Storage System Optimization, Improving CPU Efficiency in I/O bounded

I. Cabrillo Bartolomé

A.Y. Rodríguez Marrero

Instituto de Física de Cantabria (IFCA), Spain

May 25th, 2010

Instituto de Física de Cantabria

CSIC

UNIVERSIDAD DE CANTABRIA

Outline

- Motivation
- IFCA Site
- CPU Performance
- Storage System, Optimization
- Results
- Conclusions

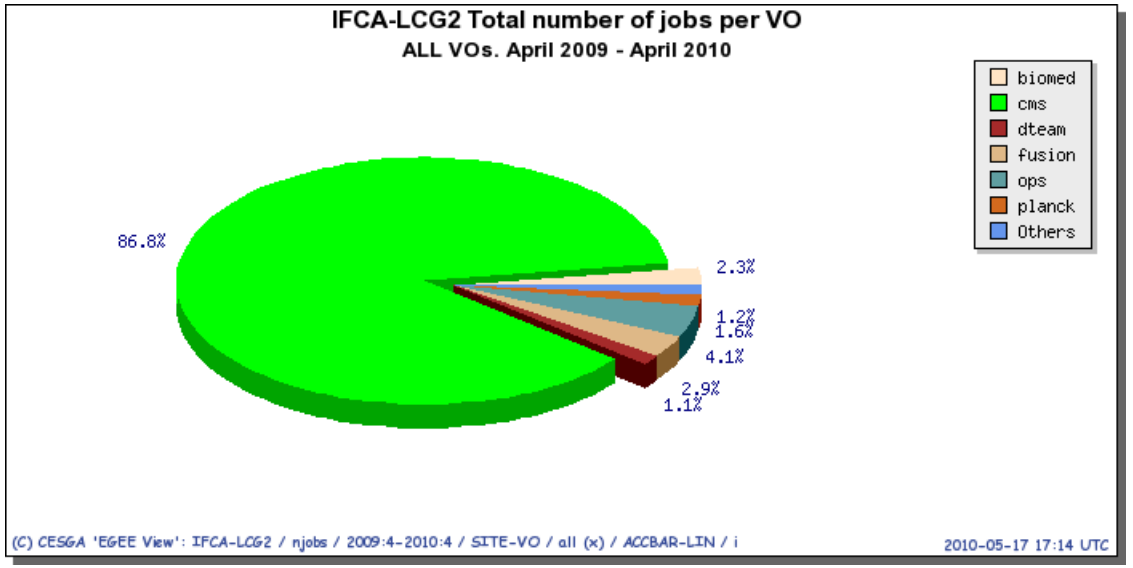
Motivation

- Why the CPU efficiency is an important parameter?
- CPU Efficiency is the ratio between the CPU time and the total execution time
- The shorter the total execution time:
  - the higher the CPU Efficiency
  - the larger the number of jobs that can be executed in a given period
  - the better the resources usage

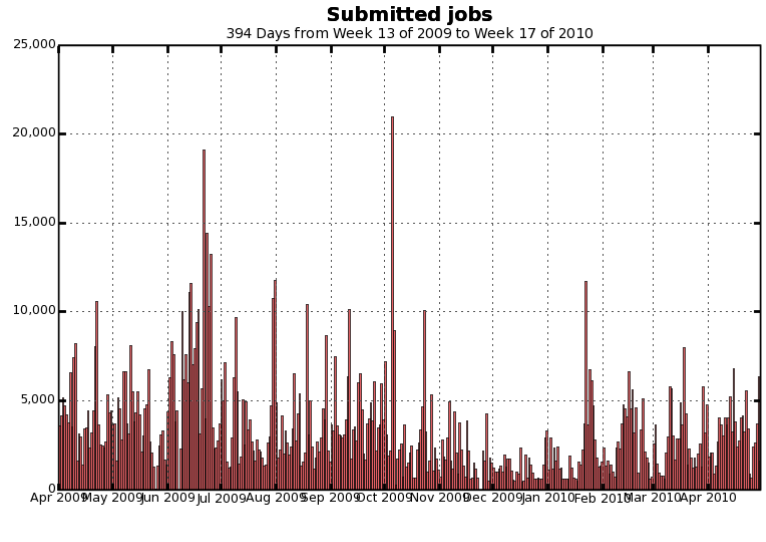
IFCA Site

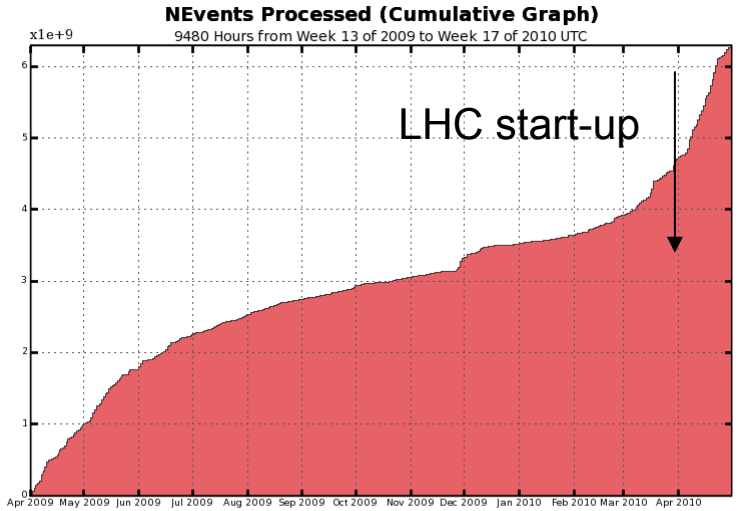
- Join resources from LHC-CMS Tier-2 & GRID-CSIC
- Support ~40 VOs, being CMS the most demanding one
- 210 worker nodes with 2 quad-core CPUs: total 1680 cores
- ~60 local users; 310 TB for storage

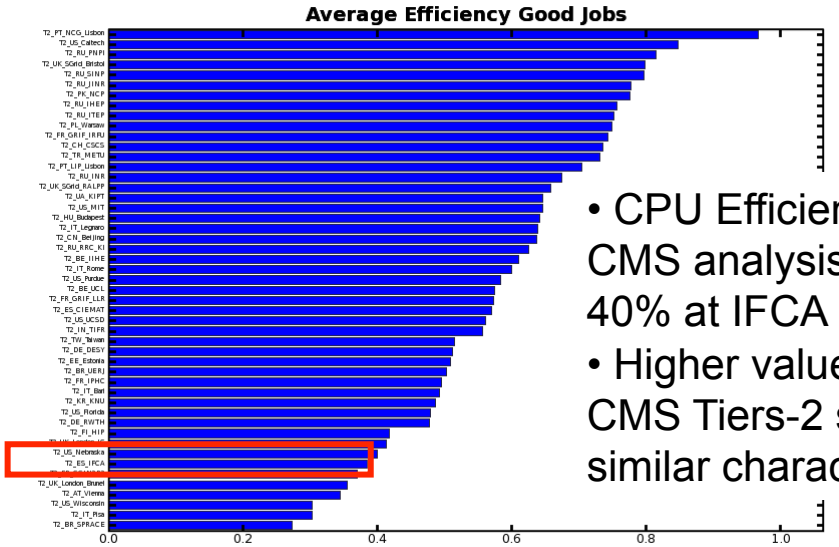
CMS case:
1400 slots, ~20 users
plus 200 via GRID,
220 TB for storage



IFCA Site: CMS activity







- CPU Efficiency for CMS analysis jobs below 40% at IFCA
- Higher values at other CMS Tiers-2 sites with similar characteristics

CPU Perfomance (I)

- The CPU Efficiency depends on many factors:

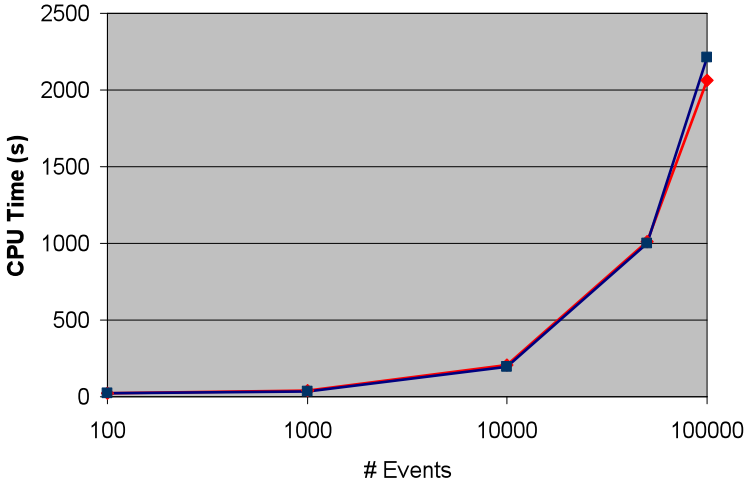
User related:
kind of activity
software
data format

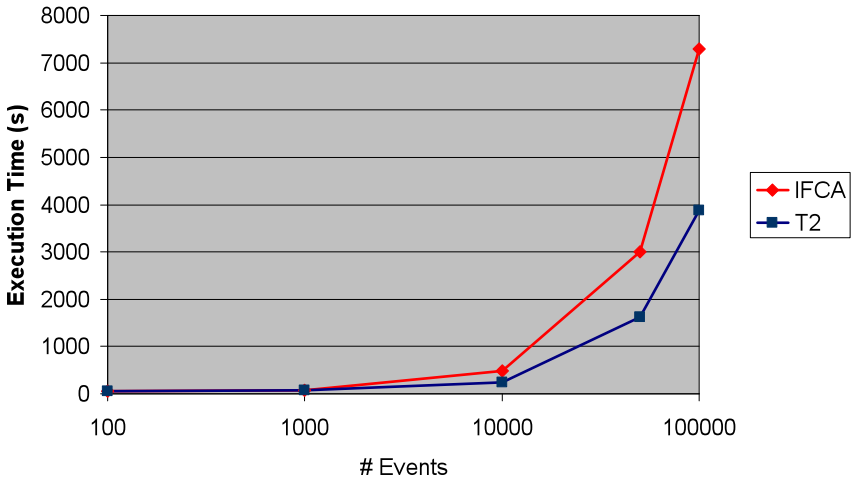
Site related:
Configuration
Status

- We study the CPU efficiency of a typical CMS job that makes an important demand on the storage resources (skimming job)
  - Extracts a small data sample from a larger one with limited calculations
- Execute, in production mode, skimming jobs as a function of the number of analyzed events at IFCA CMS Tier-2 and at a better performing CMS Tier-2 (T2)
- The jobs run over 6 GB of data

CPU Performance (II)

- For each set, we submitted 300 jobs at IFCA using two dedicated worker nodes (16 slots) and 10 jobs at T2





Analysis over # evts:	100	1,000	10,000	50,000	100,000
CPU Eff. (%) at IFCA	45	57	44	34	28
CPU Eff. (%) at T2	45	53	83	62	57

- CPU Times are similar at both sites
- BUT, execution times are considerable larger at IFCA: I/O activity increases with the number of events

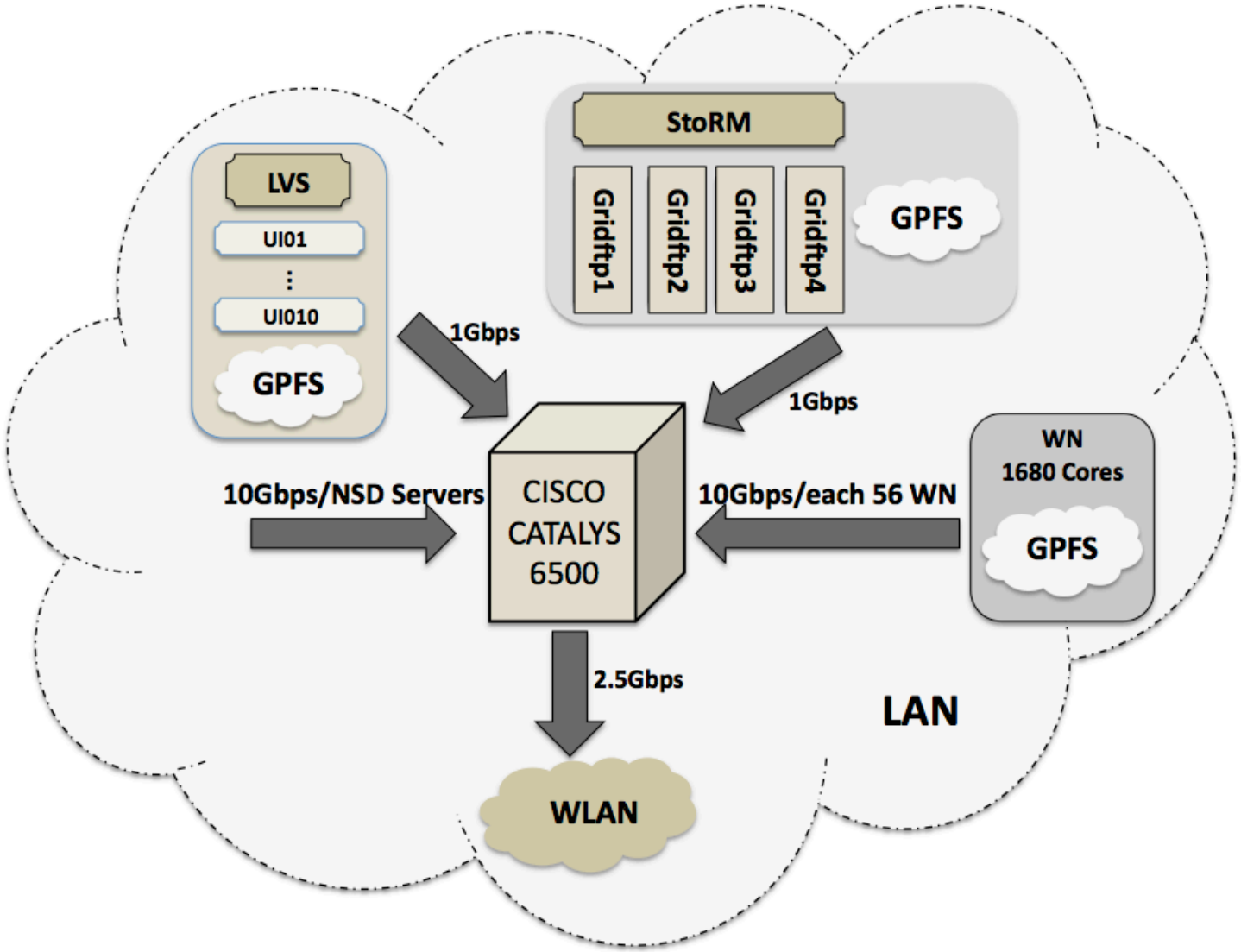
Storage System

- Could the storage system configuration be related to the poor performance?
  - File System
  - LAN Network
  - Storage Hardware
- In production mode, a set of parameters was controlled and tuned in order to optimize the configuration
  - After each modification the CPU efficiency was checked for skims with 100,000 events (6 GB of data)
  - Positive modifications were incorporated in a progressive way

File System

- General Parallel File System (GPFS):
- It allocates its own cache (Pagepool):
  - Default size: 512 MB
  - From 512 MB to 1 GB, the CPU efficiency increases about 2-3%
- Allows the control of the maximum number of threads dedicated to prefetch data and the number of concurrent operations (PrefetchThreads & Worker1Threads)
  - Related parameters changed in the allowed range
  - Not found to have an impact on the CPU efficiency

LAN Network


StoRM
Gridftp1
Gridftp2
Gridftp3
Gridftp4
GPFS
LVS
UI01
UI010
GPFS
1Gbps
1Gbps
10Gbps/NSD Servers
CISCO
CATALYS
6500
10Gbps/each 56 WN
WN
1680 Cores
GPFS
2.5Gbps
LAN
WLAN

LAN Network Optimization (I)

- Measure the maximum bandwidth between nodes of the different components of the LAN (Iperf test): GPFS disk servers and worker nodes

Between GPFS's			Between WN's			Between GPFS's and WN's		
From	To	Mbps	From	To	Mbps	From	To	Mbps
GPFS01	GPFS02	~ 1900	WN01	WN02	~ 1000	GPFS01	WN01	~ 400
GPFS02	GPFS01	~ 2000	WN02	WN01	~ 1000	GPFS02	WN02	~ 350
GPFS03	GPFS01	~ 2100	WN03	WN01	~ 1000	GPFS01	WN02	~ 400
GPFS04	GPFS01	~ 2000	WN04	WN01	~ 1000	GPFS02	WN01	~ 450
GPFS02	GPFS03	~ 2200	WN02	WN03	~ 1000	WN01	GPFS01	~ 1000
GPFS03	GPFS02	~ 1900	WN03	WN02	~ 1000	WN02	GPFS02	~ 1000
GPFS04	GPFS02	~ 2000	WN04	WN02	~ 1000	WN01	GPFS02	~ 1000
GPFS03	GPFS04	~ 2200	WN03	WN04	~ 1000	WN02	GPFS01	~ 1000
GPFS04	GPFS03	~ 2100	WN04	WN03	~ 1000	———	———	———

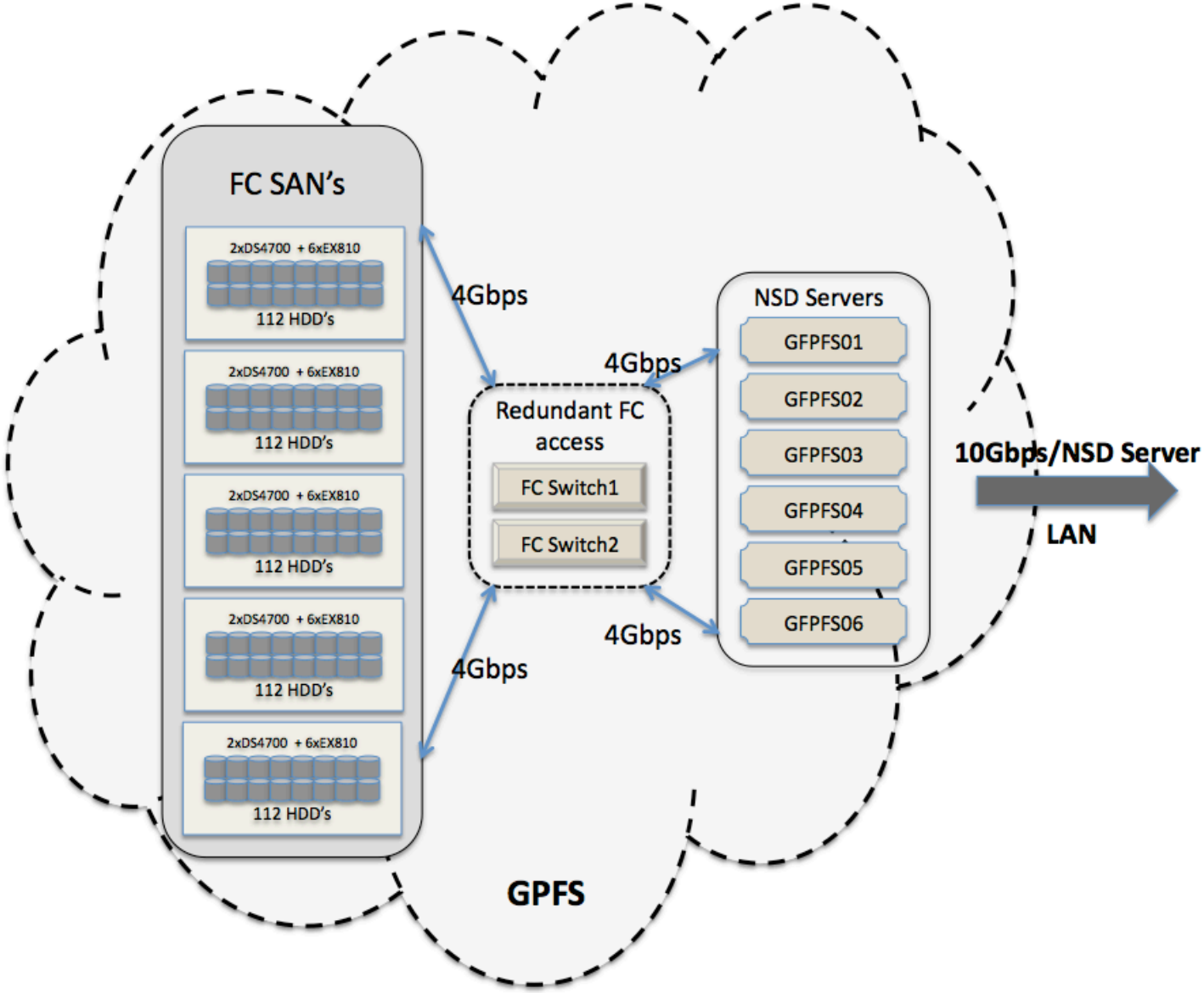
- According to specifications:
  - connection between GPFS servers: 4-5 Gbps
  - connection between GPFS servers and WNs: 1 Gbps

LAN Network Optimization (II)

- Modifications:
  - Disable firewall at GPFS servers:
    - They were in private LAN
    - Connection between GPFS servers is now 4-5 Gbps
  - Enable the TCP feature selective acknowledgements:
    - The data receiver can inform the sender about the segments that have arrived successfully, then the sender only needs to retransmit the segments that have actually been lost
    - Connection between GPFS servers and WNs is now 1 Gbps

- From an initial value of 28%, the CPU efficiency grew up to 35%

Storage Hardware



Storage Hardware Optimization

- Modifications:
  - The hardware controllers have an internal cache size of 2 GB, this is also the size of many of the CMS data files
    - To avoid continuous cache flush: the read cache feature was disabled
  - According to the modification priority variable, part of the resources is allocated to do mainly maintenance processes:
    - In production mode these kind of operations can be done in background, then the priority is lowered to the minimum
    - Then, almost all the resources are employed to do I/O user related operations
- The average CPU efficiency value is duplicated from 35% to 70%

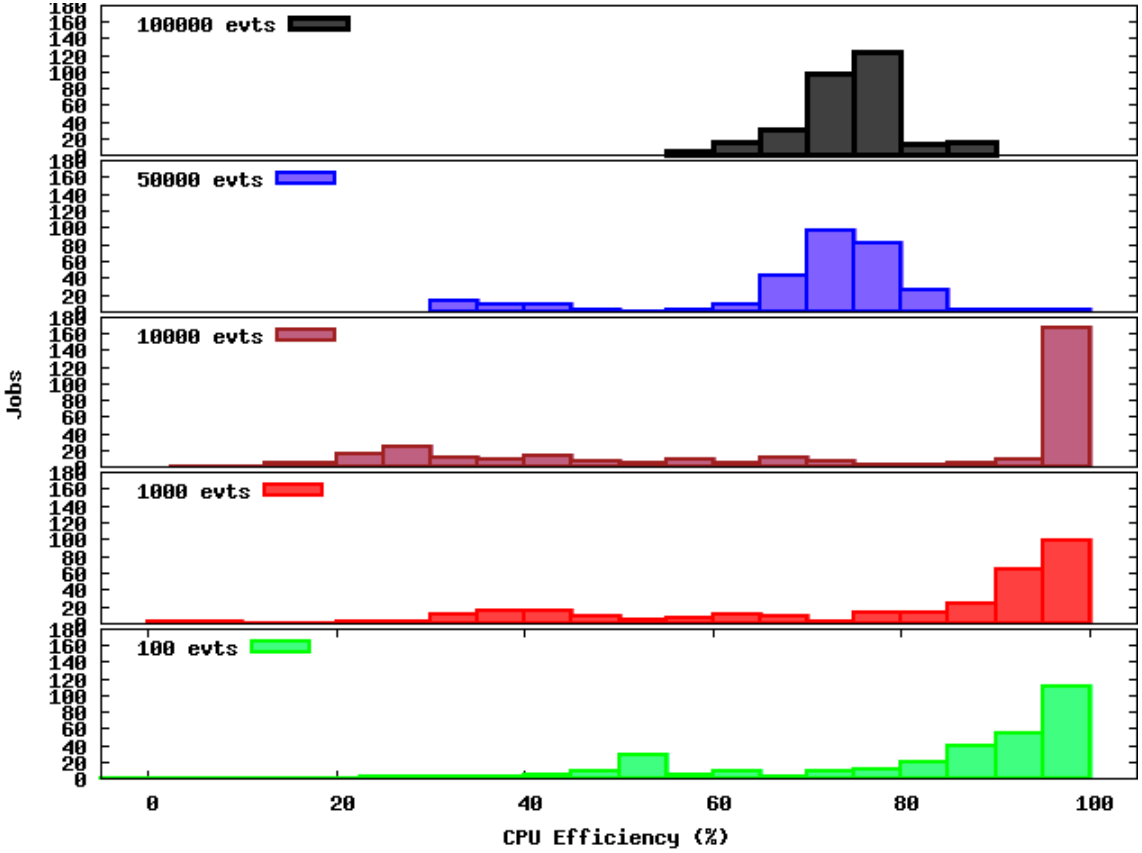
Results (I)

- Repeat the execution of the 300 jobs as a function of the number of events at IFCA

	CPU Eff. (%)	
Analysis over	**Before**	**After**
100 evts.	45	84
1,000 evts.	57	78
10,000 evts.	44	77
50,000 evts.	34	70
100,000 evts.	28	72

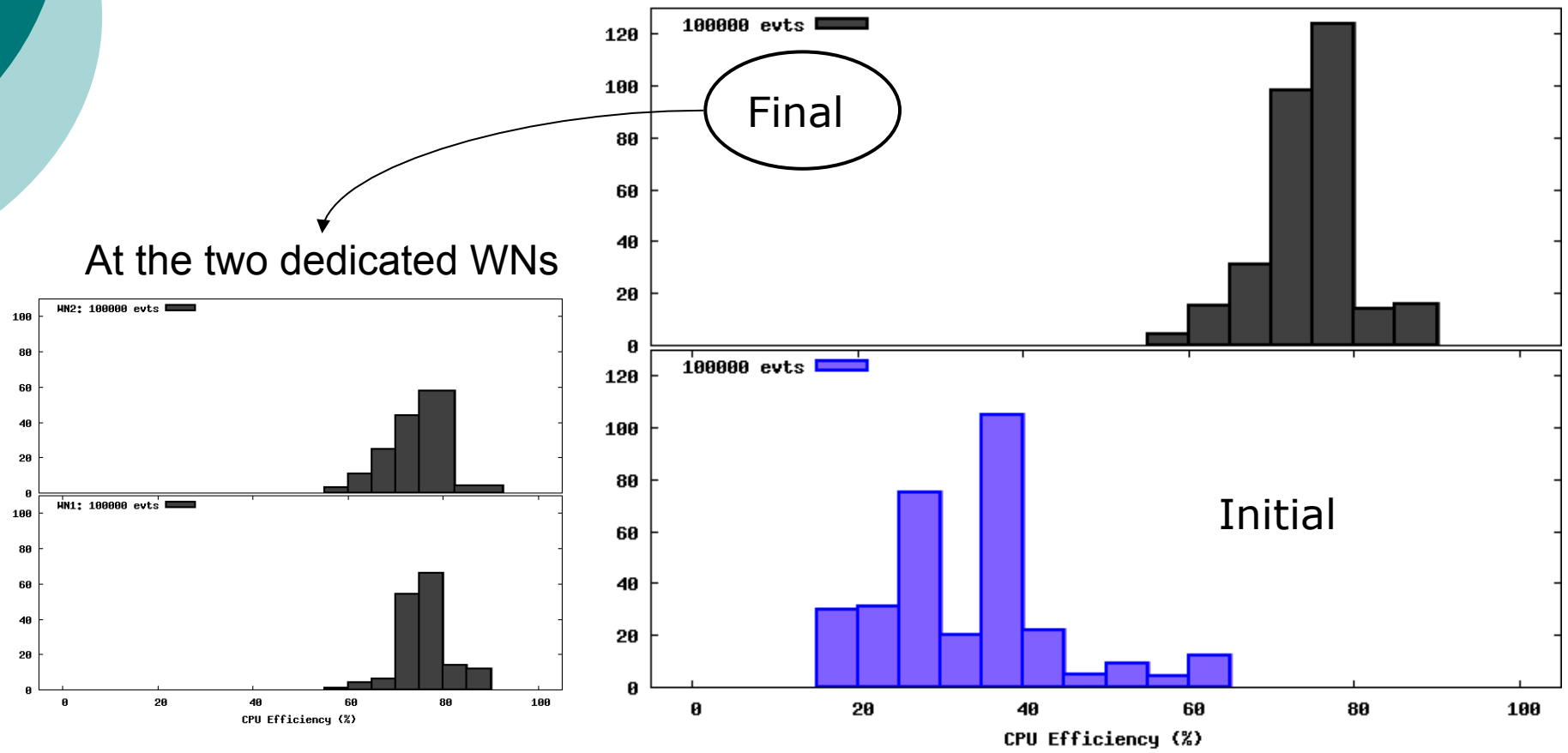
Large spread of results for the skims with a small number of events

After implemented changes



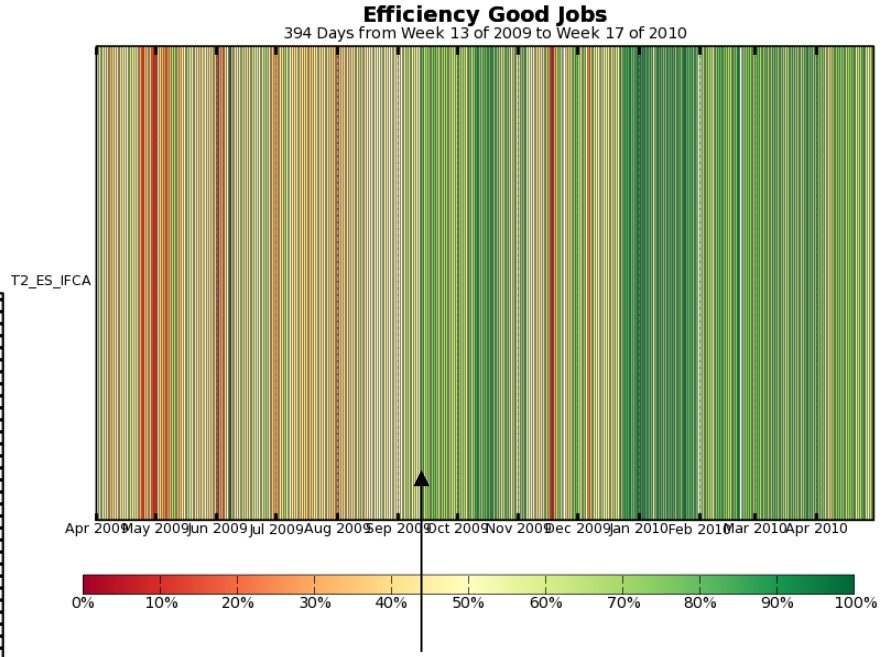
Results (II)

- For skims with 100,000 events



Current Status

- Average and evolution of the CPU efficiency after the presented modifications were implemented



Optimization of the Storage System

Conclusions

- The presented work is the first attempt to increase the CPU efficiency at IFCA
- A higher CPU efficiency implies:
  - A higher job throughput => a better usage of the resources
- The storage system configuration has an impact on the CPU efficiency, by optimizing the storage system configuration in three different areas:
  - File system
  - LAN network
  - Storage hardware

  the analysis CPU efficiency at IFCA increases from roughly 30% to 70%
- Since the implemented changes affect the general performance of the site, all communities profit from the obtained rise
- The ultimate goal of these studies is to dynamically achieve the best possible performance at the site at any time