Filesystem Comparison
NFS, GFS2, OCFS2

Giuseppe “Gippa” Paternò
Visiting Researcher
Trinity College Dublin

Who am I

- Visiting Researcher at Trinity College Dublin (Ireland)
- Solution Architect and EMEA Security Expert in Red Hat
- Previously Security Solution Architect in Sun and also in IBM
- Red Hat Certified Security Specialist (RHCSS), Red Hat Certified Architect (RHCA) and Cisco Certified Network Professinal (CCNP)
- Part of the world-wide security community (expecially SEMEA)
- Published books and whitepapers
- Forensic analisys for local govs
- More on:
  - http://www.scss.tcd.ie/Giuseppe.Paterno/
  - http://www.gpaterno.com/
  - http://www.linkedin.com/in/gpaterno

Disclaimer

I do not speak on behalf of my employer, nor I am authorized to represent it publicly.

All and any opinion and results expressed in this presentation are solely mine and do not represent my employer point-of-view.

The performance tests and their results were taken on a real project as a TCD researcher out of business hours.

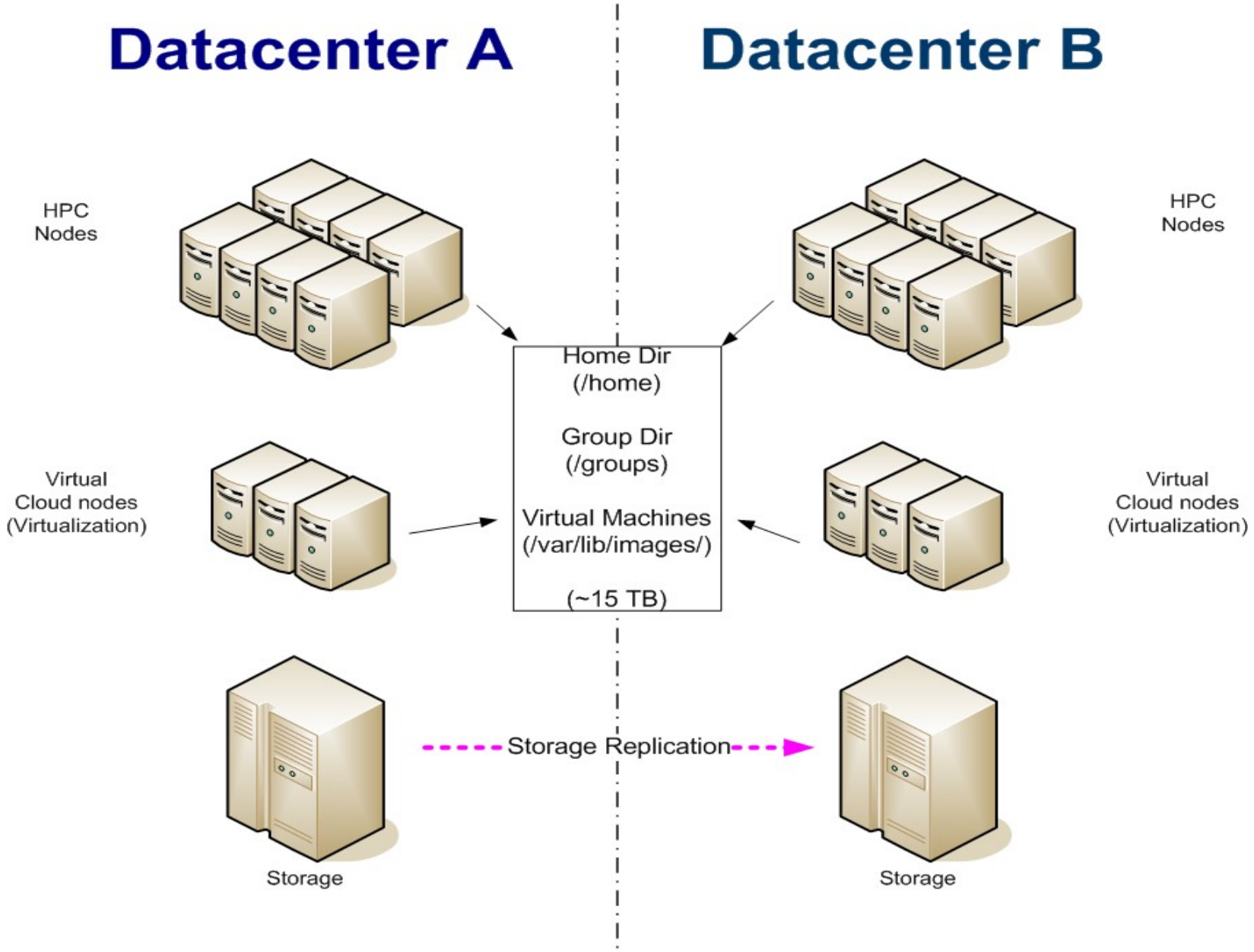
Project overview

- National importance research project
- High-Performance Computing (HPC) cluster with utility nodes
  - Split in two datacenters 25km distance in “active-active” mode
  - 8 nodes to a “private virtual cloud”
  - 16 nodes to number crunching
  - Storage (Hitachi) data replication

Project overview (2)

- High bandwidth:
  - 512 gb/s switch fabric
  - 60 gb/s cluster inter-site link
  - 20 gb/s inter-site admin link
  - 16 gb/s + 16 gb/s SAN inter-site links
  - Each node has 2x10gb/s ethernet adapter in link aggregation

Architecture Overview



Typical researcher usage

Connect to a “master node” and submit a job that:

- Downloads around 4gb data from mainframe (IBM DB2)
- User upload custom data via Samba share
- Creates his algorithm using mathlabs, SPSS or other statistics programs (even FORTRAN)
- Number crunching
- Re-iterate if needed
- Creates an automatic document
- User download results via Samba (SMB)

The filesystem is structured in homes and group directories

Issue:

Having a common filesystem across physical nodes and virtual nodes to share users' data with the maximum performance

Selection phase 1

- Objective: compare a network file system to a cluster file system (NFS vs GFS2)
- Generic load simulation:
  - Command “dd” and “rm” on 1 and 2 gb datafile size
  - Step-by-Step concurrent nodes: 2, 6, 10, 14

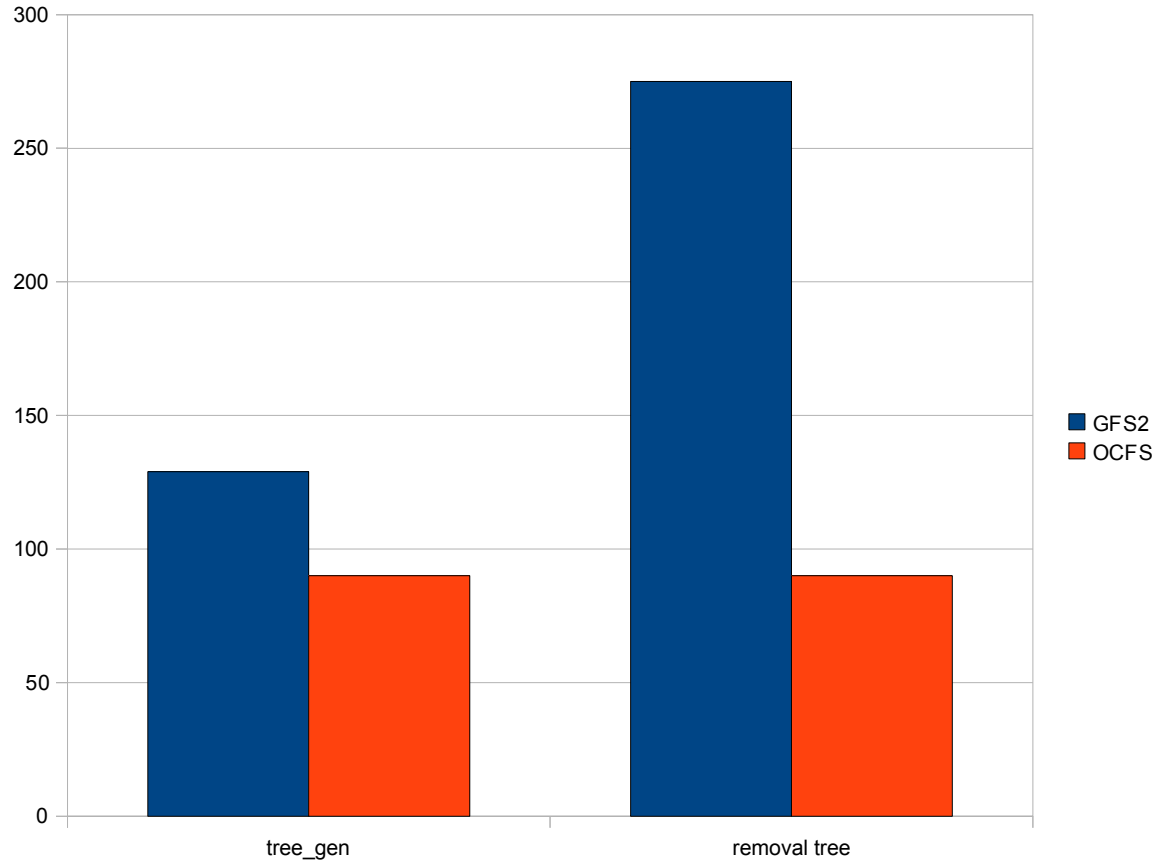
NFS vs GFS2 (generic load)

Nodes	I/O rate NFS (MB/s)	NFS avg transfer rate (MB/s)	I/O rate GFS (MB/s)	GFS avg transfer rate (MB/s)
2	21	2	43	2
6	11	6	46	4
10	8	6	45	5
14	0.5	0.1	41	8

Selection Phase 2

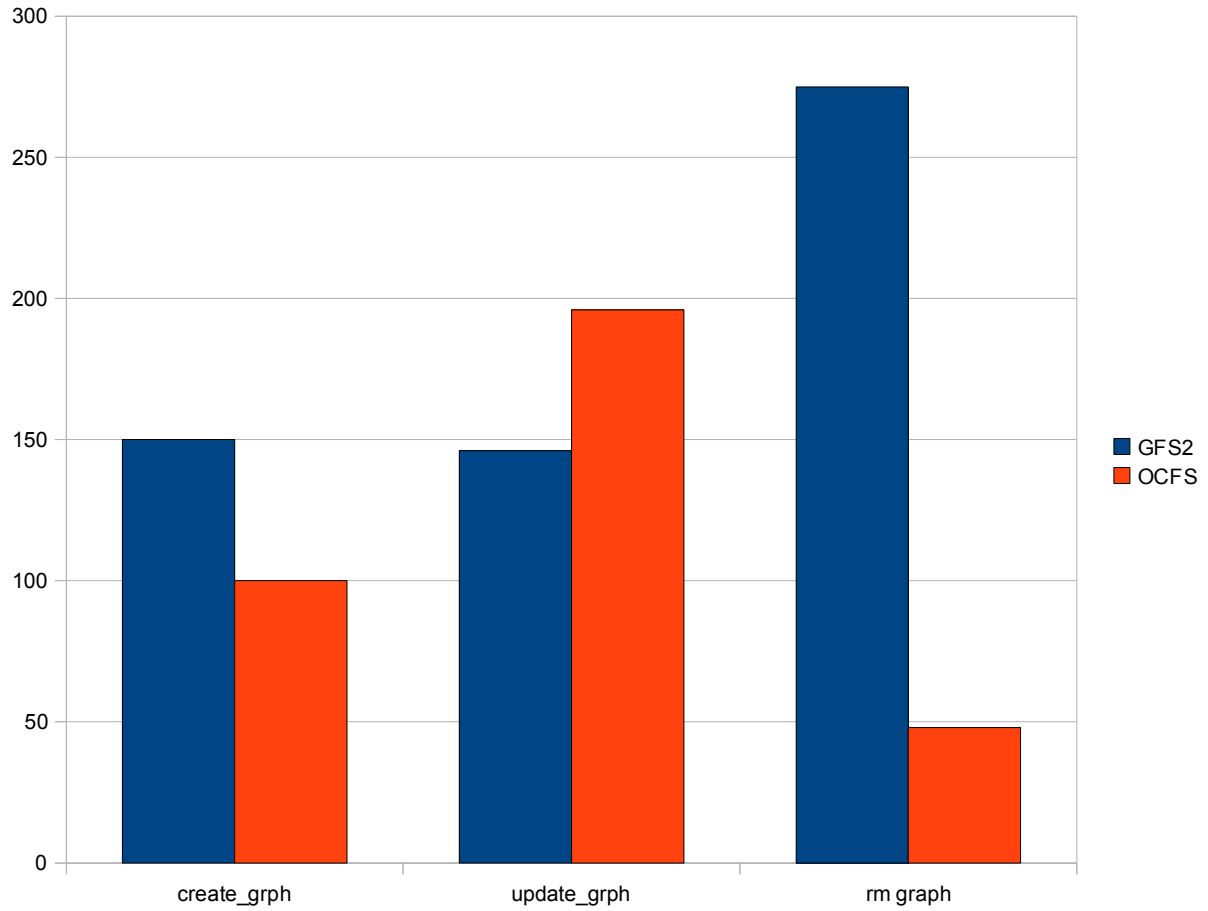
- Objective: select the best cluster filesystem for the specific load (GFS2 vs OCFS2)
- Created a custom set of scripts to simulate researchers' load:
  - creation of about 10.000 directory trees, 8 levels with 3 subdiretory each (tree_gen.sh)
  - creation of one file for each leaf directory of 1600 bytes (crea_grf.sh)
  - read/update/write of each of the above file with 20 bytes more (aggiorna_grf.sh)
  - change group ownership in the above subtree (chgrp -R)
  - removal of the subtree (rm -rf)
- **Each operation is done on a different node of the cluster**

Standard tree generation



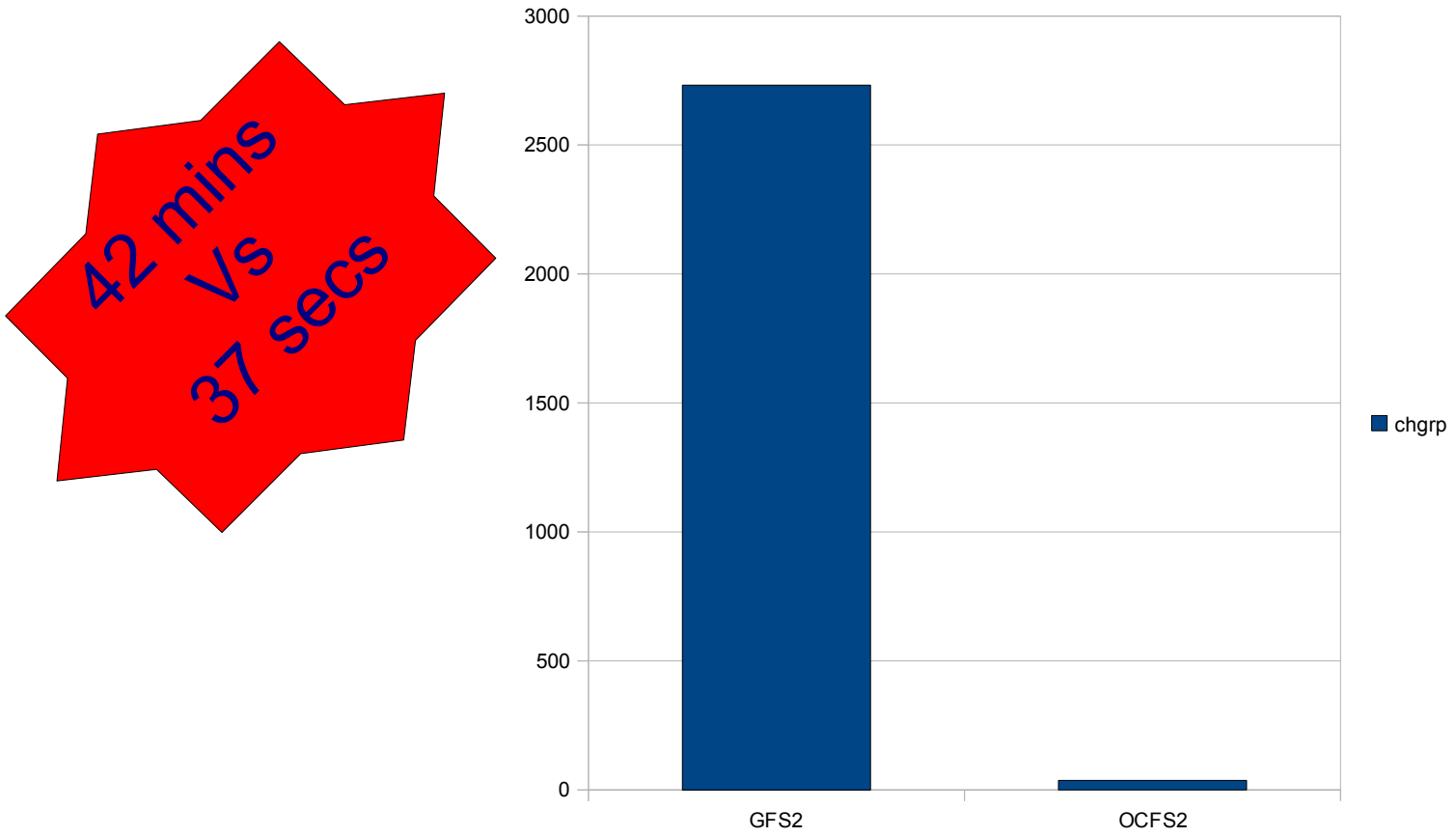
(operation timings in Seconds)

Graph structure generation



(operation timings in Seconds)

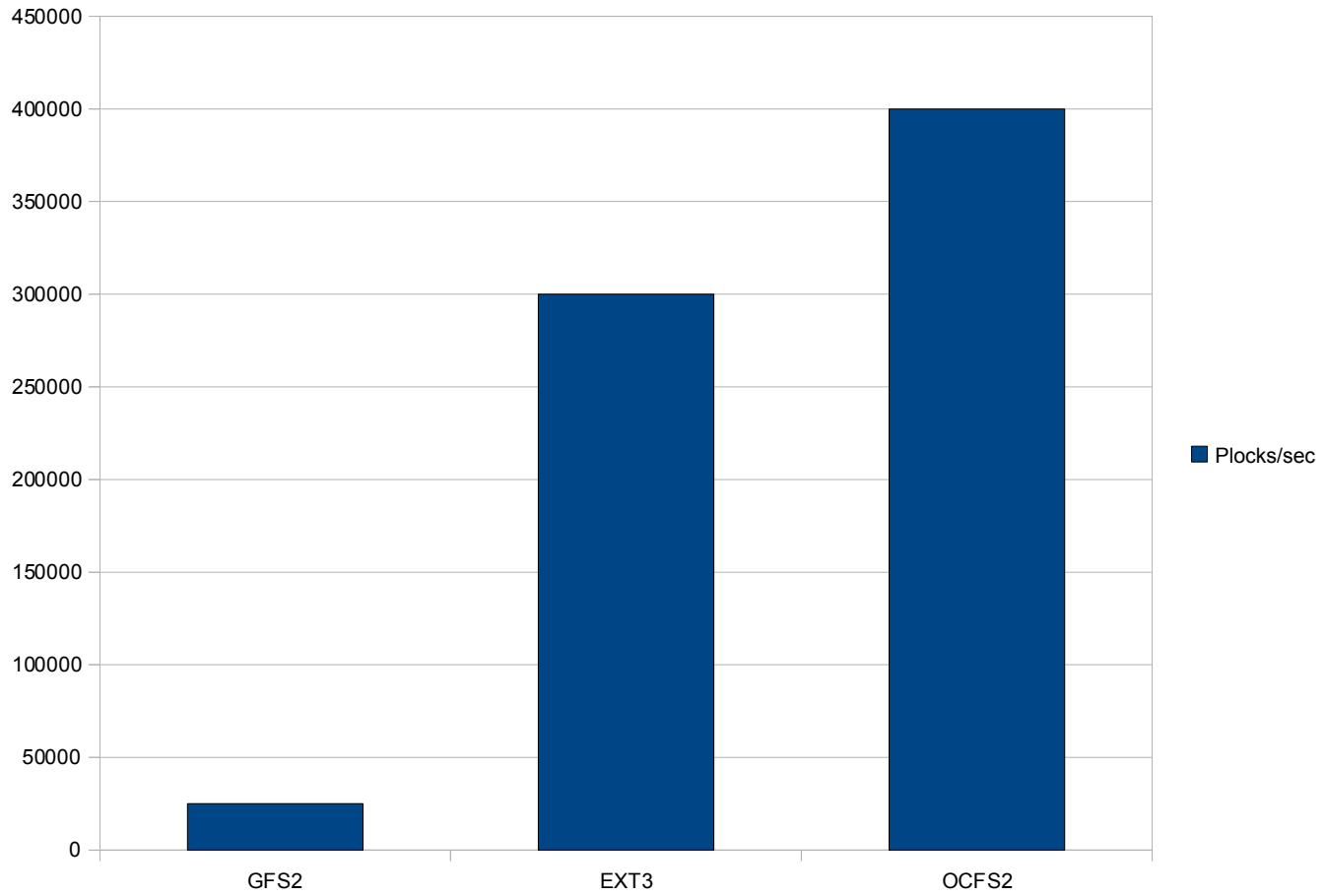
Change group (chgrp)



(operation timings in Seconds)

Operation needed to share data across the working group

POSIX locks



GFS2 vs EXT3 vs OCFS2
(plocks in a second with ping-pong test tool)

Conclusions

- NFS
  - Pro: standard, cross-platform, easy to implement
  - Con: Poor performance, single point of failure (single locking manager, even in HA)
- GFS2
  - Pro: Very responsive on large datafiles, works on physical and virtual, quota and SE-Linux support, faster than EXT3 when I/O operations are on the same node
  - Con: Only supported with Red Hat, Performance issues on accessing small files on several subdirectory on different nodes

Conclusions

- OCFS2
  - Pro: Very fast with large and small datafiles on different node with two types of performance models (mail, datafile). Works on a physical and virtual.
  - Con: Supported only through contract with Oracle or SLES, **no quota support, no on-line resize**

Questions?

Thank you!!

Giuseppe “Gippa” Paternò
Visiting Researcher
Trinity College Dublin

paternog@cs.tcd.ie
http://www.scss.tcd.ie/Giuseppe.Paterno/