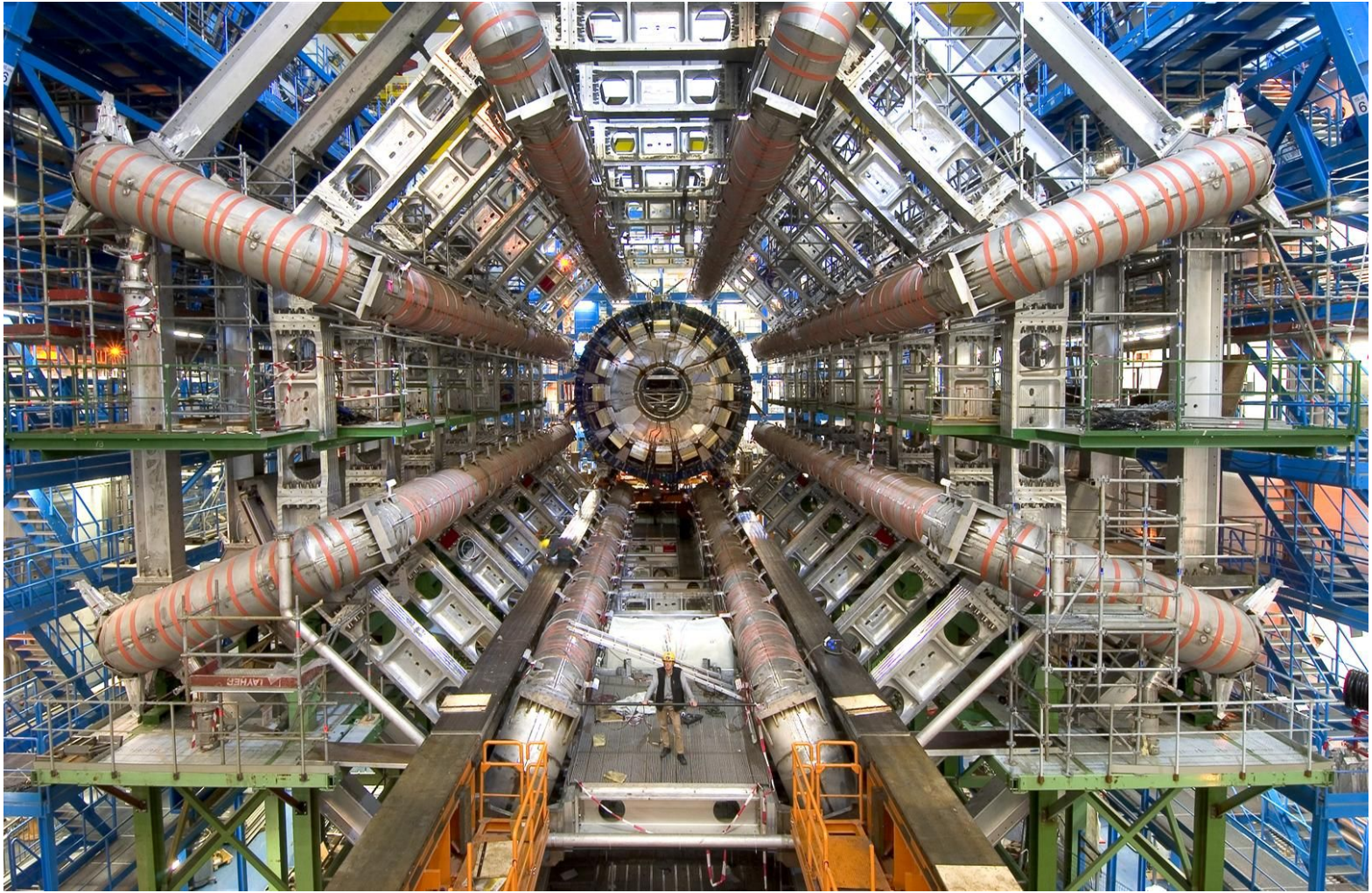
Barcelona Supercomputer Center Integration in the computing of ATLAS

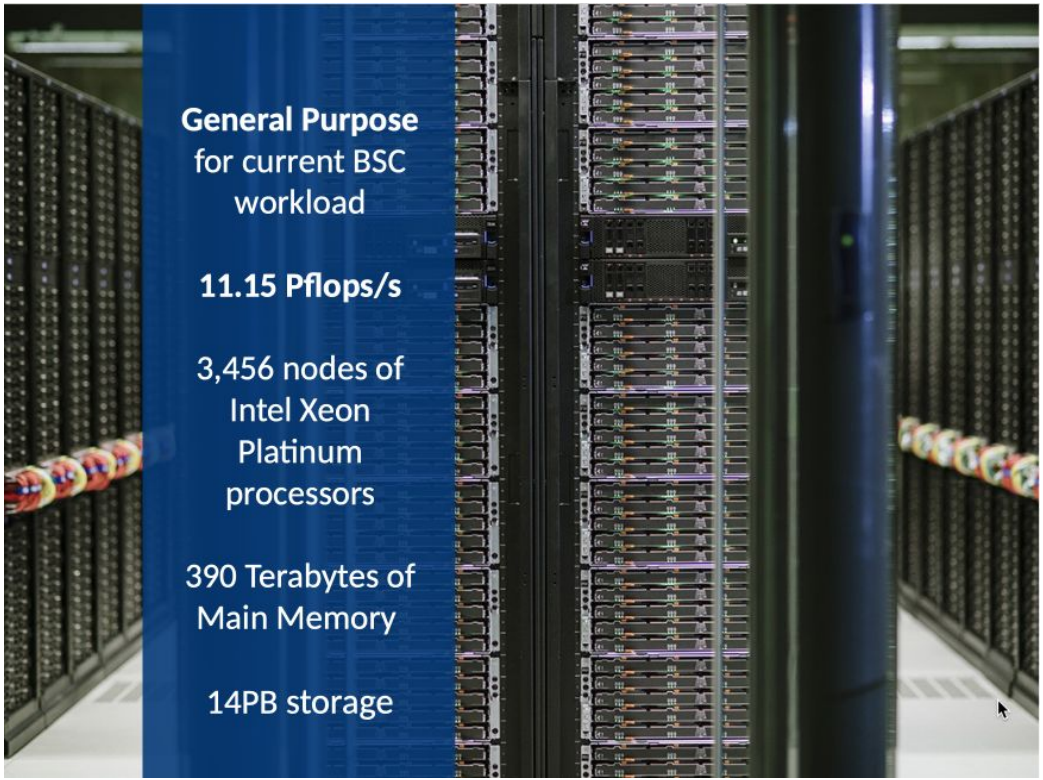
Andrés Pacheco Pages

IFAE Pizza Seminar - Wednesday 29 April 2020

IFAE Institut de Física d'Altes Energies
EXCELENCIA SEVERO OCHOA
BIST Barcelona Institute of Science and Technology
Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas
EXCELENCIA MARÍA DE MAEZTU
CIEMAT física de partículas
PIC port d'informació científica



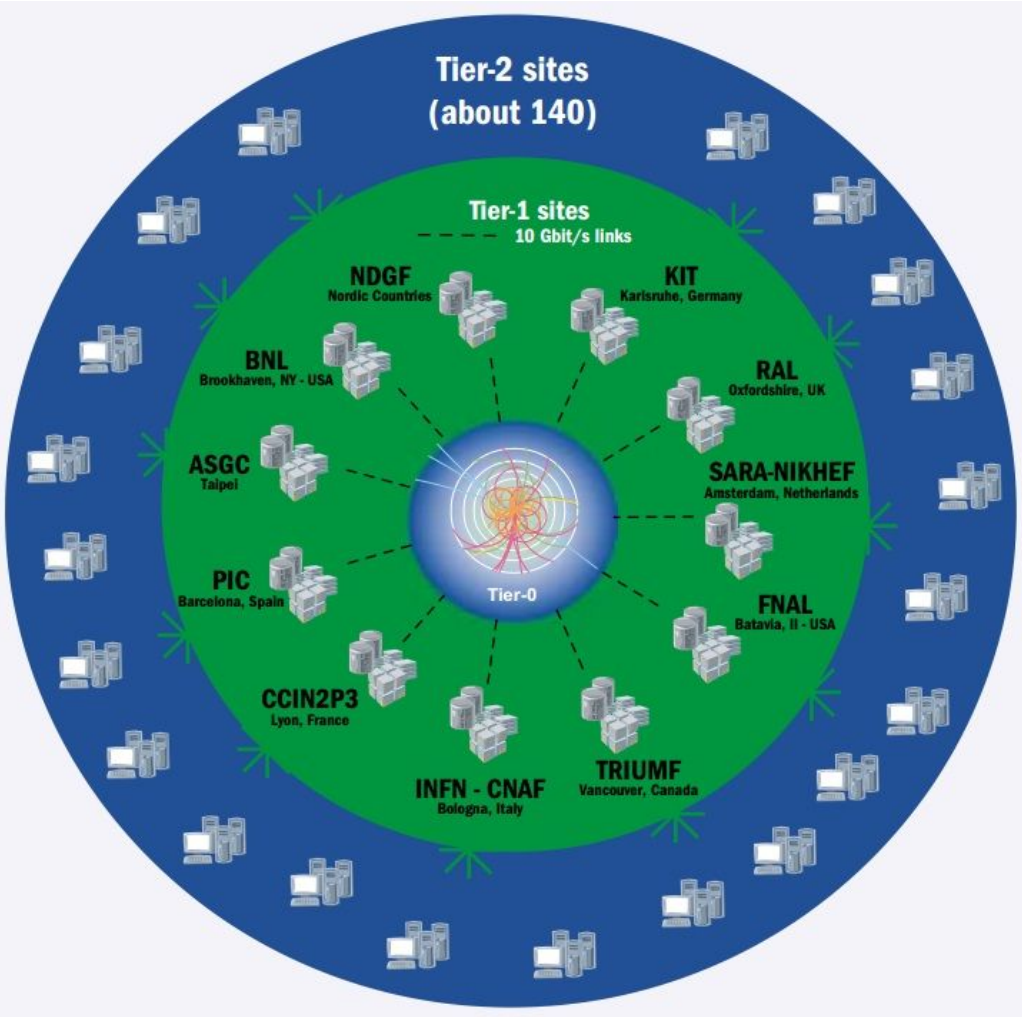
MareNostrum4 Picture

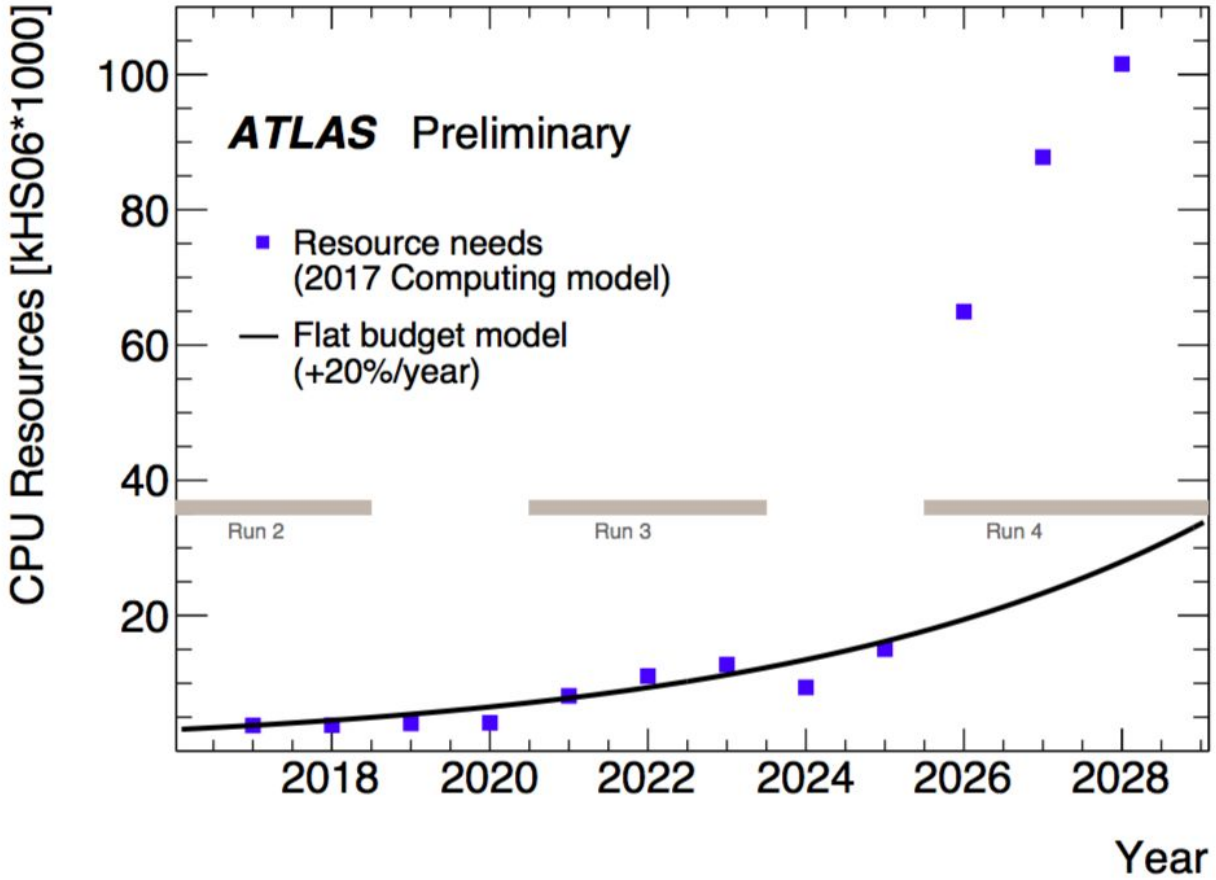


- Each node has two Intel Xeon Platinum chips, each with 24 processors, amounting to a **total of 165,888 processors** and a main memory of **2 GB RAM per processor**.
- Batch system: **SLURM**
- Operating system:**SUSE Linux 6**
- Shared file system: **GPFS**

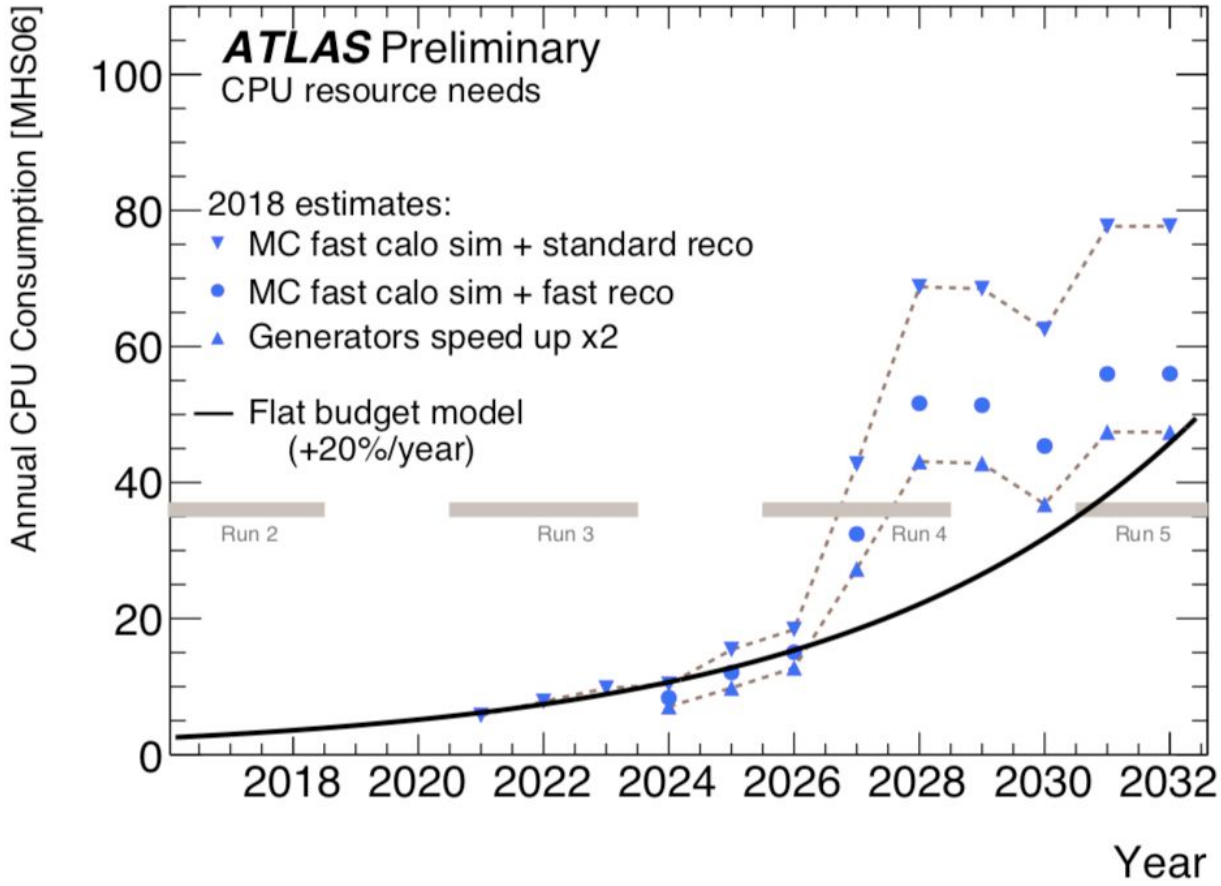
https://www.bsc.es/es/marenostrum/minotauro

Minotauro at BSC


Tier-2 sites
(about 140)
Tier-1 sites
10 Gbit/s links
NDGF
Nordic Countries
KIT
Karlsruhe, Germany
BNL
Brookhaven, NY - USA
RAL
Oxfordshire, UK
ASGC
Taipei
SARA-NIKHEF
Amsterdam, Netherlands
PIC
Barcelona, Spain
Tier-0
FNAL
Batavia, Il - USA
CCIN2P3
Lyon, France
INFN - CNAF
Bologna, Italy
TRIUMF
Vancouver, Canada



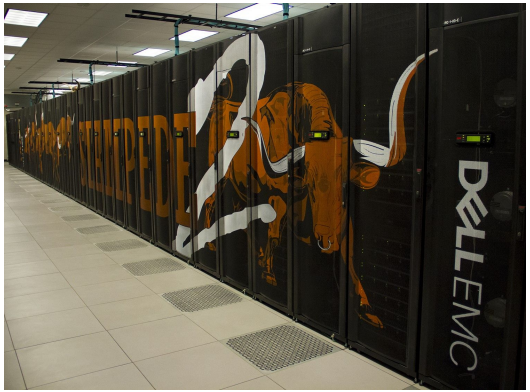
OLD ESTIMATES

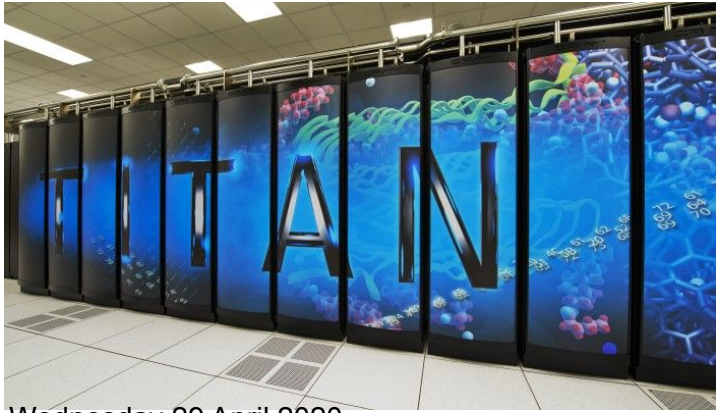


NEW ESTIMATES

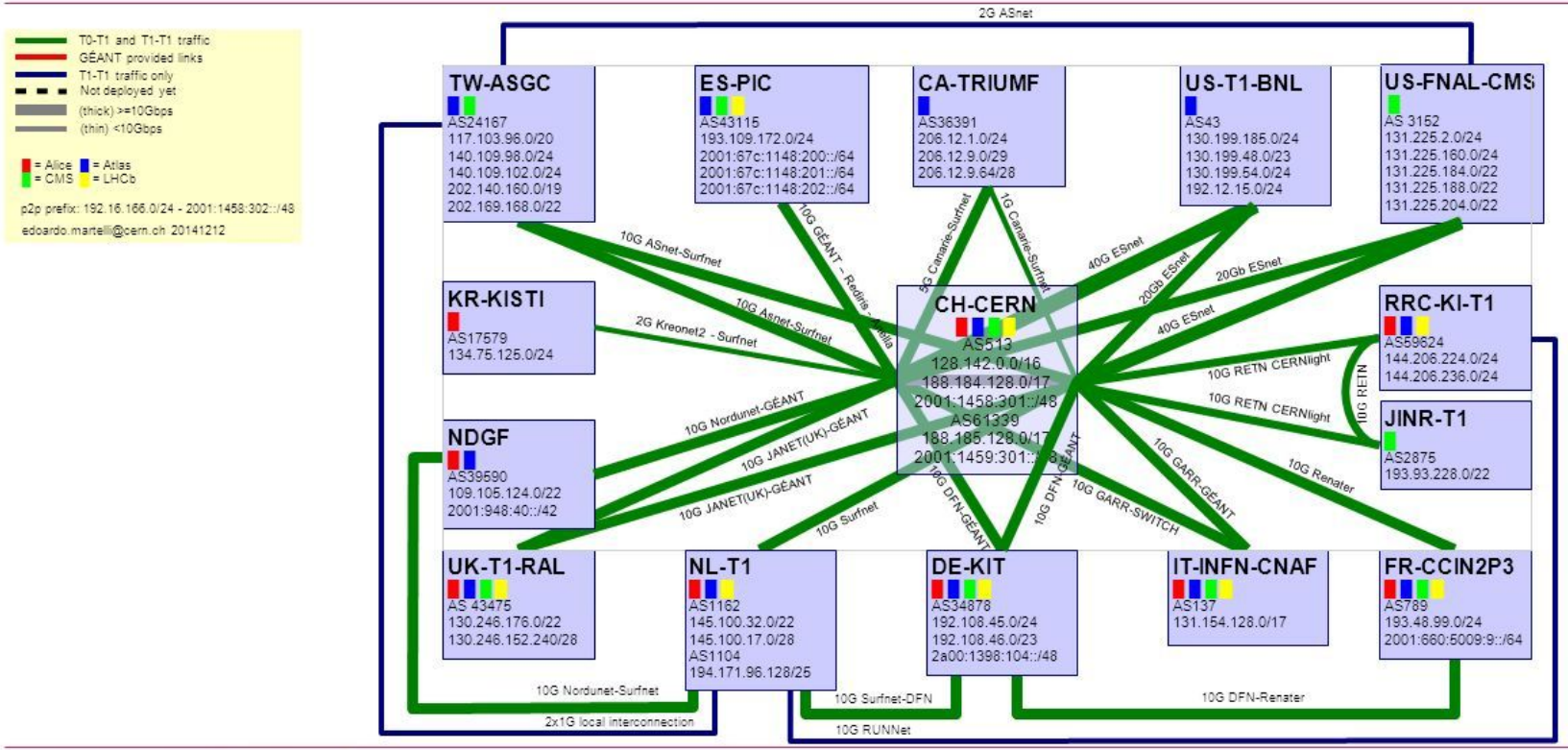

CRAY
THETA
Argonne
ENERGY
Office of Science
intel
CRAY


CRAY
Cori


STAMPEDE2
DELL EMC

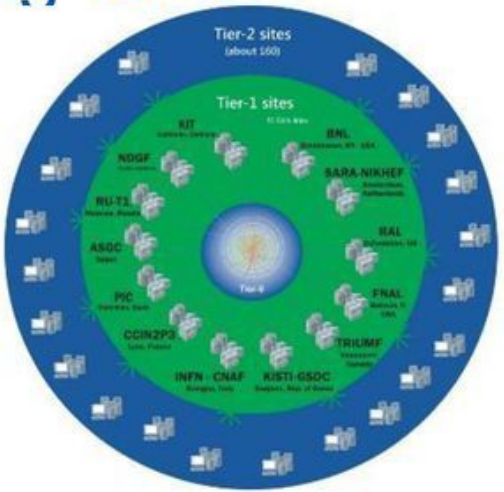

TITAN

LHCOPN


T0-T1 and T1-T1 traffic
GÉANT provided links
T1-T1 traffic only
Not deployed yet
(thick) >=10Gbps
(thin) <10Gbps
= Alice = Atlas
= CMS = LHCb
p2p prefix: 192.16.166.0/24 - 2001:1458:302::/48
edoardo.martelli@cern.ch 20141212
2G ASnet
TW-ASGC
AS24167
117.103.96.0/20
140.109.98.0/24
140.109.102.0/24
202.140.160.0/19
202.169.168.0/22
ES-PIC
AS43115
193.109.172.0/24
2001:67c:1148:200::/64
2001:67c:1148:201::/64
2001:67c:1148:202::/64
CA-TRIUMF
AS36391
206.12.1.0/24
206.12.9.0/29
206.12.9.64/28
US-T1-BNL
AS43
130.199.185.0/24
130.199.48.0/23
130.199.54.0/24
192.12.15.0/24
US-FNAL-CMS
AS 3152
131.225.2.0/24
131.225.160.0/24
131.225.184.0/22
131.225.188.0/22
131.225.204.0/22
KR-KISTI
AS17579
134.75.125.0/24
CH-CERN
AS513
128.142.0.0/16
188.184.128.0/17
2001:1458:301::/48
AS61339
188.185.128.0/17
2001:1459:301::/48
RRC-KI-T1
AS59624
144.206.224.0/24
144.206.236.0/24
JINR-T1
AS2875
193.93.228.0/22
NDGF
AS39590
109.105.124.0/22
2001:948:40::/42
UK-T1-RAL
AS 43475
130.246.176.0/22
130.246.152.240/28
NL-T1
AS1162
145.100.32.0/22
145.100.17.0/28
AS1104
194.171.96.128/25
DE-KIT
AS34878
192.108.45.0/24
192.108.46.0/23
2a00:1398:104::/48
IT-INFN-CNAF
AS137
131.154.128.0/17
FR-CCIN2P3
AS789
193.48.99.0/24
2001:660:5009:9::/64
10G ASnet-Surfnet
10G GÉANT - Reditis - Nordia
5G Canarie-Surfnet
1G Canarie-Surfnet
40G ESnet
20Gb ESnet
20Gb ESnet
40G ESnet
2G Kreonet2 - Surfnet
10G ASnet-Surfnet
10G RETN CERNlight
10G RETN
10G RETN CERNlight
10G Nordunet-GÉANT
10G JANET(UK)-GÉANT
10G JANET(UK)-GÉANT
10G Surfnet
10G DFN-GÉANT
10G DFN-GÉANT
10G GARR-SWITCH
10G GARR-GÉANT
10G Renater
10G Nordunet-Surfnet
2x1G local interconnection
10G Surfnet-DFN
10G RUNNet
10G DFN-Renater

LHC Data Distribution: WLCG
Worldwide LHC Computing Grid

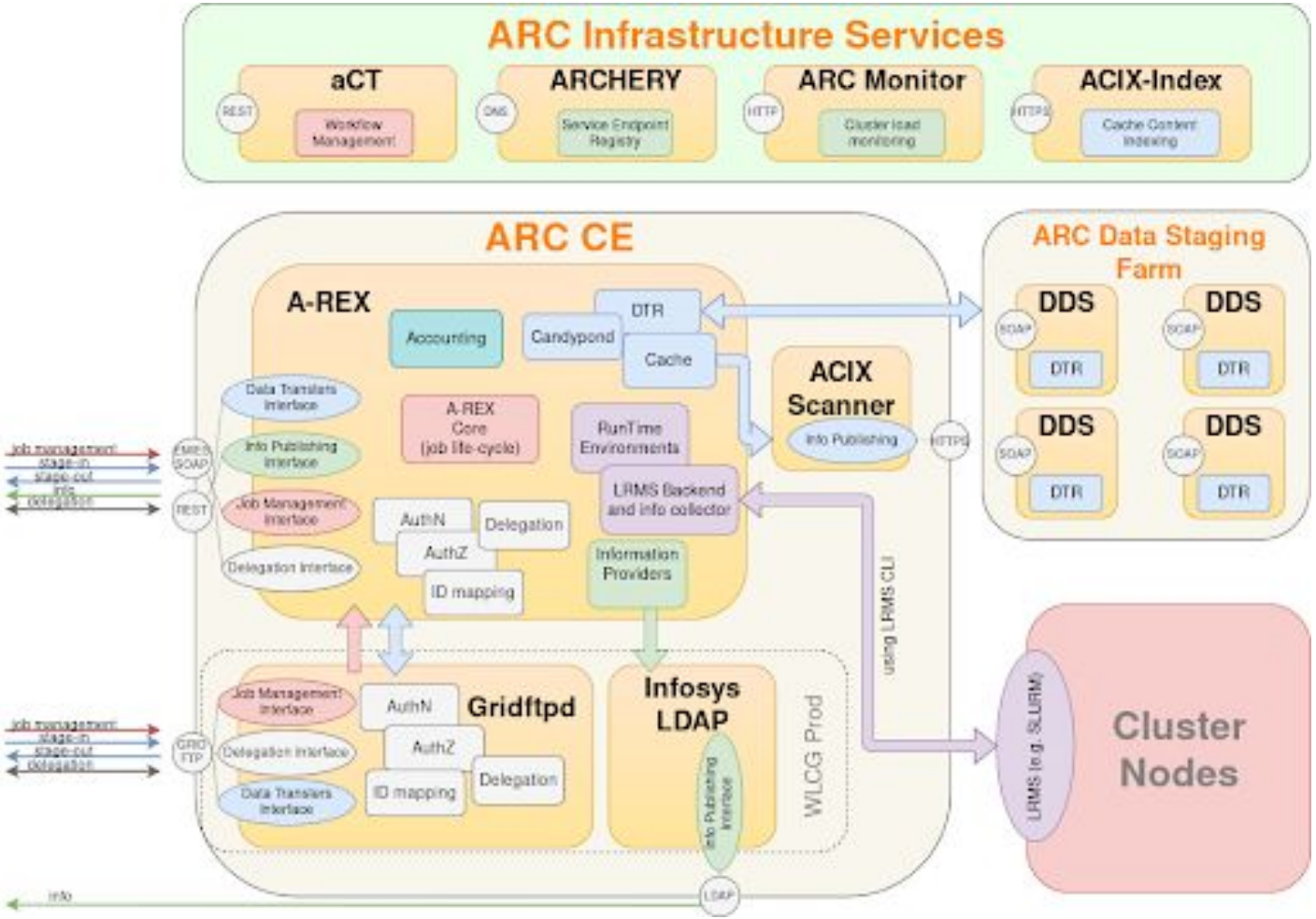
- The Worldwide LHC Computing Grid (WLCG) is a global collaboration of 170 data centres around the world, in 42 countries
- The CERN data centre (Tier-0) distributes the LHC data worldwide to the other WLCG sites (Tier-1 and Tier-2)
- WLCG provides global computing resources to store, distribute and analyse the LHC data
- The resources are distributed – for funding and sociological reasons



Tier-0 (CERN):
- Initial data reconstruction
- Data distribution
- ***Data recording & archiving***

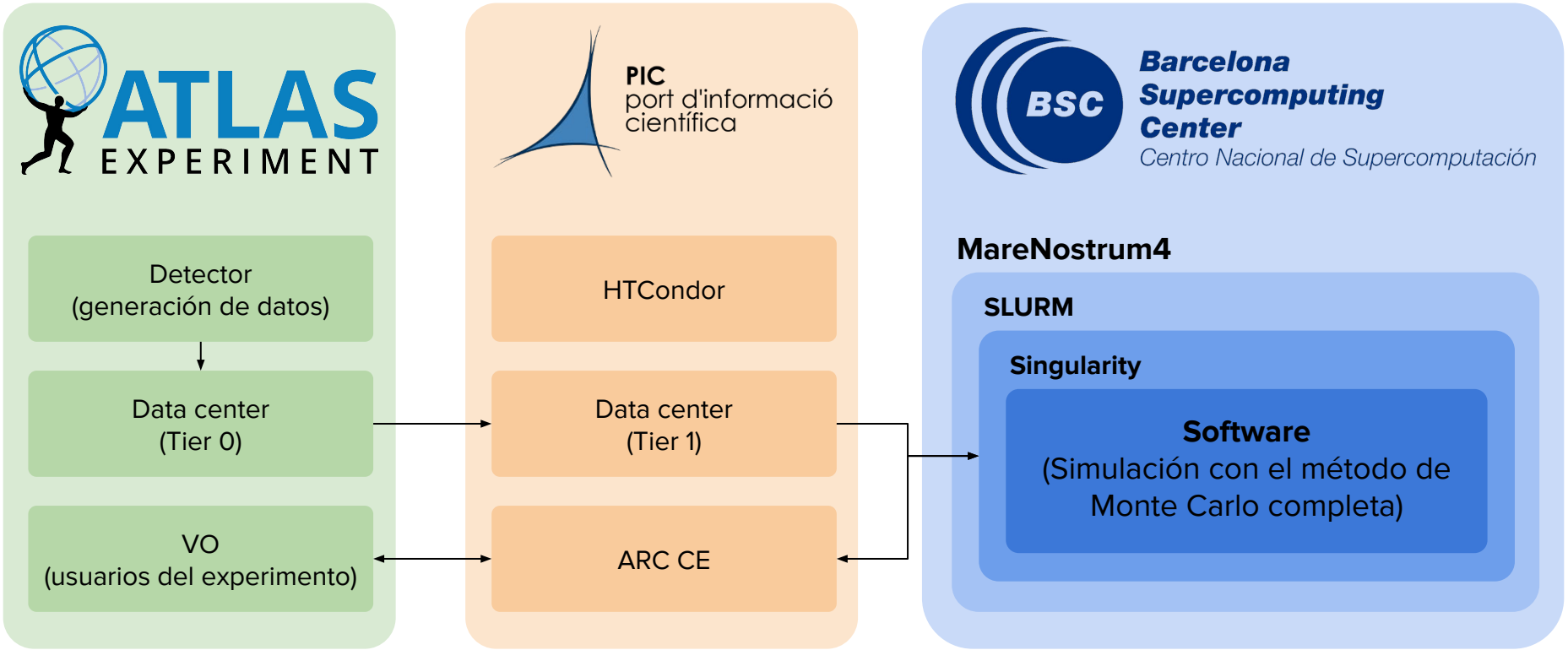
Tier-1 (13 centres):
- ***Permanent storage***
- Re-processing
- Analysis

Tier-2 (~140 centres):
- Simulation
- End-user analysis


ARC Infrastructure Services
aCT
Workflow Management
ARCHERY
Service Endpoint Registry
ARC Monitor
Cluster load monitoring
ACIX-Index
Cache Content Indexing
ARC CE
A-REX
Accounting
Candypond
DTR
Cache
Data Transfers Interface
Info Publishing Interface
Job Management Interface
Delegation Interface
A-REX Core (job life-cycle)
RunTime Environments
LRMS Backend and info collector
AuthN
AuthZ
ID mapping
Delegation
Information Providers
ACIX Scanner
Info Publishing
ARC Data Staging Farm
DDS
DTR
SOAP
job management
stage-in
stage-out
info
delegation
A-REX SOAP
REST
GRID FTP
Gridftpd
Infosys LDAP
WLCG Prod
LDAP
using LRMS CLI
LRMS (e.g. SLURM)
Cluster Nodes

Workflow at MareNostrum4?

- We must **copy all the input files** using the DTN by mounting a sshfs file system between PIC and BSC.
- We must **submit the jobs** using the login nodes and running on validated Singularity images with all the software preloaded.
- We must **check the status of the jobs** using the login nodes.
- We must **retrieve the output files** using the sshfs filesystem.

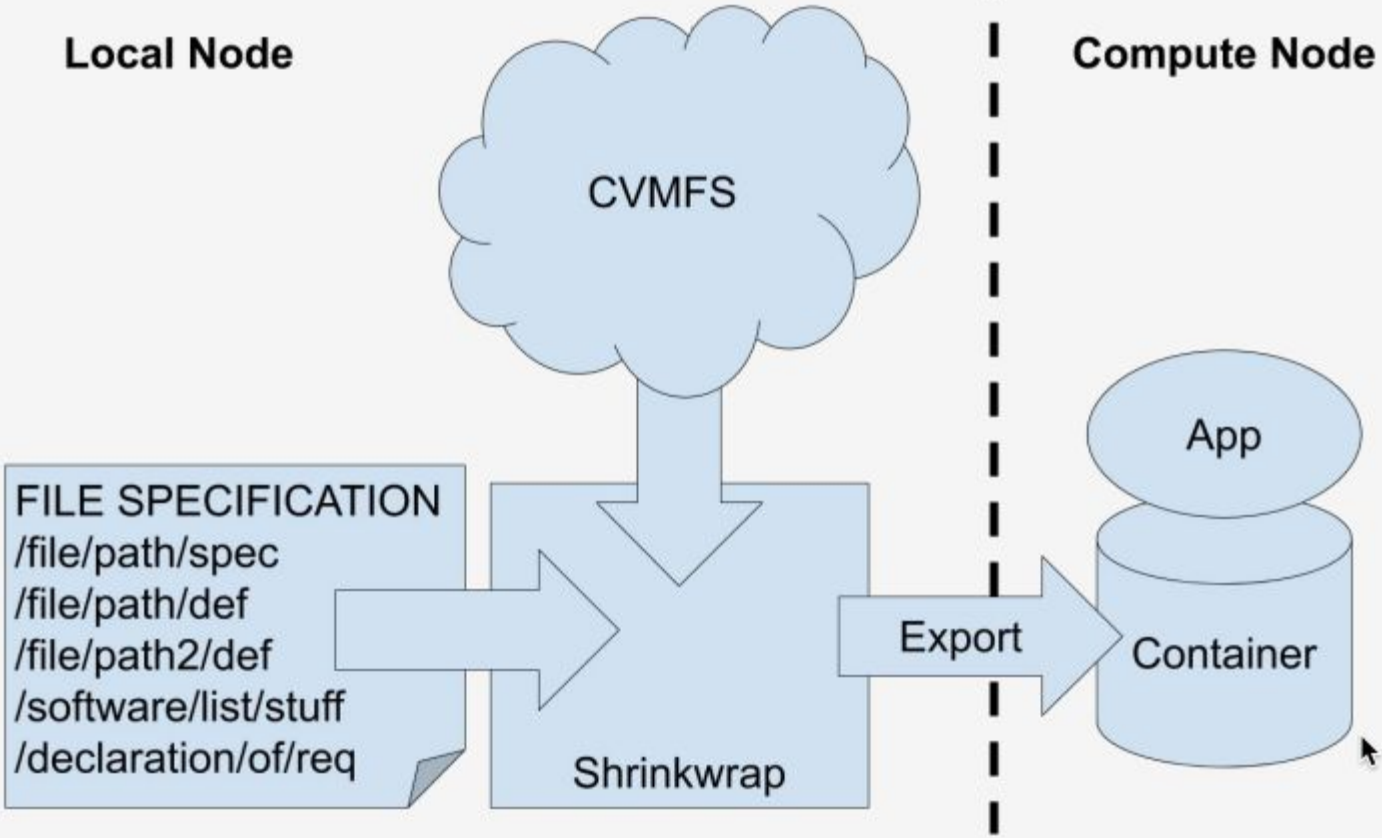

ATLAS
EXPERIMENT
PIC
port d'informació
científica
BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación
MareNostrum4
Detector
(generación de datos)
Data center
(Tier 0)
VO
(usuarios del experimento)
HTCondor
Data center
(Tier 1)
ARC CE
SLURM
Singularity
Software
(Simulación con el método de
Monte Carlo completa)

Pipeline

How to solve the problem of running on isolated worker nodes?

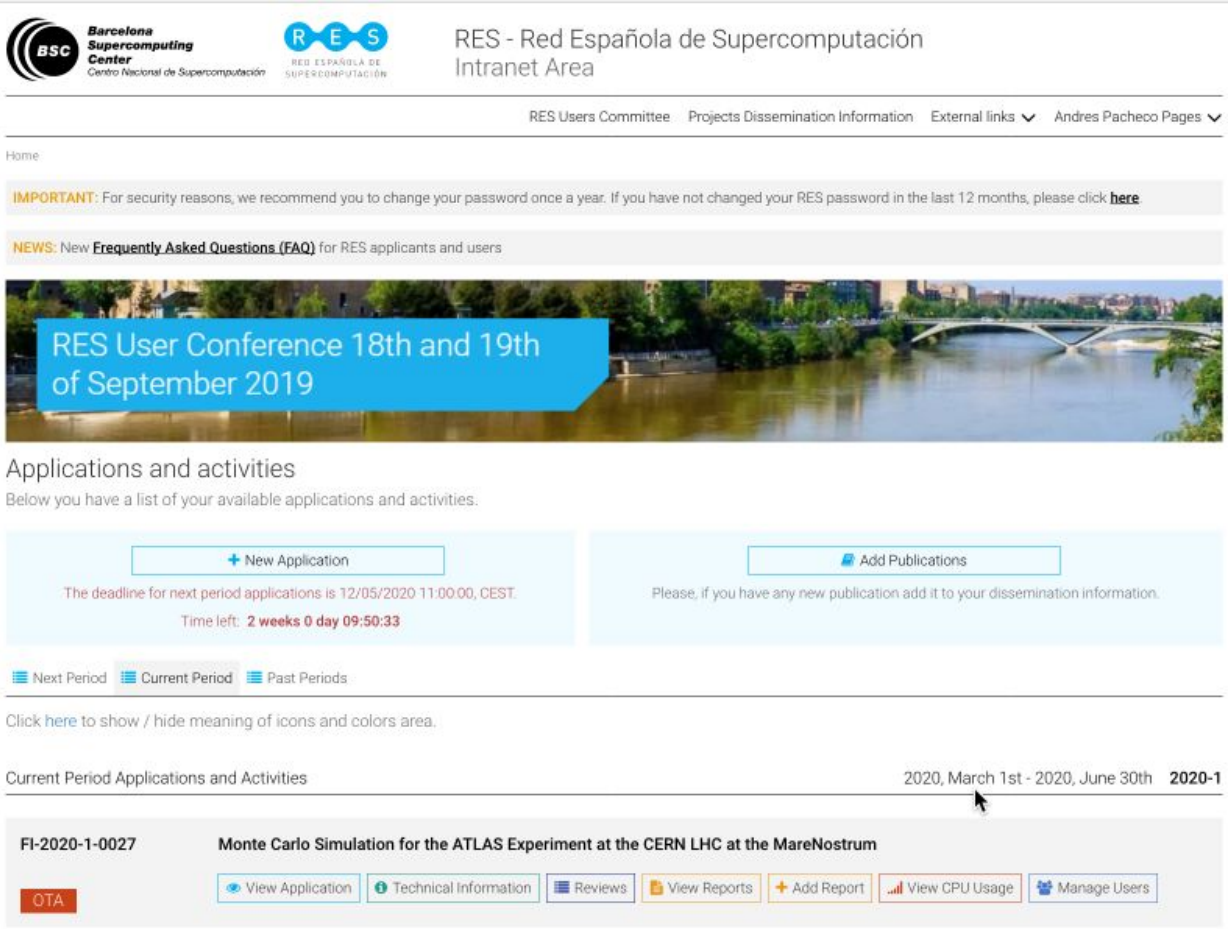
- The **working solution** now is **to create a filesystem with a partial copy of the ATLAS CVMFS filesystem repository and** including files containing **detector conditions**. The latest tool is called Shrinkwrap.
- This works because the **releases used for simulation are very few**.
- Then the filesystem is copied inside a **Singularity image** running a validated operating system (**CC7**).
- The **“problem” is to find the right list of files to be copied** to the image and the balance of the number of images to maintain: one image per ATLAS release, per workflow,... just run the parrot utility on a workflow to get an idea of the list of files accessed from cvmfs… thousands.

Use Case: HPC Environments



How do we get grants at MareNostrum4? RES

- The main **source of allocation of cpu hours** come from the **“Red Española de Supercomputación” (RES)** competitive program.
- Web is www.bsc.es/res
- You enter, you register, you request the time and then **you get approved or denied every 4 months**.
- You can **get hours** allocated in **any center of the RES**.

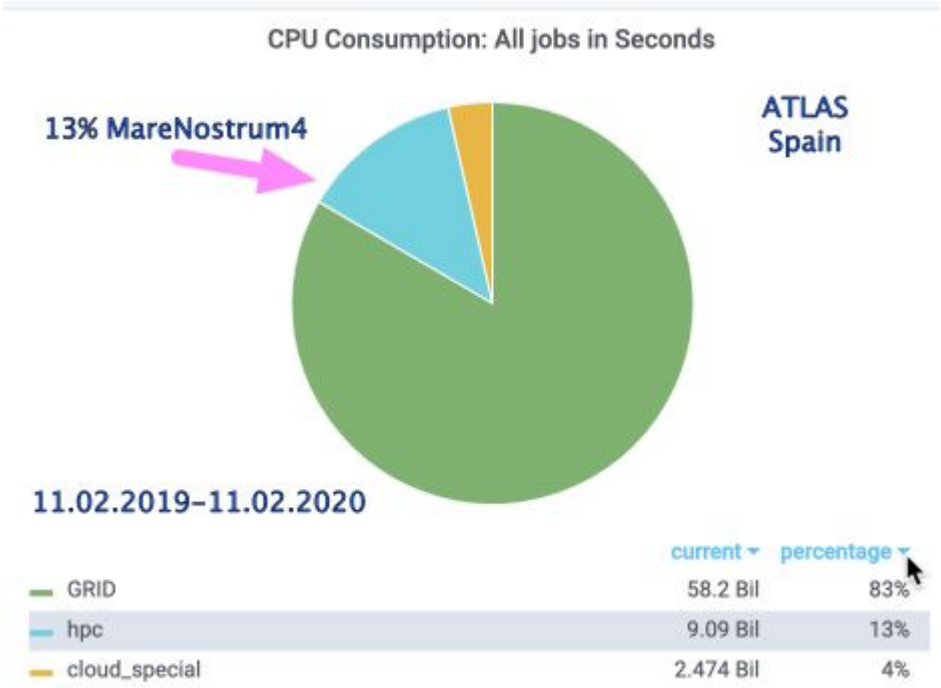

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
RES
RED ESPAÑOLA DE SUPERCOMPUTACIÓN
RES - Red Española de Supercomputación
Intranet Area
RES Users Committee
Projects Dissemination Information
External links
Andres Pacheco Pages
Home
IMPORTANT: For security reasons, we recommend you to change your password once a year. If you have not changed your RES password in the last 12 months, please click here
NEWS: New Frequently Asked Questions (FAQ) for RES applicants and users
RES User Conference 18th and 19th of September 2019
Applications and activities
Below you have a list of your available applications and activities.
New Application
The deadline for next period applications is 12/05/2020 11:00:00, CEST.
Time left: 2 weeks 0 day 09:50:33
Add Publications
Please, if you have any new publication add it to your dissemination information.
Next Period
Current Period
Past Periods
Click here to show / hide meaning of icons and colors area.
Current Period Applications and Activities
2020, March 1st - 2020, June 30th
2020-1
FI-2020-1-0027
Monte Carlo Simulation for the ATLAS Experiment at the CERN LHC at the MareNostrum
OTA
View Application
Technical Information
Reviews
View Reports
Add Report
View CPU Usage
Manage Users

How do we get grants at MareNostrum4? PRACE

- **Another program** we can apply for resources at BSC is **PRACE (Partnership for Advanced Computing in Europe)**
- Web is: http://www.prace-ri.eu/how-to-apply/
- There are **several types of calls from 2 months till 1 year**. You can get the allocation at the MareNostrum4 or at any of the HPCs in Europe. You select which you want explicitly.
- The **smallest grant is 2 months and 50 khours**. PRACE Preparatory Access type A.

	Distribution	Description		Access Criteria
Peer Reviewed evaluation	40%	PRACE		PRACE Access commitee
	40%	RES	General access	RES Access Committee
			Time in advance	
			Unexperienced users	
		Strategic projects (up to 7%)		Board of Trustees + RES annual evaluation
Expost evaluation	5%	Director time		Director BSC
	15%	BSC		Internal BSC

CPU from ATLAS jobs in Spanish sites: 13% correspond to jobs in MN4



- On the left, we have the CPU consumption pie chart of ATLAS jobs by resource type 1 year to date.
- ATLAS has already got off-pledge 13% of the Spanish contribution to the CPU from MareNostrum4 using queues at IFIC and PIC.

Source: ATLAS Job Accounting

Plans and the next move

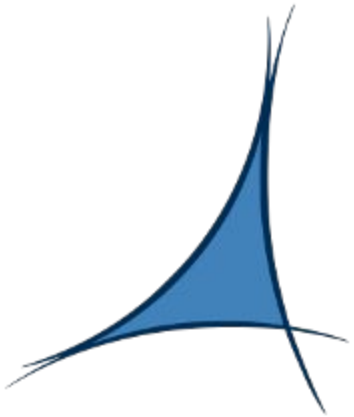
- Current plans is to increase the use of BSC thanks to the strategic program.
- We plan at PIC to run 1 million hours per month and increase each quadrimester.
- We need some work to increase the types of simulations we can run.
- After simulation the next target are the analysis jobs in containerized images.
  - Useful for analysis using GPUs

Can we replace the LHC computer centers?

- The answer is not.
- We need at least grid centers to receive the data from the experiment, store it on disk and **tape**, distribute, and **reprocess** the data. As well as simulate and analyze.
- The same is valid for simulated data once is produced, needs to be archived.
- The reconstruction of the data needs access to the databases of detector information, which is hard to upload to any supercomputer center.

Summary and conclusions

- We have managed to integrate the ATLAS Simulation jobs into the MareNostrum 4.
- The BSC has included the LHC computing in the list of strategic projects.
- We expect that the transition to MareNostrum 5 can be straightforward with 17 times more computing power in 2021.
- We still need grid computing for the LHC
  - Still many workflows cannot run in the BSC due to the lack of connectivity
  - We need to store, distribute and archive to tape the data.
- Thanks to the work of Carlos Acosta (PIC) and Elvis Diaz (UAB Student), all the PIC team and the collaboration with IFIC.



PIC
port d'informació
científica

FASES DE LA DESESCALADA

MAYO

L	M	X	J	V	S	D
	28	29	30	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

JUNIO

L	M	X	J	V	S	D
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

FASE 0 FASE 1 FASE 2 FASE 3

Salida controlada de menores	Apertura de comercio excepto centros comerciales.	Apertura del interior de locales a ⅓ de su aforo	Flexibilización de la movilidad general.
Deporte individual al aire libre	Apertura de restauración con un 30% de ocupación en terrazas.	Apertura excepcional de centros escolares para clases de refuerzo o la selectividad.	Centros comerciales a un 50% de su capacidad y con una distancia de 2 metros.
Locales con cita previa para tener llevar comida a domicilio	Apertura de hoteles y alojamientos turísticos excluyendo zonas comunes.	Cines, teatros y similares a ⅓ de su aforo.	Mayor aforo en restauración, preservando las distancias de seguridad.
Entrenamiento individual de deportistas profesionales	Lugares de culto a ⅓ de su capacidad	Equipamientos culturales a ⅓ de su aforo. Si es al aire libre, máximo 400 personas sentadas.	
Preparación de todos los locales públicos con medidas de protección	Apertura de centros de alto rendimiento deportivo.	Lugares de culto al 50% de su aforo.	

Conceptos generales:

- La provincia es la unidad territorial de medición.
- La duración de las fases será como mínimo de dos semanas cada una y el avance irá condicionado a indicadores de salud pública y a la evolución de los datos.
- La desescalada será asimétrica según la evolución de cada provincia.
- Inicio de la nueva normalidad como mínimo a partir del 25 de junio: Se permite el desplazamiento entre provincias y se siguen manteniendo las normas de seguridad y distancia social.