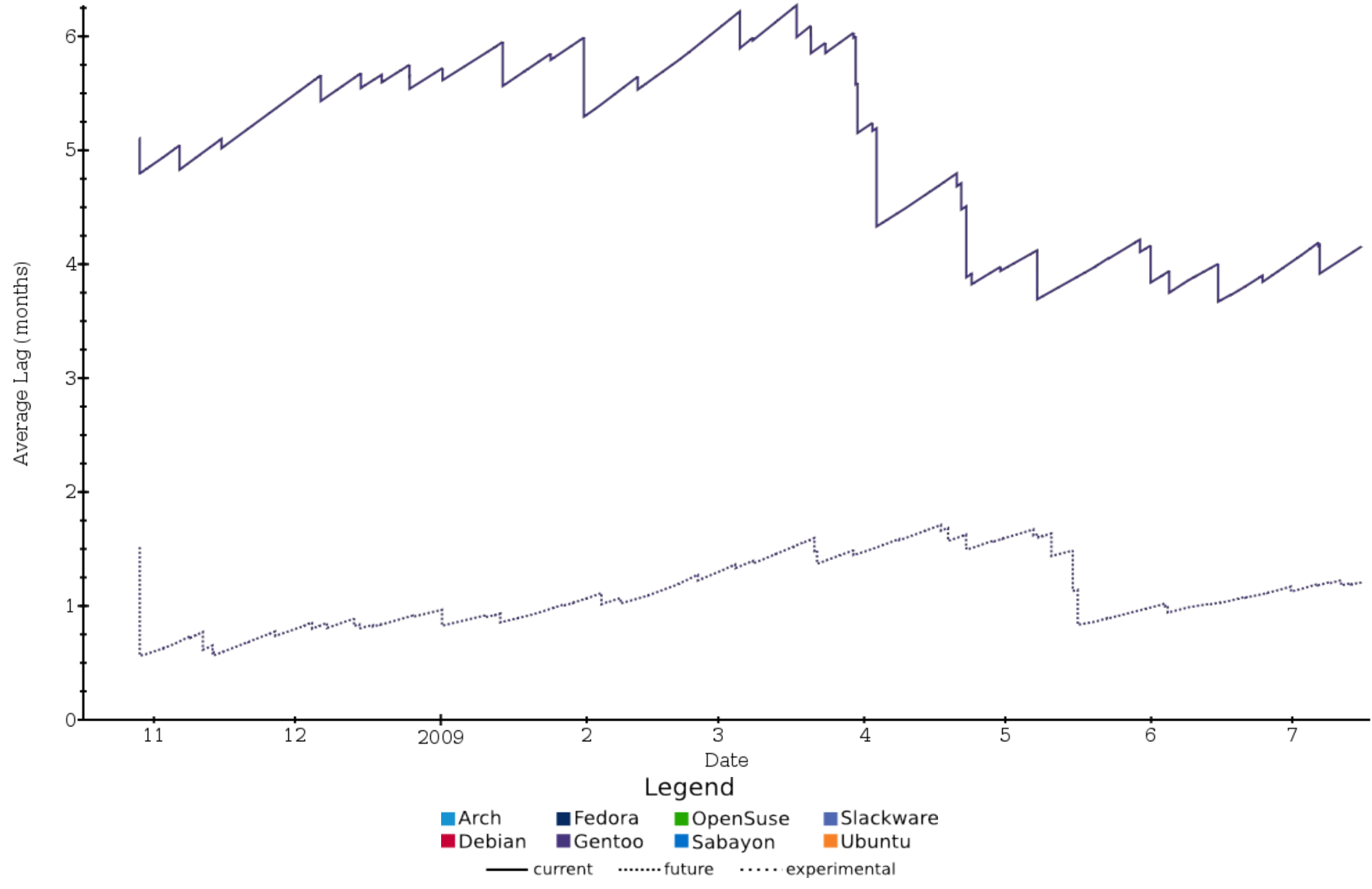
Debian


Average Lag (months)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
11
12
2009
2
3
4
5
6
7
Date
Legend
Arch
Fedora
OpenSuse
Slackware
Debian
Gentoo
Sabayon
Ubuntu
current
future
experimental

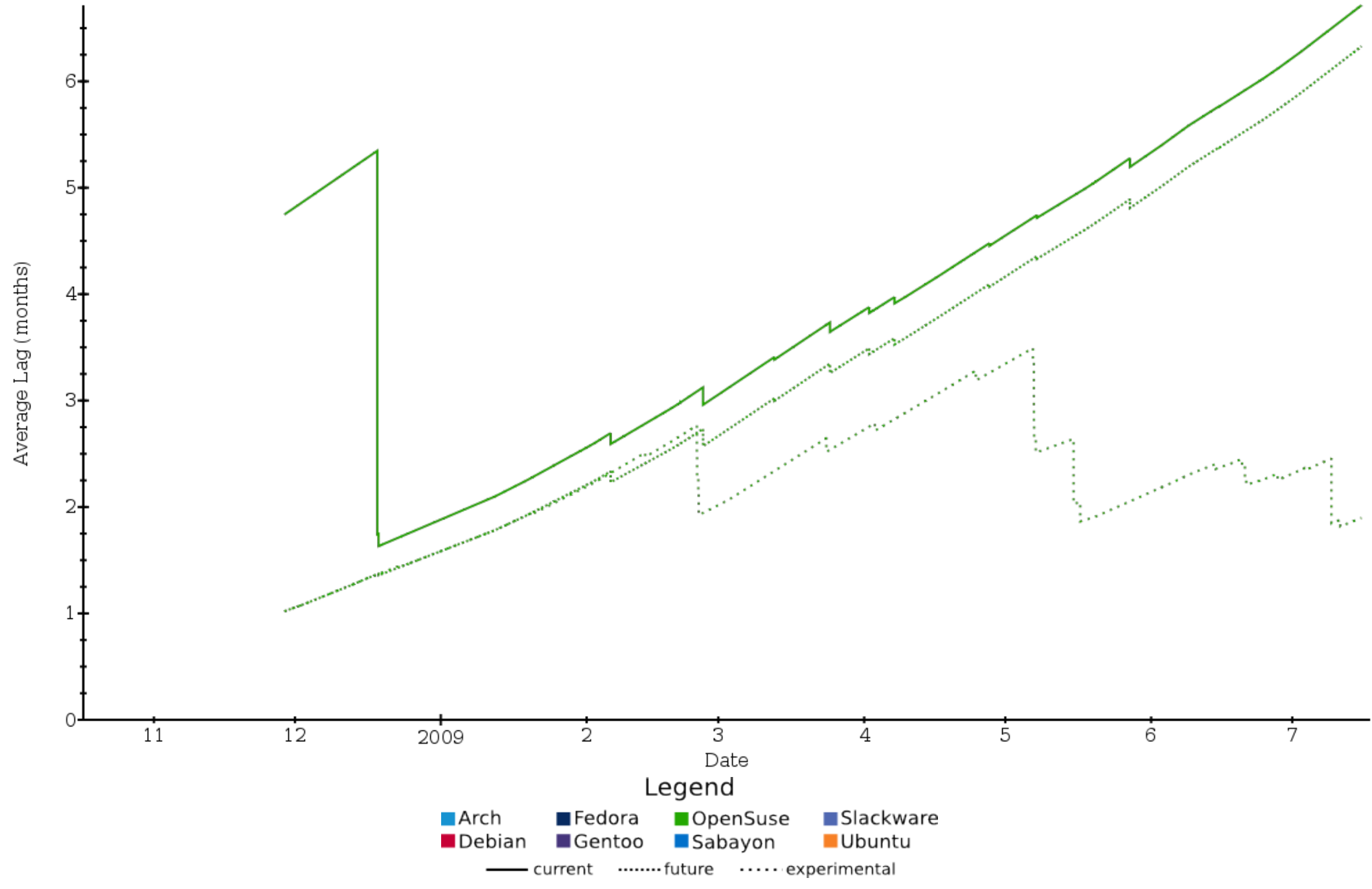
Fedora


Average Lag (months)
0
1
2
3
4
11
12
2009
2
3
4
5
6
7
Date
Legend
Arch
Fedora
OpenSuse
Slackware
Debian
Gentoo
Sabayon
Ubuntu
current
future
experimental

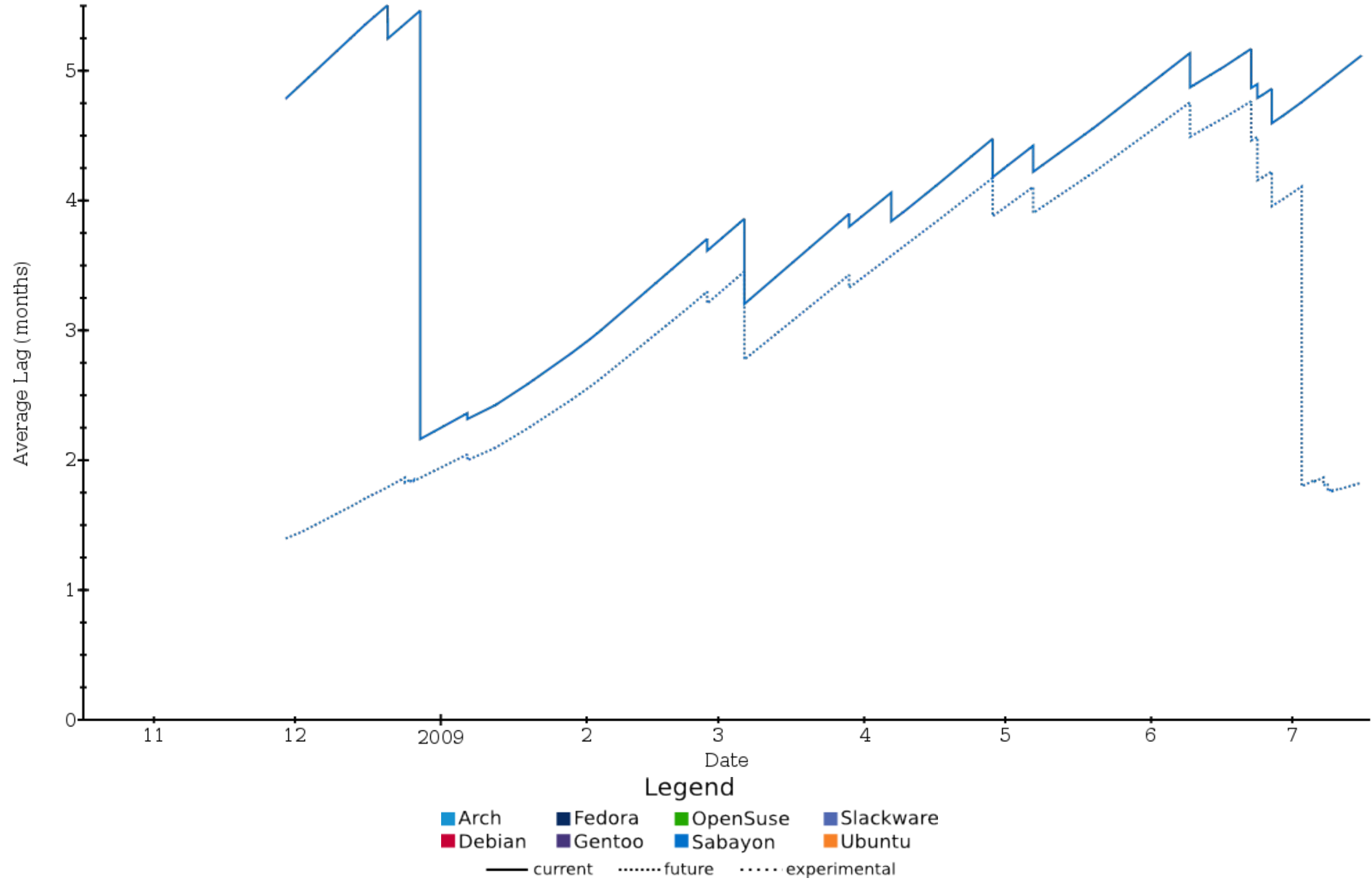
Gentoo


Average Lag (months)
0
1
2
3
4
5
6
11
12
2009
2
3
4
5
6
7
Date
Legend
Arch
Fedora
OpenSuse
Slackware
Debian
Gentoo
Sabayon
Ubuntu
current
future
experimental

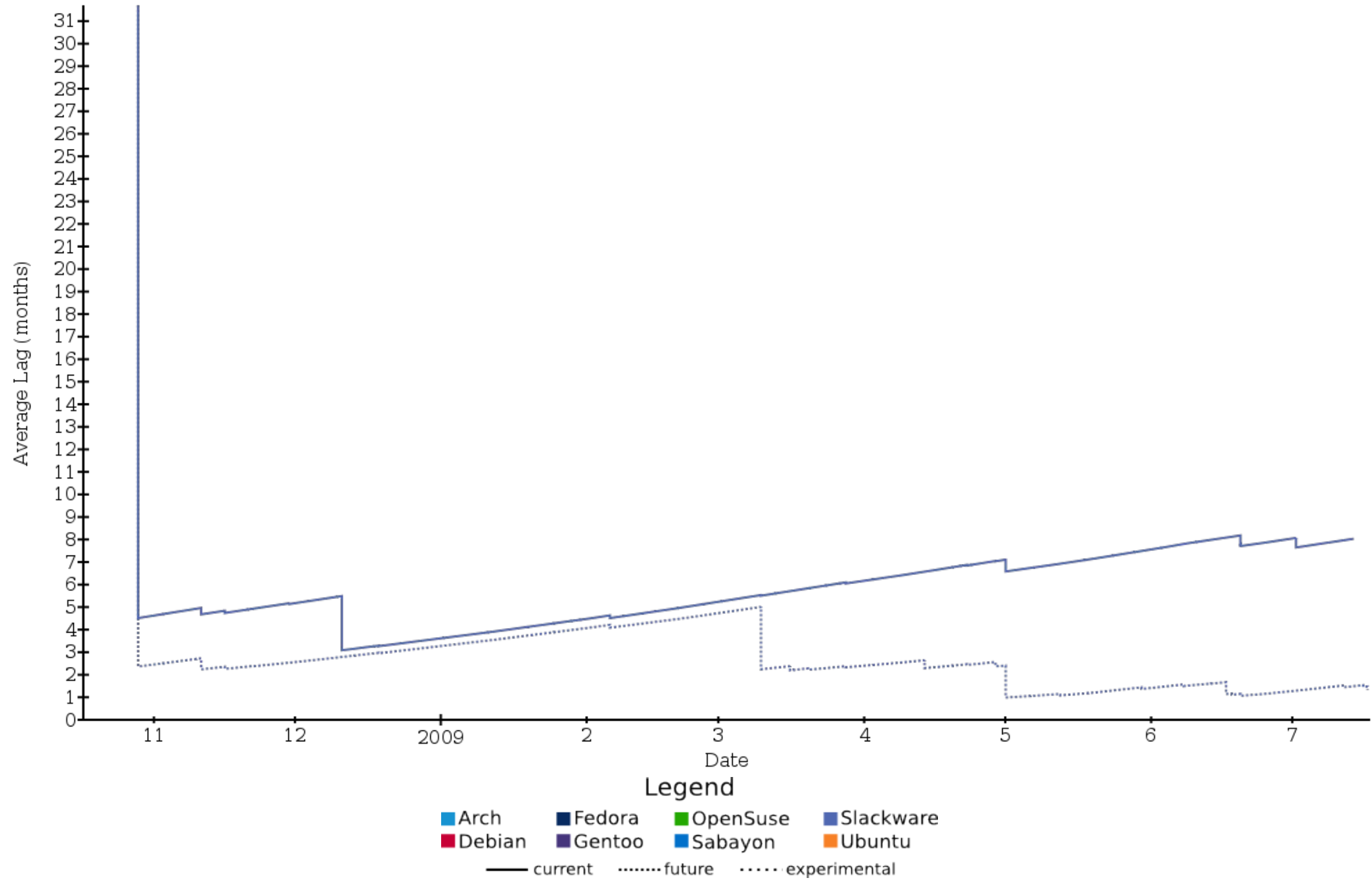
OpenSuse


Average Lag (months)
0
1
2
3
4
5
6
11
12
2009
2
3
4
5
6
7
Date
Legend
Arch
Fedora
OpenSuse
Slackware
Debian
Gentoo
Sabayon
Ubuntu
current
future
experimental

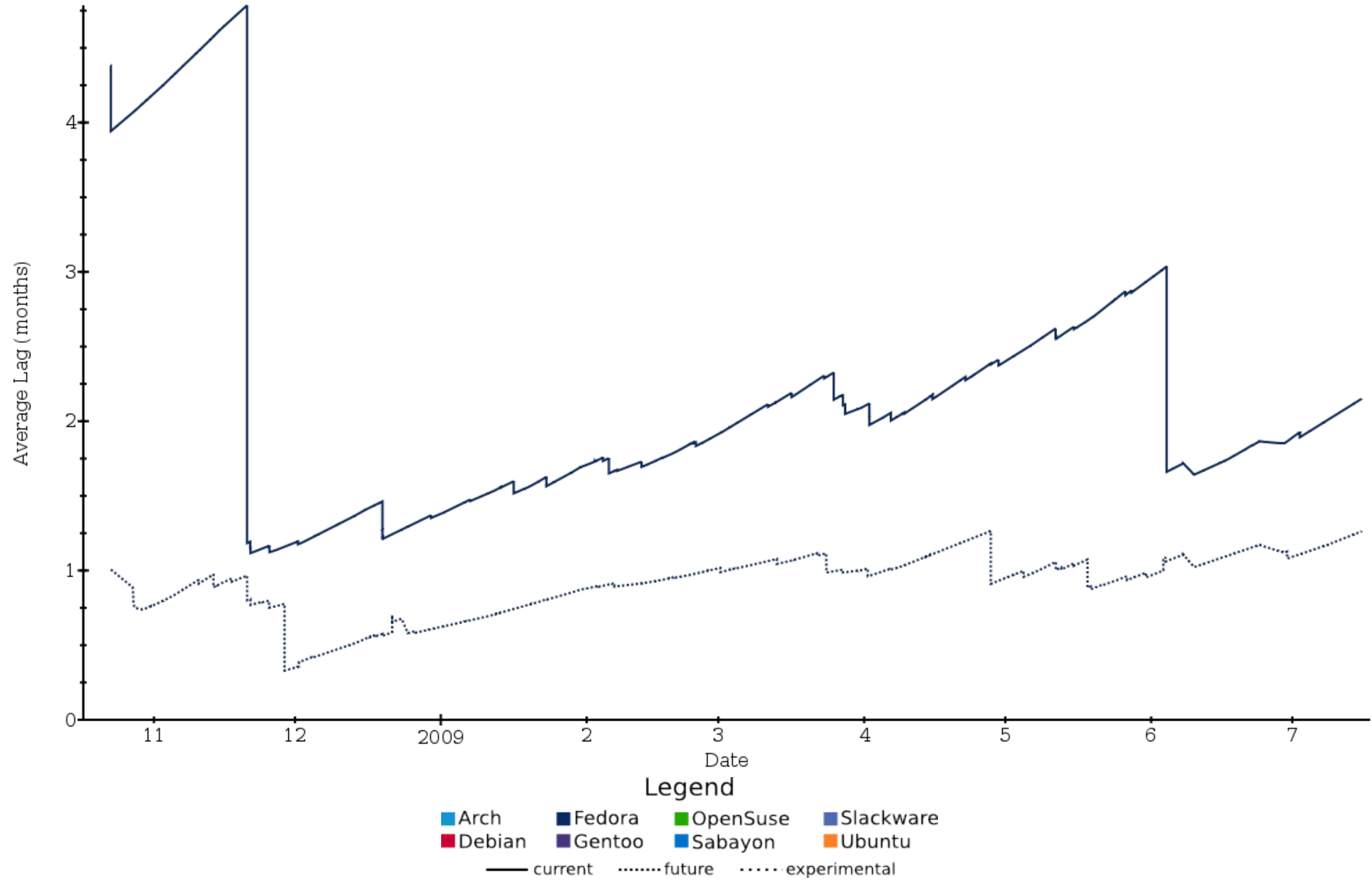
Sabayon


Average Lag (months)
0
1
2
3
4
5
11
12
2009
2
3
4
5
6
7
Date
Legend
Arch
Fedora
OpenSuse
Slackware
Debian
Gentoo
Sabayon
Ubuntu
current
future
experimental

Slackware


Average Lag (months)
31
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
11
12
2009
2
3
4
5
6
7
Date
Legend
Arch
Fedora
OpenSuse
Slackware
Debian
Gentoo
Sabayon
Ubuntu
current
future
experimental

Ubuntu


Average Lag (months)
0
1
2
3
4
11
12
2009
2
3
4
5
6
7
Date
Legend
Arch
Fedora
OpenSuse
Slackware
Debian
Gentoo
Sabayon
Ubuntu
current
future
experimental