Publication patterns in HEP computing

M. G. Pia¹, T. Basaglia², Z. W. Bell³, P. V. Dressendorfer⁴

¹INFN Genova, Genova, Italy
²CERN, Geneva, Switzerland
³ORNL, Oak Ridge, TN, USA
⁴IEEE, Piscataway, NJ, USA

CHEP 2012, NYC
Analysis topics

- General tools
  - Geant4
  - ROOT

- HEP experiments
  - LEP
    - ALEPH, DELPHI, L3, OPAL
  - BaBar
  - LHC
    - ALICE, ATLAS, CMS, LHCb, TOTEM

- Grid computing
  - LCG

What they publish
- How much
- Where
- Citations
- Technology vs physics
- Software vs hardware
- Software/DAQ-trigger
Data sources

Thomson-Reuters: ISI Web of Knowledge
- CERN subscription: since 1970, conference database not included
- Search by keywords, collaboration name

Journal web sites
- IEEE TNS
- NIM, Comp. Phys. Comm. (Elsevier)
- JINST (IOP/SISSA)
  ➤ Full-text searches

CERN databases
- CERN Document System
- Greybook

Years: 1982-2011 (LEP), 1992-2011 (BaBar, LHC)
- Reproducible sample
Data sample

- **Contamination**
  - Non-pertinent entries in the data sample

- **Omission**
  - Pertinent papers are not included in the data sample

- **Cross-checks**
  - WoS/CDS, WoS/publishers’ web sites

- **WoS inconsistencies and errors**
  - Total number of citations includes Conference database
  - Proceedings papers: false classifications and omissions
  - Manually corrected whenever possible

- **Automated analysis** (whenever possible)

- **Manual evaluation**: abstracts and full-text papers
  - Some degree of subjectivity
Geant 4

S. Agostinelli et al.
Geant4: a simulation toolkit

2934 citations
(14 May 2012)

Most cited CERN publication in WoS
(excluding Rev. Part. Properties)

J. Allison et al.
Geant4 Developments and Applications

574 citations
(14 May 2012)

Many papers cite the NIM paper, but they omit citing the TNS one, even though both are indicated in http://cern.ch/geant4.

Many papers that use Geant4 do not cite either reference.

Maria Grazia Pia, INFN Genova

Citation analysis: until 2011 (reproducibility)
Geant 4
Born from LHC experimental requirements
Multidisciplinary sources of citations

Geant4 NIM: Citing Journals

Geant4 NIM: Citing Collaborations

75% citations (plot)
30% Physics

16% citations (plot)
19% citations from collaborations

LHC
HEP
Other

Maria Grazia Pia, INFN Genova
R. Brun and F. Rademakers

**ROOT - An object oriented data analysis framework**

*NIM A*, vol. 389, no. 1-2, pp. 81-86, 1997

Citation analysis: until 2011 (reproducibility)

540 citations
(14 May 2012)

347 citations
excluding proceedings

I. Antcheva et al.

**ROOT - A C++ framework for petabyte data storage, statistical analysis and visualization**

*Comp. Phys Comm.*, vol. 180, no. 12, pp. 2499-2512, 2009

Citation analysis: until 2011 (reproducibility)

27 citations
(14 May 2012)

20 citations
excluding proceedings
ROOT

ROOT Proc.: Citing Journals

- Astropart. Phys.
- Lect. Notes Comp.
- NIM B
- JHEP
- EPJC
- JINST
- Phys. Rev. D
- Phys. Rev. C
- TNS
- NIM A

75% citations

Field of citing journals

<table>
<thead>
<tr>
<th>Field</th>
<th>Geant4 %</th>
<th>ROOT %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>30.3</td>
<td>49.6</td>
</tr>
<tr>
<td>Physics</td>
<td>29.9</td>
<td>18.2</td>
</tr>
<tr>
<td>BioMedical</td>
<td>13.9</td>
<td>6.0</td>
</tr>
</tbody>
</table>

Maria Grazia Pia, INFN Genova
HEP experiments

LEP
- ALEPH
- DELPHI
- L3
- OPAL

BaBar

LHC
- ALICE
- ATLAS
- CMS
- LHCb
- TOTEM

Start of run
- LEP: 1989
- BaBar: 1999
- LHC: 2008

Collaboration members
Time distribution

Run start
- LEP: 1989
- BaBar: 1999
- LHC: 2008

Publication year

Number of publications vs operation year

Publication year

Rescaled w.r.t. year of start run

Maria Grazia Pia, INFN Genova
Time distribution

Run start

LEP: 1989
BaBar: 1999
LHC: 2008

Publications/member vs. year

Year


Publications/member vs. operation year

Year

-20 -10 0 10 20

Same as previous slide, rescaled by the number of experiment members
Publications

Share of hardware, software and DAQ-trigger publications

Maria Grazia Pia, INFN Genova
Physics publications

LEP experiments completed their life-cycle
LHC experiments: at an early stage of their physics production

Maria Grazia Pia, INFN Genova
Roughly constant trends, once the number of publications is normalized to the number of collaborators.
Software vs. hardware

Hardware/Software publications

Experiment

Ratio

daq-trigger/software publications

Experiment

Ratio

Hardware publications: approximately 4 times more than software
DAQ-trigger publications: approximately 1.3 times more than software
Journals: LEP and LHC

Dominated by physics publications

Still dominated by technological publications

Maria Grazia Pia, INFN Genova
IEEE TNS is the most popular journal for HEP technological publications in recent years
The most cited papers are often the general reference papers about the detector published by each experiment.

Citations of the most cited paper:

- **ALEPH**: 340
- **DELPHI**: 309
- **L3**: 509
- **OPAL**: 473
- **BaBar**: 859
- **ALICE**: 116
- **CMS**: 129
- **LHCb**: 101
- **TOTEM**: 35

**ATLAS**: ATLAS pixel detector electronics and sensors: 185
Physics papers cite more references than technological papers.

More references → more citations.

Bibliographical entries in software papers are often web sites.
The number of pages of a paper depends on the format of the journal:
- $1 \text{ page}_{TNS} \approx 2.5 \text{ pages}_{JINST}$

Different journal formats in the same category

Evolutions of the format of some journals (e.g. NIM)
Sources of citations to physics papers

Samples in plots account for >90% of citations

Citations to HEP physics papers mostly come from journals specialized in HEP and a few related fields (astroparticle and nuclear physics)

Maria Grazia Pia, INFN Genova
Sources of citations to technological papers

Citations from HEP physics and technology journals

Maria Grazia Pia, INFN Genova
More refined analysis of technological papers published since start of LHC run

TNS 2008-2011

NIM 2008-2011
Citations 2008-2011

Self-citations
Outside citations

Number of self-citations
Number of outside citations

Hardware  Software  DAQ-trigger
Hardware  Software  DAQ-trigger

TNS

NIM A

Maria Grazia Pia, INFN Genova
### LCG – LHC Computing Grid

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Title</th>
<th>Journal</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sakamoto, H</td>
<td>Data grid deployment for high energy physics in Japan</td>
<td>CPC</td>
<td>2007</td>
</tr>
<tr>
<td>Belov, S et al.</td>
<td>LCG MCDB - a knowledgebase of Monte-Carlo simulated events</td>
<td>CPC</td>
<td>2008</td>
</tr>
<tr>
<td>Yin, Fet al.</td>
<td>Grid resource management policies for load-balancing and energy-saving by vacation queuing theory</td>
<td>CPC</td>
<td>2009</td>
</tr>
<tr>
<td>Chytracek, R et al.</td>
<td>POOL development status and production experience</td>
<td>TNS</td>
<td>2005</td>
</tr>
<tr>
<td>Hatlo, M et al.</td>
<td>Developments of mathematical software libraries for the LHC experiments</td>
<td>TNS</td>
<td>2005</td>
</tr>
<tr>
<td>Pfeiffer, A et al.</td>
<td>The LCG PI project: Using interfaces for physics data analysis</td>
<td>TNS</td>
<td>2005</td>
</tr>
<tr>
<td>Munro, C et al.</td>
<td>Measurement of the LCG2 and gLite File Catalogue's performance</td>
<td>TNS</td>
<td>2006</td>
</tr>
<tr>
<td>Kim, BK et al.</td>
<td>A Composition of Monitoring Services for the LHC Computing Grid</td>
<td>J. Grid Comp.</td>
<td>2009</td>
</tr>
</tbody>
</table>
Small sample of publications

Hard to perform any statistical analysis
Conclusions

- Software is largely **underrepresented** in HEP scholarly literature w.r.t. hardware
- **Publication patterns** appear similar in the LEP and LHC era
- **Citation patterns** are different for publications by HEP experiments and about general software tools

Publish!

...and don’t forget to cite