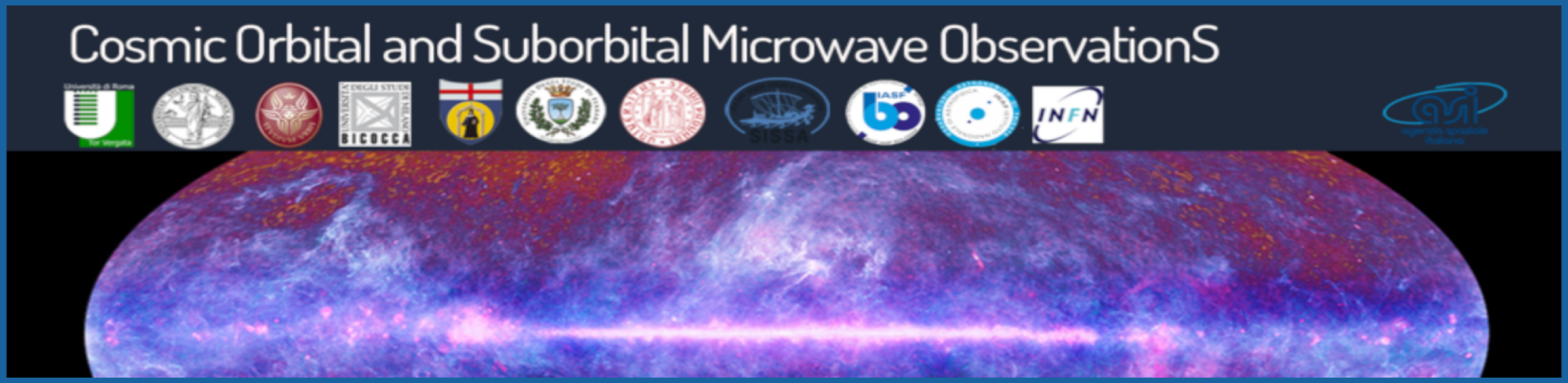



4th ASI/COSMOS Workshop: Ground-based CMB experiments

4-5 March 2019
Dipartimento di Fisica - Università degli Studi di Milano

# Computing infrastructure in Europe, challenges in data analysis

Francesco Piacentini

Sapienza - University of Rome - Physics Department

**With contribution from Paolo Natoli, Maurizio Tomasi, Andrea Zacchei**

# Outline

- Needs
- Available resources
- Evolution of algorithms
- Conclusion

# Increasing computational requirements

- Computational need for CMB experiments are driven by
  - Montecarlo realization of the observation, for:
    - Bias estimation
    - Uncertainty propagation

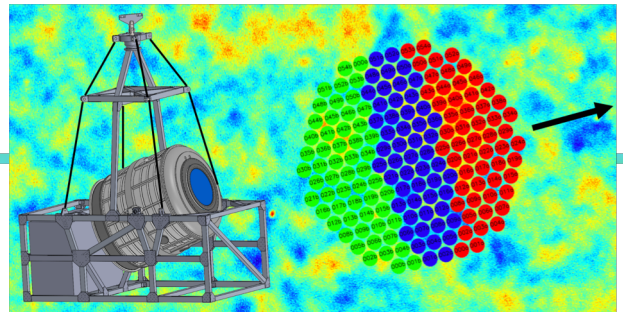
      Statistical impact (noise propagation)

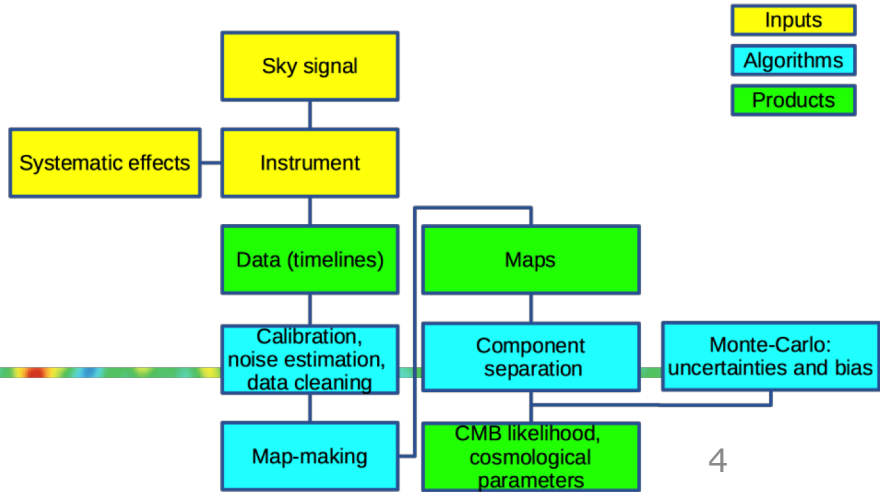
      Systematic effects impact (with demanding thresholds)

      including covariances
  - Analysis of timeline
    - Increasing number of detectors
    - Increasing integration time
    - Increasing angular resolution
    - Increasing spectral resolution
  - Specific to ground (and balloon)
    - Simulation and removal of atmospheric contamination
    - Simulation and removal of sidelobe pick-up

# Example: LSPE SWIPE example (CORI Nersc)

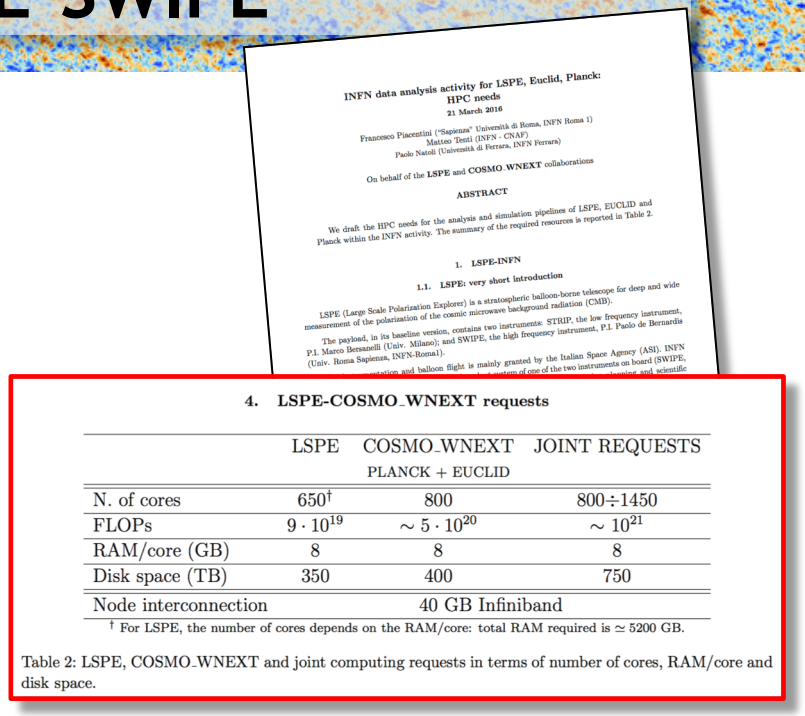
- SWIPE Instrument simulator:
  - Parallel Fortran code
  - The more complex version requires **1.3 k core hr** for one simulation
  - 326 detectors (~100 per band); 16 GB RAM for each detector
  - **1000 simulations require 1.3 M core hr**
- Mapmaking, with correlated noise (G. De Gasperis):
  - Parallel Fortran code
  - Can produce pixel-pixel inverse noise covariance matrix
  - 320 cores x 10 hr with 10 GB RAM per core, for one map
  - **300 maps require 1 M core hr**
- Power spectrum estimation (QML) – **Bolpol**
  - MPI/OpenMP
  - Flexible (32 to 1000+ cores)





# Example: resources for LSPE-SWIPE

- INFN is willing to support the data analysis effort
- Coordination activity since 2016
  - in collaboration with INFN-Euclid
  - at CNAF first
  - at CINECA then
- No data activity yet!



**INFN data analysis activity for LSPE, Euclid, Planck: HPC needs**

21 March 2016

Francesco Piacentini ("Sapienza" Università di Roma, INFN Roma 1)
Matteo Tenti (INFN - CNAF)
Paolo Natoli (Università di Ferrara, INFN Ferrara)

On behalf of the **LSPE** and **COSMO_WNEXT** collaborations

**ABSTRACT**

We draft the HPC needs for the analysis and simulation pipelines of LSPE, EUCLID and Planck within the INFN activity. The summary of the required resources is reported in Table 2.

**1. LSPE-INFN**

**1.1. LSPE: very short introduction**

LSPE (Large Scale Polarization Explorer) is a stratospheric balloon-borne telescope for deep and wide measurement of the polarization of the cosmic microwave background radiation (CMB).

The payload, in its baseline version, contains two instruments: STRIP, the low frequency instrument, P.I. Marco Bersanelli (Univ. Milano); and SWIPE, the high frequency instrument, P.I. Paolo de Bernardis (Univ. Roma Sapienza, INFN-Roma1).

**4. LSPE-COSMO_WNEXT requests**

| | LSPE | COSMO_WNEXT | JOINT REQUESTS |
|---|---|---|---|
| | | PLANCK + EUCLID | |
| N. of cores | 650† | 800 | 800÷1450 |
| FLOPs | $9 \cdot 10^{19}$ | $\sim 5 \cdot 10^{20}$ | $\sim 10^{21}$ |
| RAM/core (GB) | 8 | 8 | 8 |
| Disk space (TB) | 350 | 400 | 750 |
| Node interconnection | | 40 GB Infiniband | |

† For LSPE, the number of cores depends on the RAM/core: total RAM required is $\simeq$ 5200 GB.

Table 2: LSPE, COSMO_WNEXT and joint computing requests in terms of number of cores, RAM/core and disk space.

- NERSC support going-on in the meanwhile
- Post-Planck experiments: mp107
- Allocation 2019:
  - 186 users
  - 70 Mcore hrs
  - 14 LSPE users
  - 700 kcore hrs each
  - **10 Mcore hrs for LSPE**

# Example: AME in Andromeda from SRT

- Proposal 33-18 (E. Battistelli et al.)
  - K band
  - 7 horns,
  - x 4 Stokes,
  - x 16384 spectral bins,
  - x 33 Hz sampling,
  - x 18 bits

  => 34.5 MB/s
- 10.0 hr / day => 1.2 TB/day
- 28 days total => **35 TB total**

- Data to be combined in a single map, after atmosphere subtraction,

  + maser search

# Example: PIXIE timeline simulation

- See: **Time-ordered data simulation and map-making for the PIXIE Fourier transform spectrometer** Sigurd Næss, Jo Dunkley, Alan Kogut, Dale Fixsen

  https://arxiv.org/abs/1710.06761

> However, it also has some important limitations. Because it stores the full spectrum/autocorrelation function in each pixel, its memory requirements scale poorly with resolution. This makes it impractical to investigate the effect of sub-resolution features (both spectrally and spatially) - to do so would require the data cube to be pixelized at many times higher resolution than the PIXIE output map, which would make the memory requirements of this approach prohibitively high. For example, for 0.1° spatial resolution and 5000 frequency bins, storing the full-sky autocorrelation function would need about 700 GB of RAM.

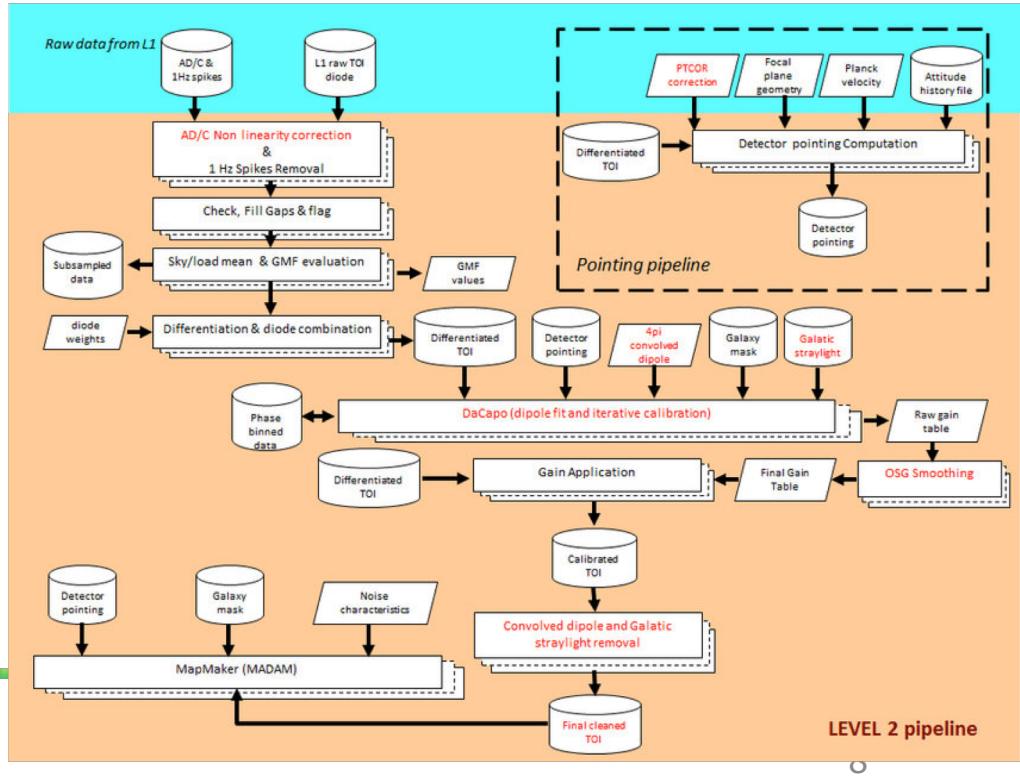
> If one assumes a frequency-independent beam, which should be a good approximation for PIXIE,

> In the end, we went for a Taylor expansion approach: The autocorrelation function is evaluated as a perturbation around a different but similar precomputed autocorrelation function. This is done differently for each sky component.

# Example: allocation for Planck LFI-DPC

- Resources at Planck-LFI Data Processing Center (A. Zacchei)
  - Testing cluster (donation by CINECA): 2008
  - First production cluster (donation by CINECA): 2008
  - Second production cluster: 2011
  - Simulations at NERSC (out of DPC budget)

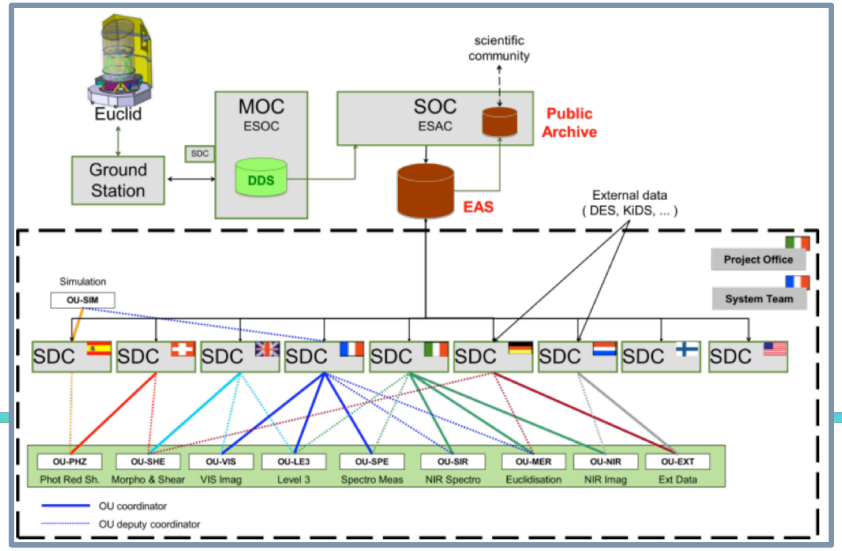
- Costs:
  - Power and maintenance 60 k€/yr
  - Cooling system: 100 k€
  - Second production cluster: 70 k€
  - Storage: 50 k€
  - Manpower ½ FTE/yr



# Example: allocation for Euclid

- More than 500,000 visible (VIS) and near infrared images (NISP imaging and NISP spectroscopy) that will be transferred to Earth daily.
- ~30 PB of images data
- ~10 billion sources
- 9 SDC (Science Data Centers)
- Algorithms in optimization phase
  - Including GPU based algorithms (presentation by Daniele Tavagnacco on the 6th)
- Italian SDC will contribute with 20% of total resources

- Costs are of **order a few millions € as Italian contribution**



# Resources NERSC

- NERSC has been the main provider of computational power for CMB experiments in the last decades

- CORI KNL

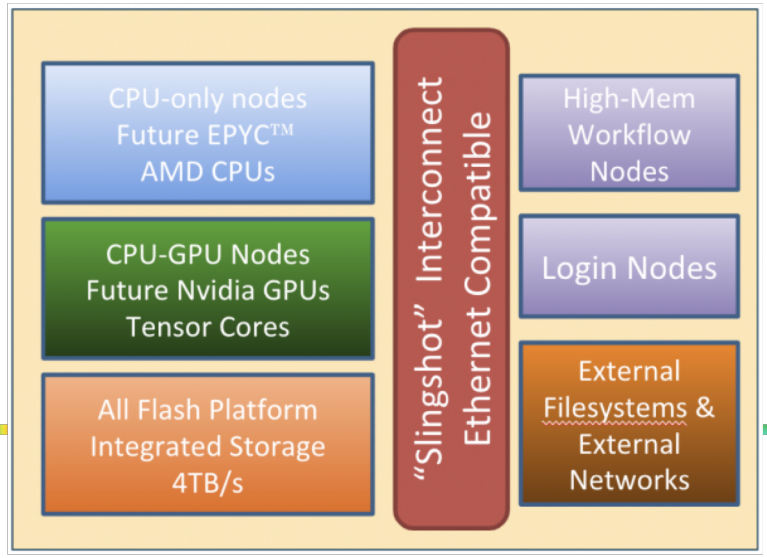
**29.5 PFlops**

| | | |
|---|---|---|
| KNL Cabinets | 54 | Each cabinet has 3 chassis; each chassis has 16 compute blades, each compute blade has 4 nodes |
| KNL Compute nodes | 9,688 | Each node is a single-socket Intel® Xeon Phi™ Processor 7250 ("Knights Landing") processor with 68 cores per node @ 1.4 GHz |
| | | Each core has two 512-bit-wide vector processing units. Each core has 4 hardware threads (272 threads total). Two cores form a tile. |
| | | 44.8 GFlops/core; 3 TFlops/node; 29.5 PFlops total (theoretical peak) |
| | | Each node has 96 GB DDR4 2400 MHz memory, six 16 GB DIMMs (102 GiB/s peak bandwidth). Total aggregate memory (combined with MCDRAM) is 1.09 PB. |
| | | Each node has 16 GB MCDRAM (multi-channel DRAM), > 460 GB/s peak bandwidth |
| | | Each core has its own L1 caches, with 64 KB (32 KiB instruction cache, 32 KB data). Each tile (2 cores) shares a 1MB L2 cache. |
| Interconnect | | Cray Aries with Dragonfly topology with 5.625 TB/s global bandwidth (Phase I). 45.0 TB/s global peak bisection bandwidth (Phase II). |

# Resources NERSC

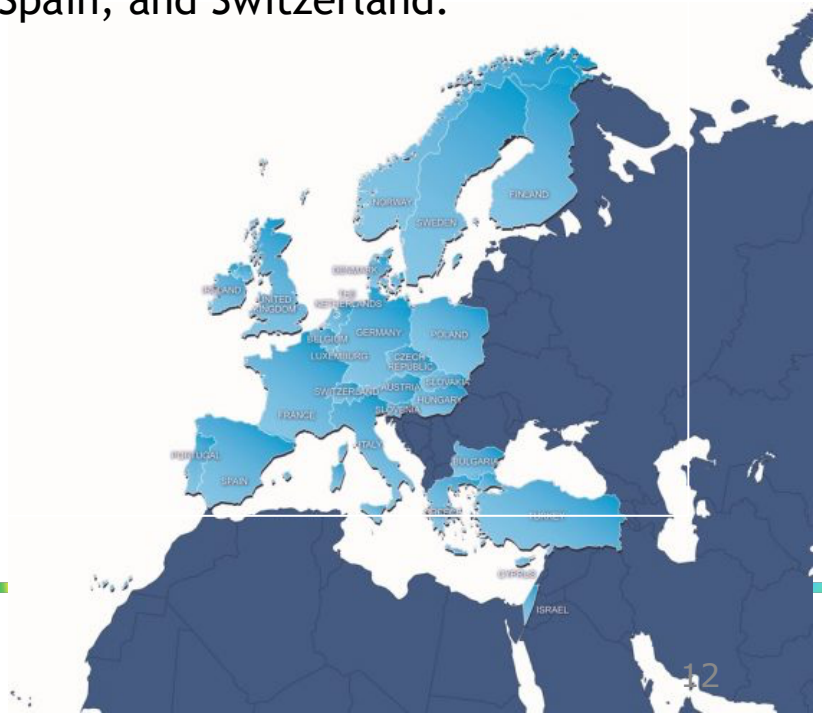
- NERSC has been the main provider of computational power for CMB experiments in the last decades

- PERLMUTTER (2020)
- Perlmutter, a Cray system, will be a heterogeneous system:
  - CPU-only nodes
  - GPU-accelerated nodes
  - performance of more than 3 times Cori, NERSC's current platform
    - **~ 100 PFlops**



# Resources in Europe: PRACE

- The mission of PRACE (Partnership for Advanced Computing in Europe) is
  - to enable high-impact scientific discovery
  - engineering research and development across all disciplines
  - to enhance European competitiveness for the benefit of society.
  - By offering world class computing and data management resources and services.
- Hosting Members: France, Germany, Italy, Spain, and Switzerland.



# Resources in Europe: PRACE

Hazel Hen, HLRS/GCS, Germany **7.42 PFlops**

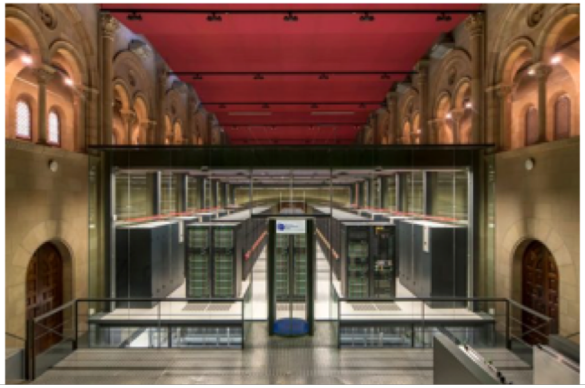

JUWELS, FZ Jülich/GCS, Germany **12 PFlops**


JUWELS Supercomputer, Copyright: Forschungszentrum Jülich / R.-U. Limbach

SuperMUC-NG, LRZ/GCS, Germany **26.9 PFlops**

JOLIOT Curie, GENCI, France **6.68 PFlops**


MareNostrum, BSC, Spain **11.4 PFlops**



Piz Daint, ETH Zurich/CSCS, Switzerland

# Resources: CINECA (in PRACE)

•**Marconi** among the most powerful supercomputer:
Rank 19 in November 2018.
**20 PFlops**

•**GALILEO**: renewed in March 2018 with **Intel Xeon E5-2697 v4 (Broadwell) nodes**, available for italian research community.

•**D.A.V.I.D.E.**: the energy-aware, High Performance Cluster, based on OpenPOWER8 servers and NVIDIA Tesla P100 GPUs.

| | CPU (mhz,core, ...) | Total cores / Total Nodes | Memory per node | Accelerator |
|---|---|---|---|---|
| ***MARCONI-A2*** | Intel Knights Landing<br>1x Intel Xeon Phi7250<br>@1.4GHz<br>68 cores each | 244800 / 3600 | 96 GB | - |
| ***MARCONI-A3*** | Intel SkyLake<br>2x Intel Xeon 8160<br>@2.1GHz<br>24 cores each | 72576+38016+43776 /<br>1512+792+912 | 192 GB | - |
| ***GALILEO*** | Intel Broadwell<br>2x Intel Xeon E5-2697<br>v4<br>@2.3GHz<br>18 cores each | 12960 / 360 | 128 GB | |
| ***D.A.V.I.D.E.*** | OpenPOWER8<br>NVIDIA Tesla P100<br>SXM2<br>@2GHz<br>16 cores each | 720 / 45 | | Tesla P100 |

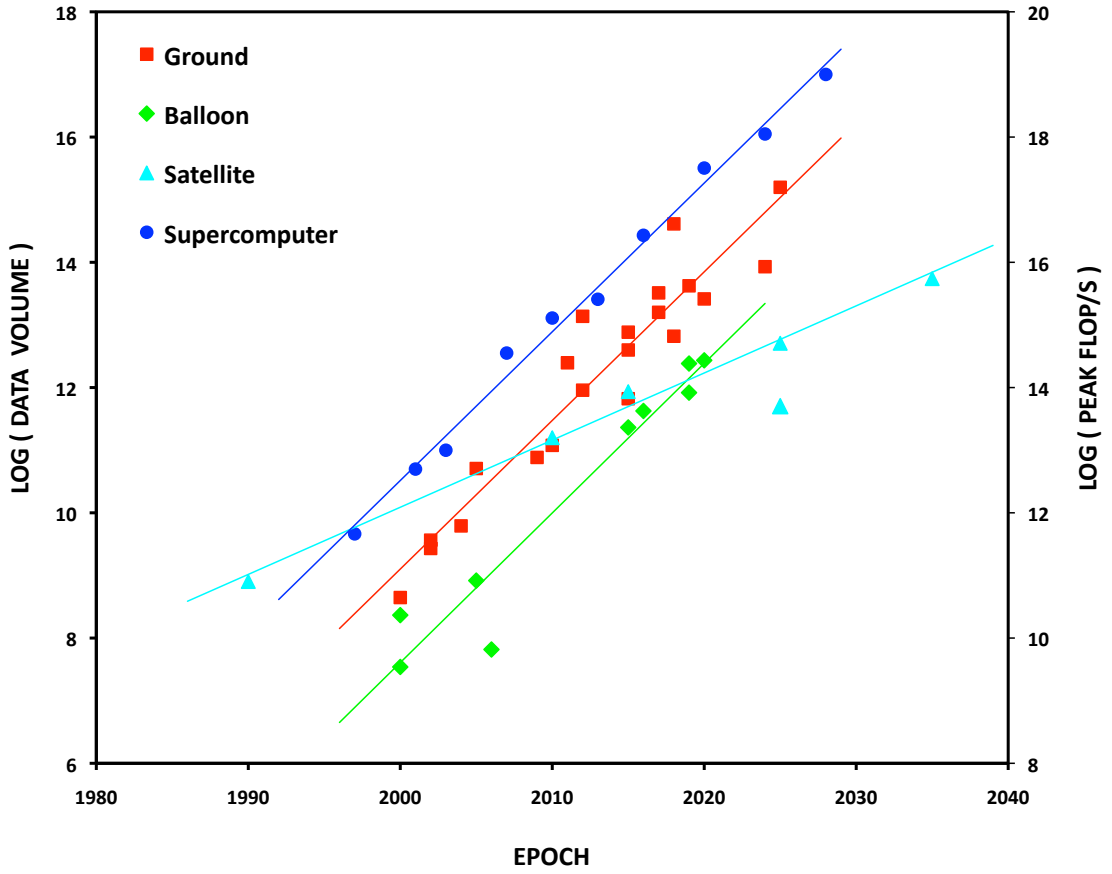
# Resources: CSC Finland

- **CSC – IT CENTER FOR SCIENCE LTD.** Espoo, Finland
  - https://www.csc.fi
  - **Sisu: Cray XC40 Supercomputer**
  - The most powerful supercomputer in Finland
  - Designed for High Performance Computing (HPC).
  - Sisu consists of nine cabinets, with a total theoretical peak performance of **1.7 PFlops**.
  - 1688 nodes, x two 12 core Intel Xeon E5-2690v3 (Haswell) 2,6 GHz CPUs
  - Total of 40512 cores
  - 64 GB of memory (2,67 GB/core) in each node

# Evolution of computer power and algorithms

# Scaling with time (Moore's law)

- Plot & data by J. Borrill 2016
- Projected scaling up of computing power (based on some version of Moore's law) allows in principle to scale up to cover forthcoming ground based experiments…



# … but… (Natoli - Villa Finaly 2016)

- Difficult to forecast the evolution of supercomputing resources in the next decade
- System are becoming severely energy constrained.
- Many experts speak openly of Moore's law coming to an end.
- Even ignoring the above, exploitation of available resources (when available) are limited by user concurrency and cost of flop unit.
  - *We are not the only community in need of significant computing power*
  - *Must find a balance between cheap flops offered on clogged computers and costly dedicated service.*
  - *Can European coordination play a role here?*
- Increasing size of data limits human direct intervention.
- Automatization is a must and complicates business.

# GPU based, smaller, performing systems

- Proposal *Grandi Attrezzature 2018*: "A SCALABLE ARTIFICIAL INTELLIGENCE SYSTEM FOR MACHINE AND DEEP LEARNING RESEARCH AND TRAINING AT SAPIENZA UNIVERSITÀ DI ROMA"
- NVIDIA DGX-2
  - 16 interconnected GPUs
  - 2 PFlops
  - 363.5 k€
- Need of software development: GPU based algorithms
  - Linear algebra with TensorFlow (www.tensorflow.org)
  - FFT with cuFFT
  - CUDA® compilation
  - …


**SYSTEM SPECIFICATIONS**

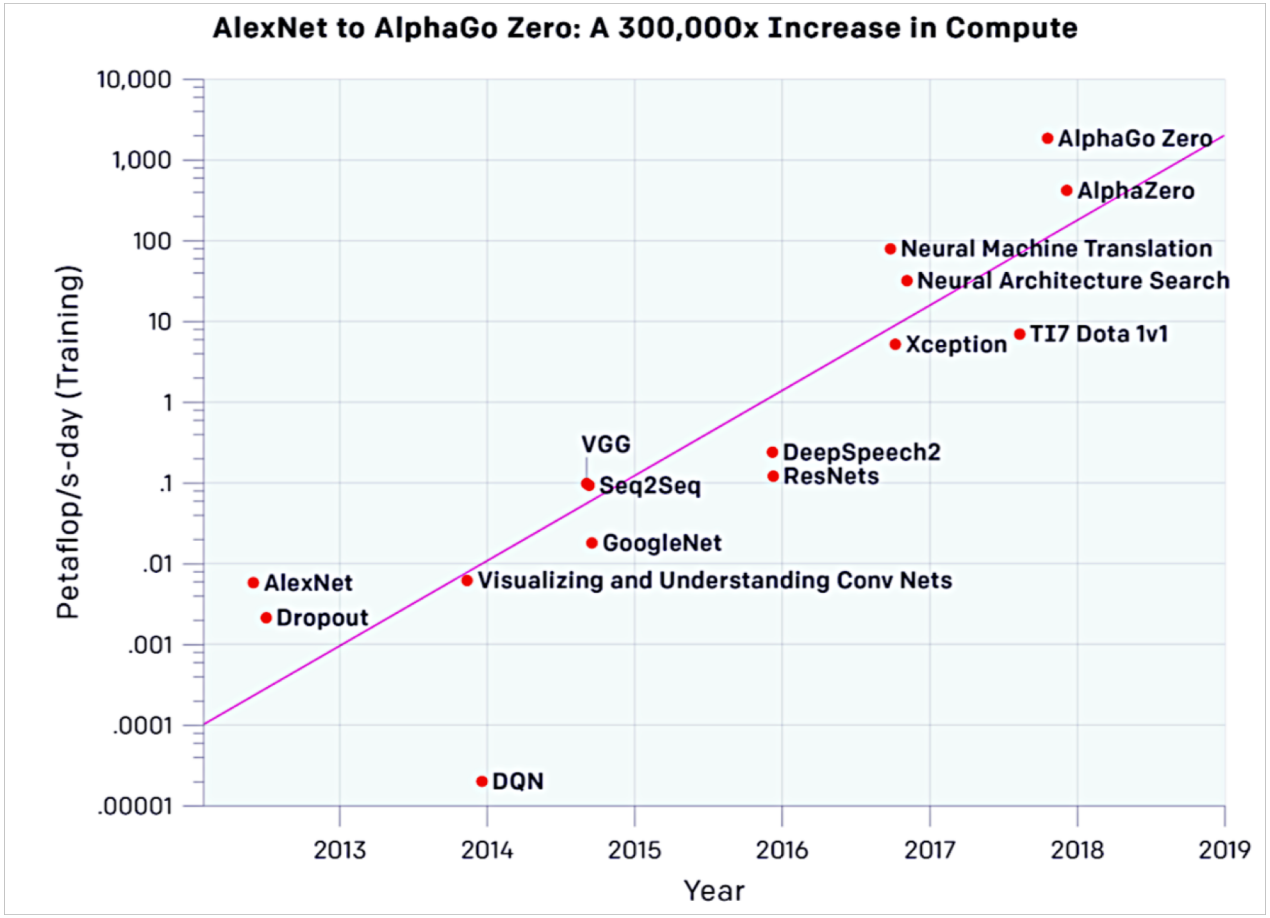
| | |
|---|---|
| GPUs | **16X NVIDIA® Tesla V100** |
| GPU Memory | **512GB total** |
| Performance | **2 petaFLOPS** |
| NVIDIA CUDA® Cores | **81920** |
| NVIDIA Tensor Cores | **10240** |
| NVSwitches | **12** |
| Maximum Power Usage | **10 kW** |
| CPU | **Dual Intel Xeon Platinum 8168, 2.7 GHz, 24-cores** |
| System Memory | **1.5TB** |
| Network | **8X 100Gb/sec Infiniband/100GigE Dual 10/25Gb/sec Ethernet** |
| Storage | **OS: 2X 960GB NVME SSDs Internal Storage: 30TB (8X 3.84TB) NVME SSDs** |
| Software | **Ubuntu Linux OS See Software stack for details** |
| System Weight | **340 lbs (154.2 kgs)** |
| System Dimensions | **Height: 17.3 in (440.0 mm) Width: 19.0 in (482.3 mm) Length: 31.3 in (795.4 mm) - No Front Bezel 32.8 in (834.0 mm) - With Front Bezel** |
| Operating Temperature Range | **5°C to 35°C (41°F to 95°F)** |

# Moore's law for Machine Learning systems

Moore's law with

$T_{x2}$ = 6 months

(instead of 18 months)



# Other paths

- In general, new algorithms will be more and more required:
  - Management of thousands of detectors
  - New techniques of component separation
  - Modelling of the Milky Way
  - Adapt to new technologies
- Avoid transmission of large amount of data
  - “move the code and not the data”

- Docker based codes (www.docker.com/)
  - Allows to distribute the code on many machines
  - Operates on virtual machines: hardware independent
  - Largely adopted in Euclid

- FPGA based systems
  - Very low cost for purchase and power
  - Require complete algorithm redesign

# Critical points / discussion

- **Prepare to cope with order 10^5 detectors**
  - Setup and optimization
  - Characterization (gain, angular, spectral, time response, polarization, …)
  - Systematics propagation
  - Data Storage and transmission
  - Combination of huge data an=mount
- **Atmospheric fluctuations (and ground pickup):**
  - Simulations for instrument design
  - Control and removal in data analysis
- **Infrastructure for European instrumentation**
  - We have “spread” resources in Italy and Europe
  - Experiments should have a well defined allocation of resources, Included in the instrument costs
  - This was proven difficult with LSPE-INFN funding
- **New algorithms development is a key**
  - Including GPU based, FPGA based, machine learning based techniques

# Thank you

# Resources NERSC

- NERSC has been the main provider of computational power for CMB experiments in the last decades

- CORI Haswell

2.81 PFlops

| | | |
|---|---|---|
| Haswell Cabinets | 14 | Each cabinet has 3 chassis; each chassis has 16 compute blades, each compute blade has 4 dual socket nodes |
| Haswell Compute nodes | 2,388 | Each node has two sockets, each socket is populated with a 16-core Intel® Xeon™ Processor E5-2698 v3 ("Haswell") at 2.3 GHz |
| | | 32 cores per node |
| | | Each core supports 2 hyper-threads, and has 2 256-bit-wide vector units |
| | | 36.8 Gflops/core; 1.2 TFlops/node; 2.81 PFlops total (theoretical peak) |
| | | Each node has 128 GB DDR4 2133 MHz memory (four 16 GB DIMMs per socket); 298.5 TB total aggregate memory. |
| | | Each core has its own L1 and L2 caches, with 64 KB (32 KB instruction cache, 32 KB data) and 256 KB, respectively; there is also a 40-MB shared L3 cache per socket |