Il modello di calcolo per l'analisi dati della missione Euclid

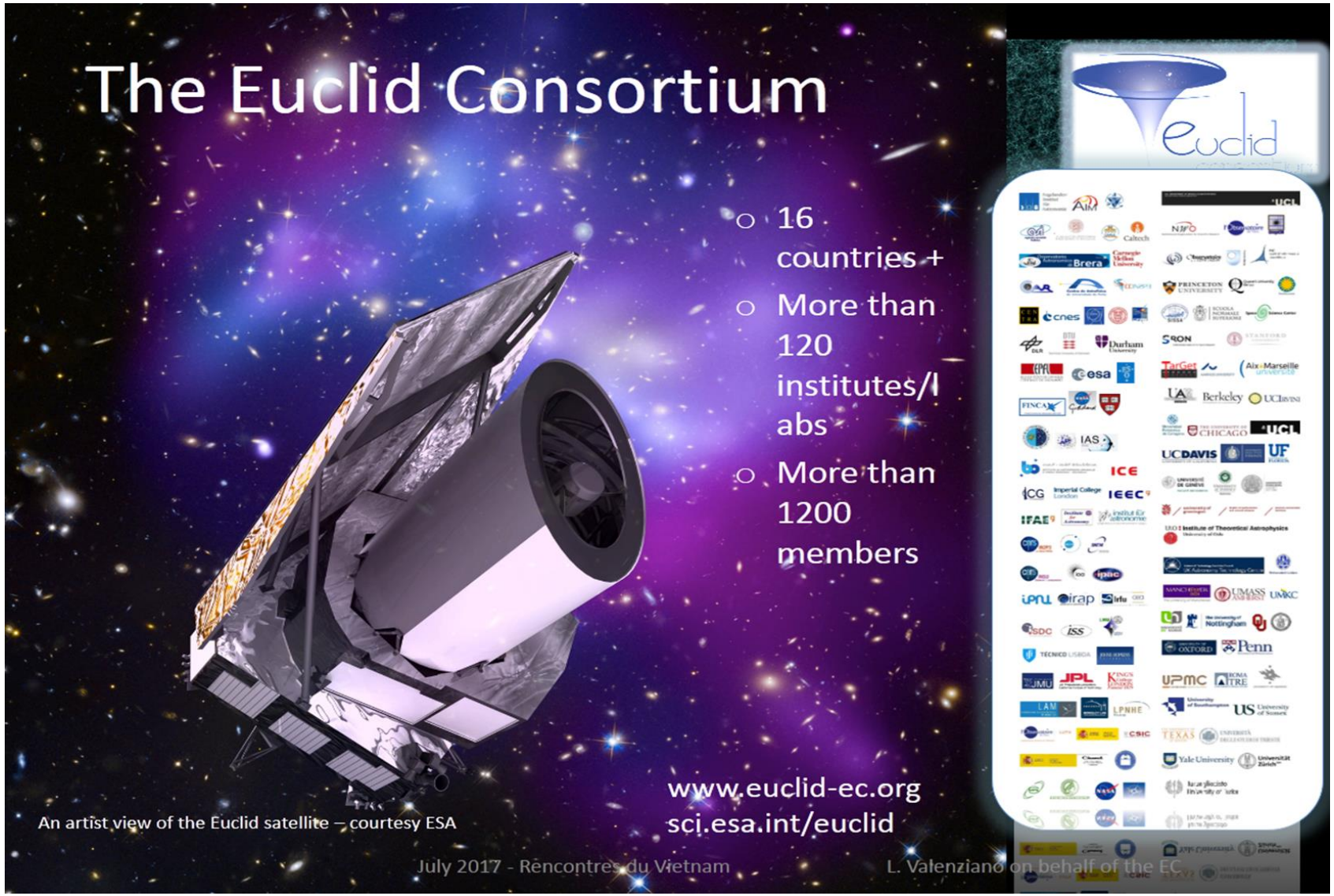
Alessandro Renzi (Univ. and INFN Padova) per Euclid-INFN

Workshop sul Calcolo nell'I.N.F.N.
Paestum, 23 - 27 maggio 2022

Outline

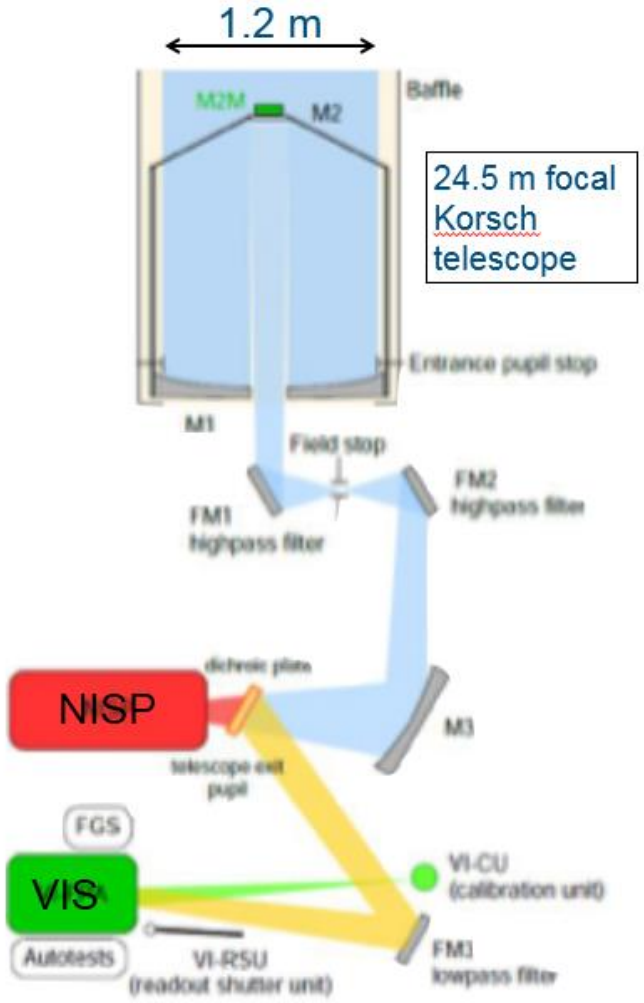
- La missione Euclid
- Il calcolo nell’esperimento Euclid
- Euclid-INFN Computing Model
- Prospettive future per il calcolo Euclid-INFN

La missione Euclid



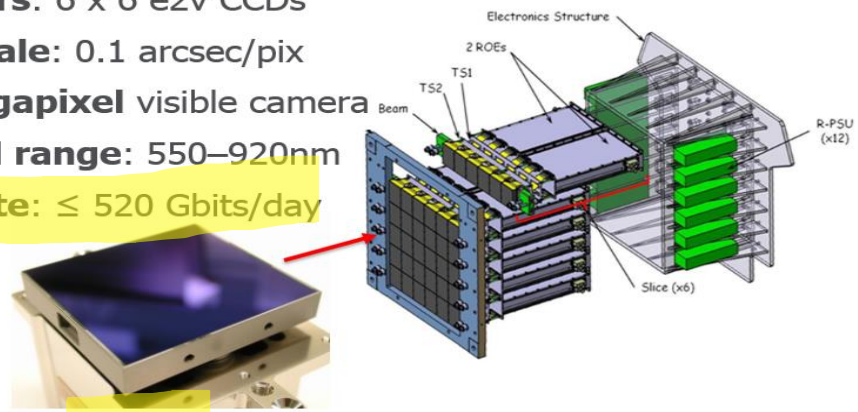
La missione si propone di mappare le strutture a grande scala dell'Universo (LSS in inglese) per studiare la fisica fondamentale dalla cosmologia, in particolare la Dark Energy (che è legata all'espansione accelerata dell'Universo) e alcune proprietà dei neutrini (somma delle masse, numero di specie, gerarchia).

La missione Euclid

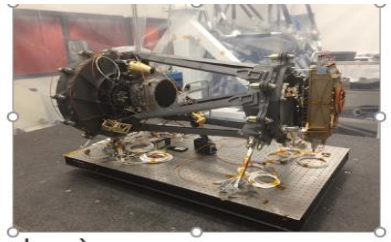


- **FoV**: 0.787 x 0.709 = 0.557 deg^2
- **Active area**: 877 cm^2
- **Detectors**: 6 x 6 e2v CCDs
- **Plate scale**: 0.1 arcsec/pix
- **609 megapixel** visible camera
- **Spectral range**: 550–920nm
- **Data-rate**: ≤ 520 Gbits/day

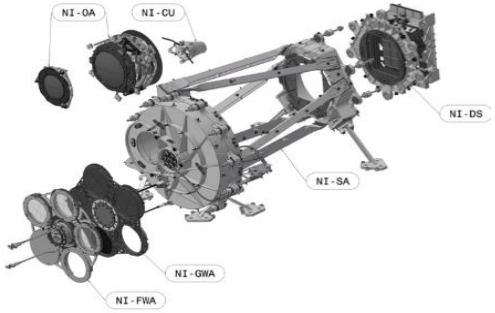
(M. Cropper & VIS Team)



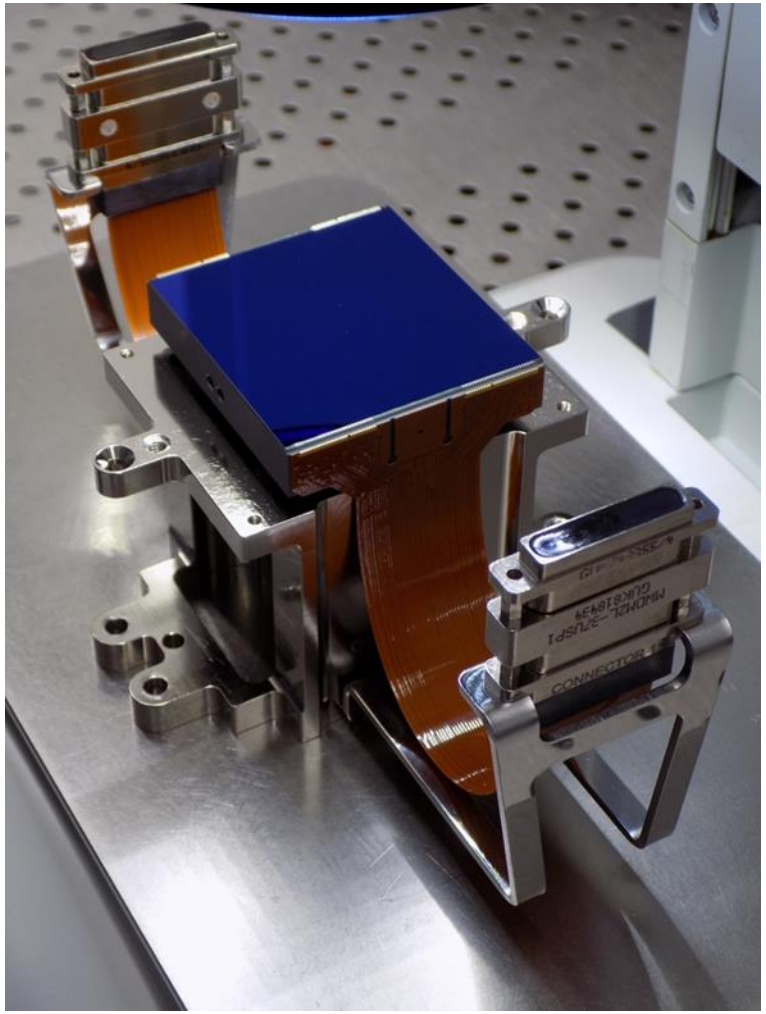
(T. Maciaszek, A. Ealet & NISP Team)



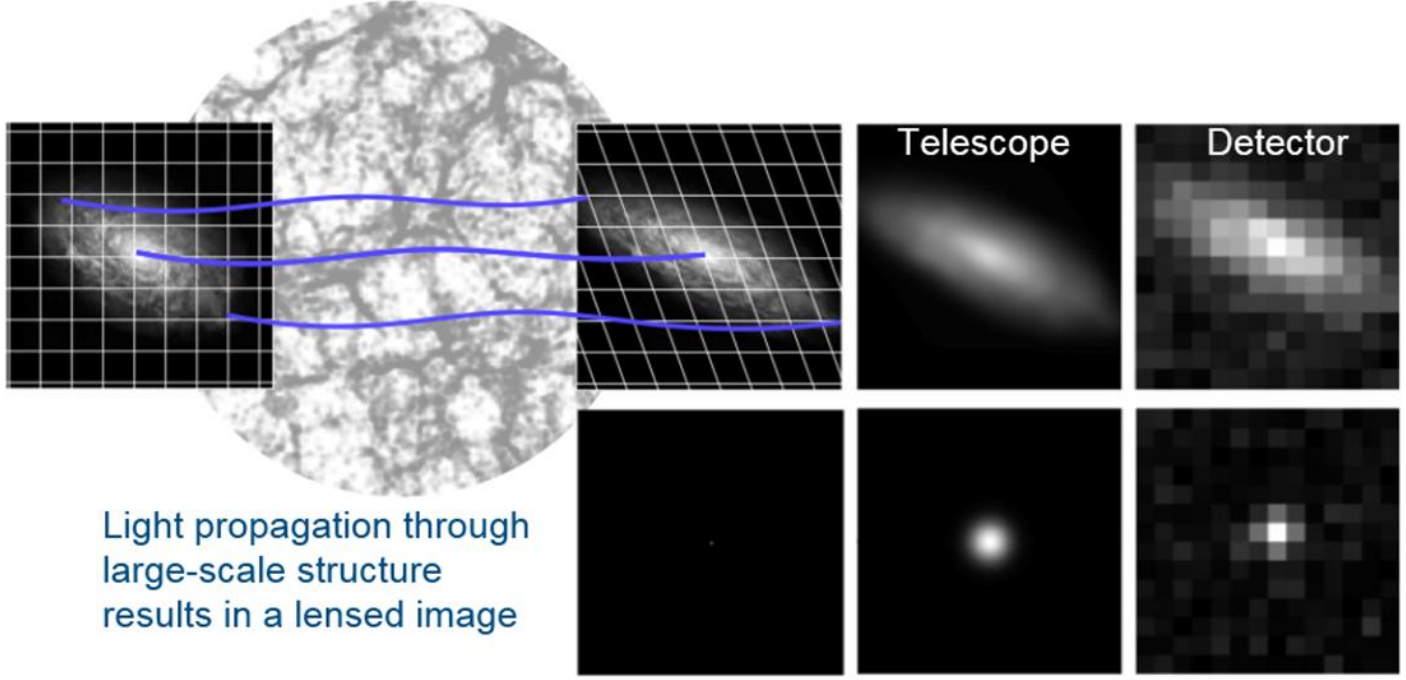
- **FoV**: 0.55 deg^2
- **Mass**: 158 Kg
- **Telemetry**: < 290Gb/day
- **Size**: 1m x 0.5m x 0.5m
- **IR detectors**: 16x 2Kx2K H2RG (Teledyne)
- **Plate scale**: 0.3 arcsec/pix
- **3 Filters:**
  Y (950-1192nm)
  J (1192-1544nm)
  H (1544-2000nm)
- **4 Grisms**:
  1B (920-1300), 1 orient. 0^0
  3R (1250-1850), 3 orient. 0^0,90^0,180^0

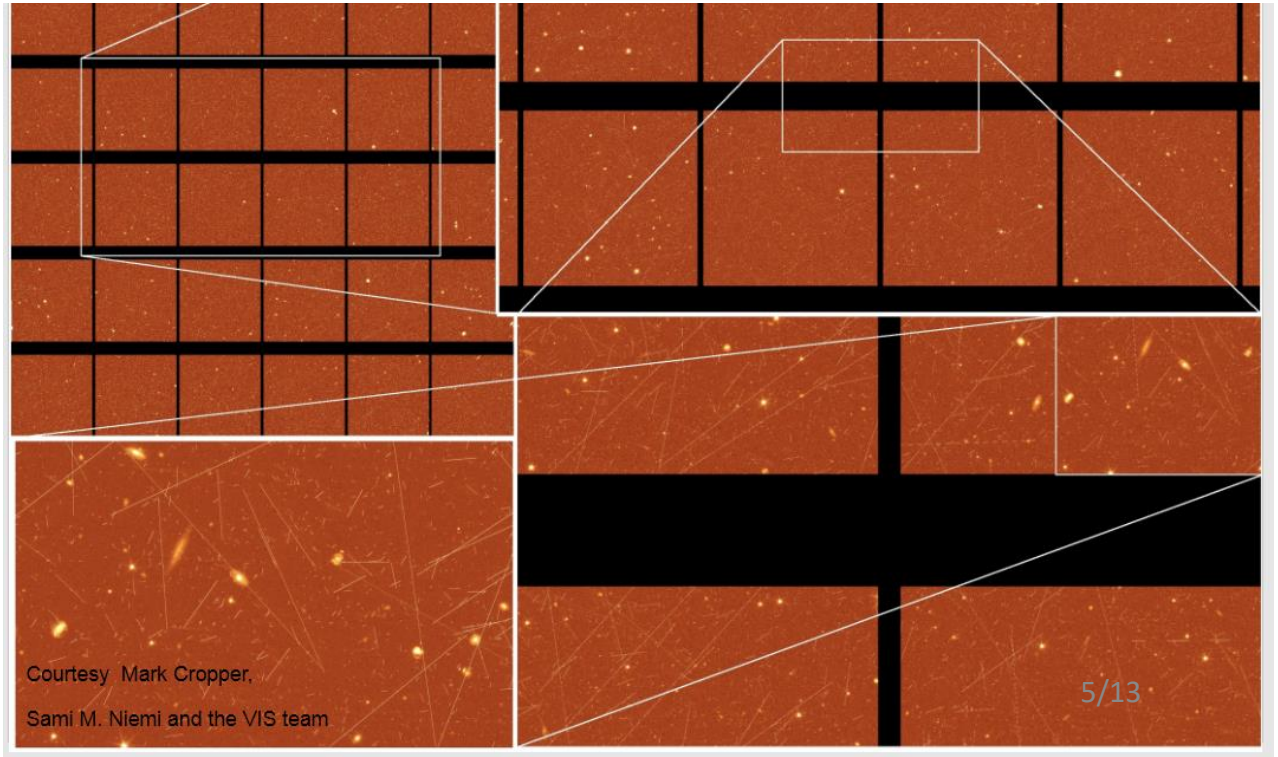


La missione Euclid

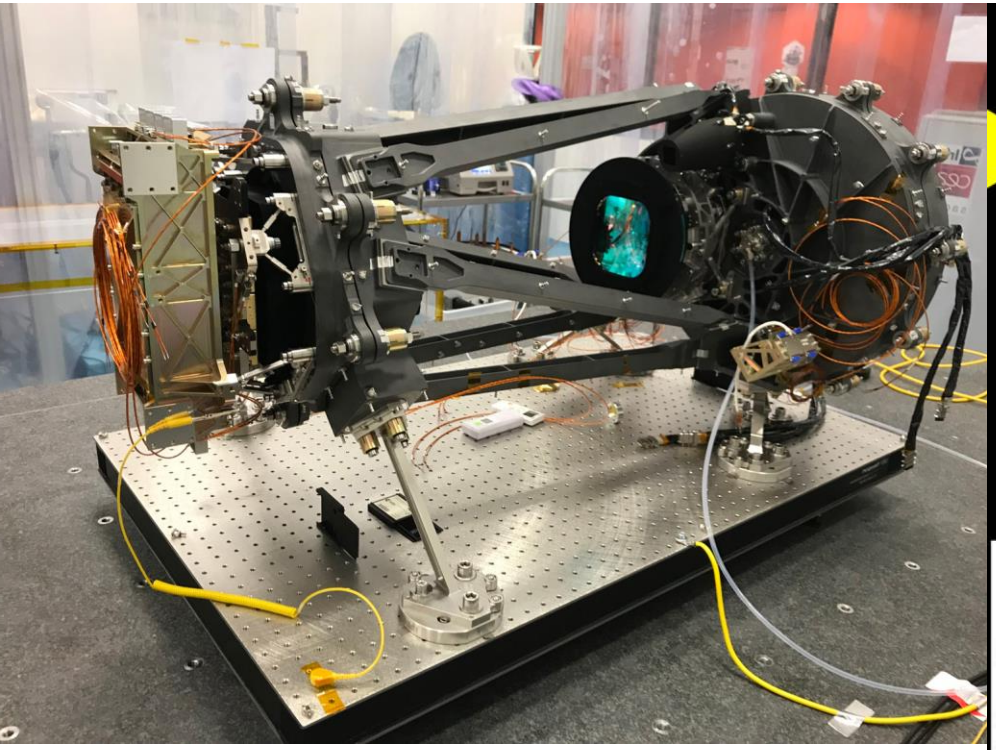


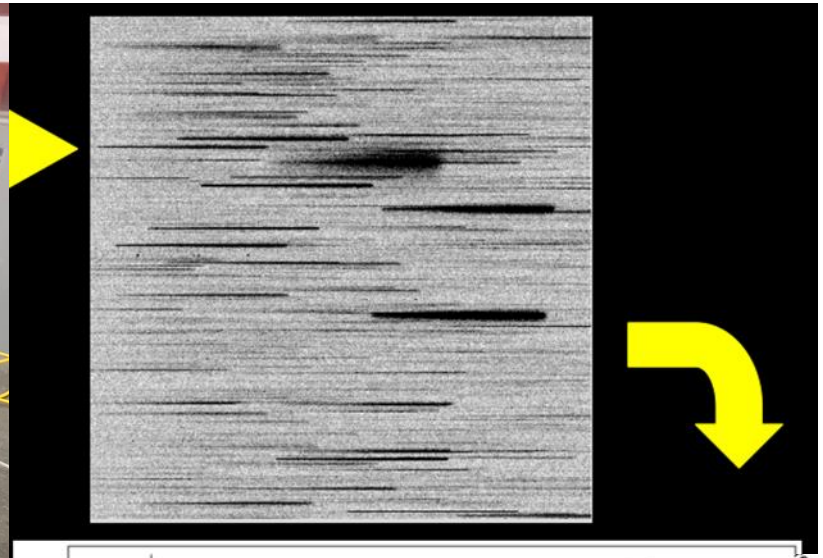
VIS (Visible Instrument): raccoglie immagini di galassie nello spettro visibile.

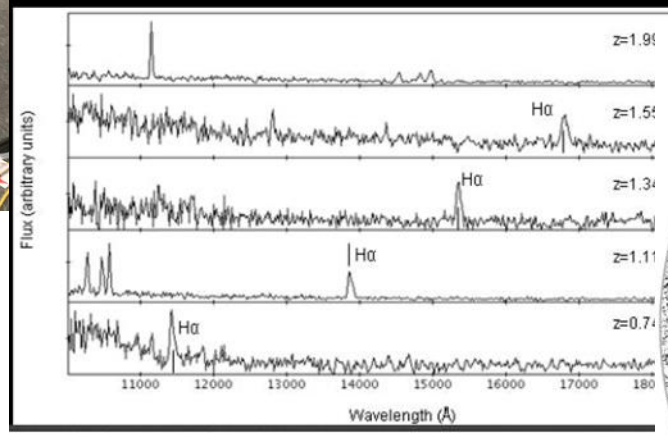


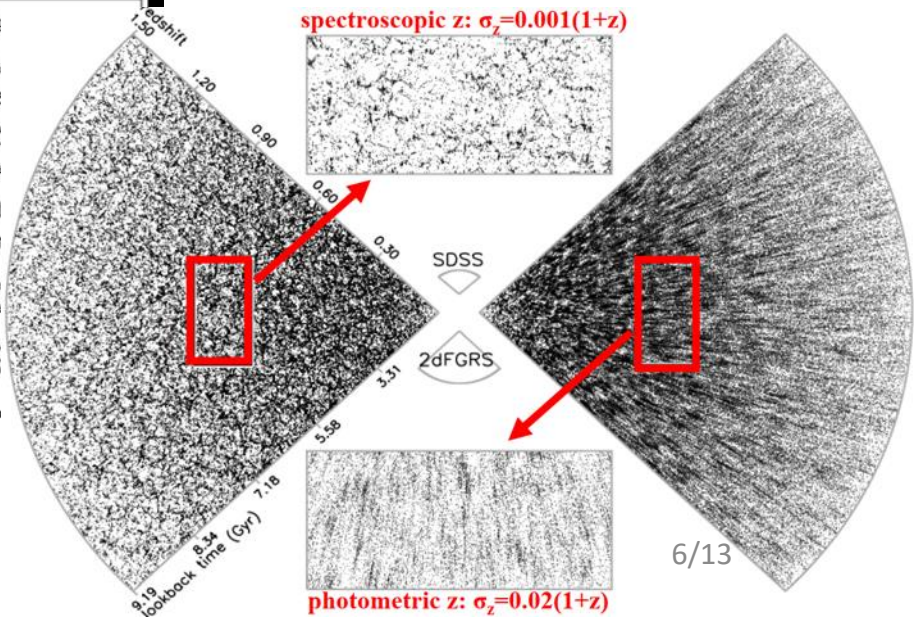


La missione Euclid



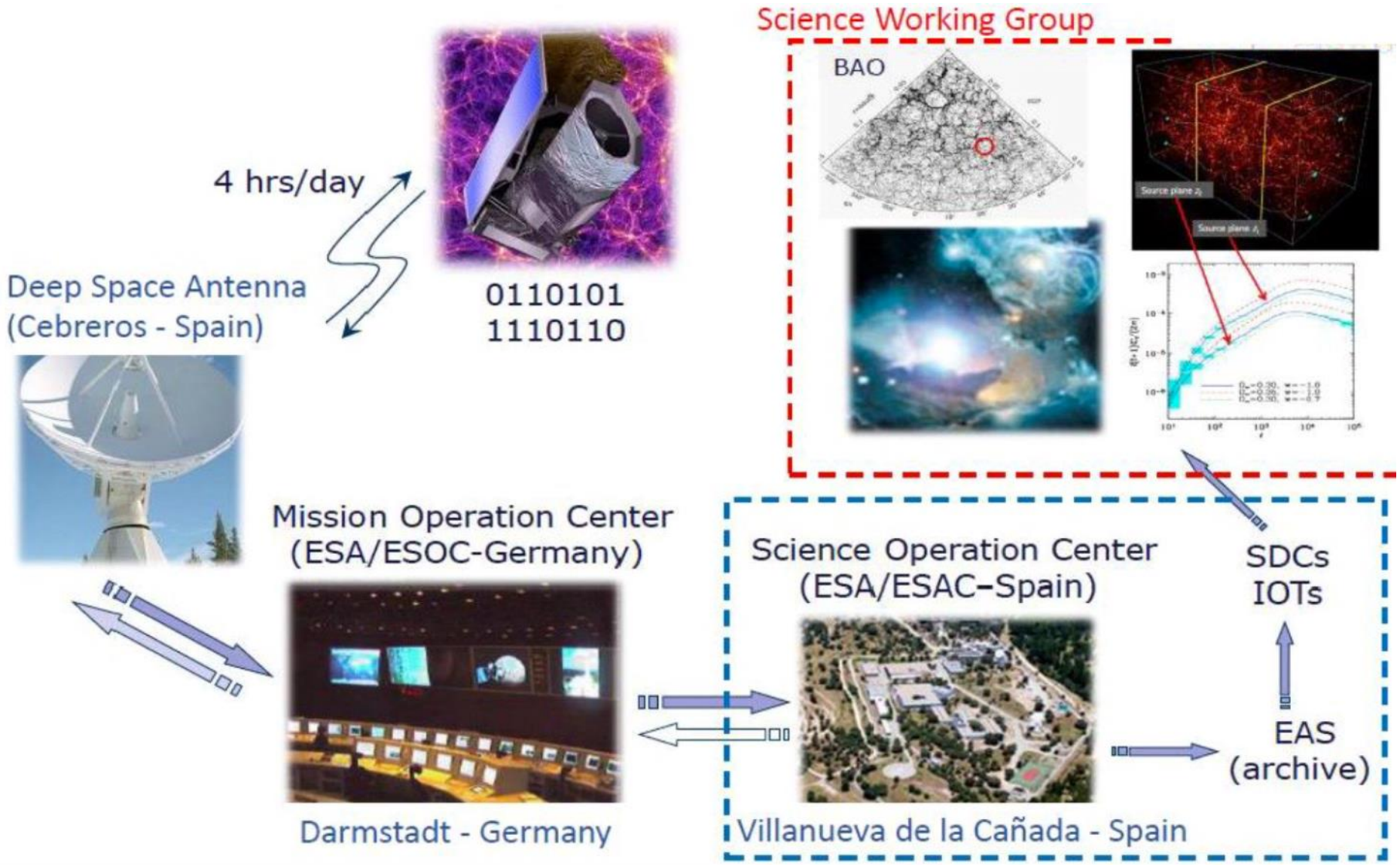






NISP (Near Infrared Spectrometer and Photometer): raccoglie spettri e fotometria di galassie nell'infrarosso [redshifts]

Il calcolo nell'esperimento Euclid



Il calcolo nell'esperimento Euclid

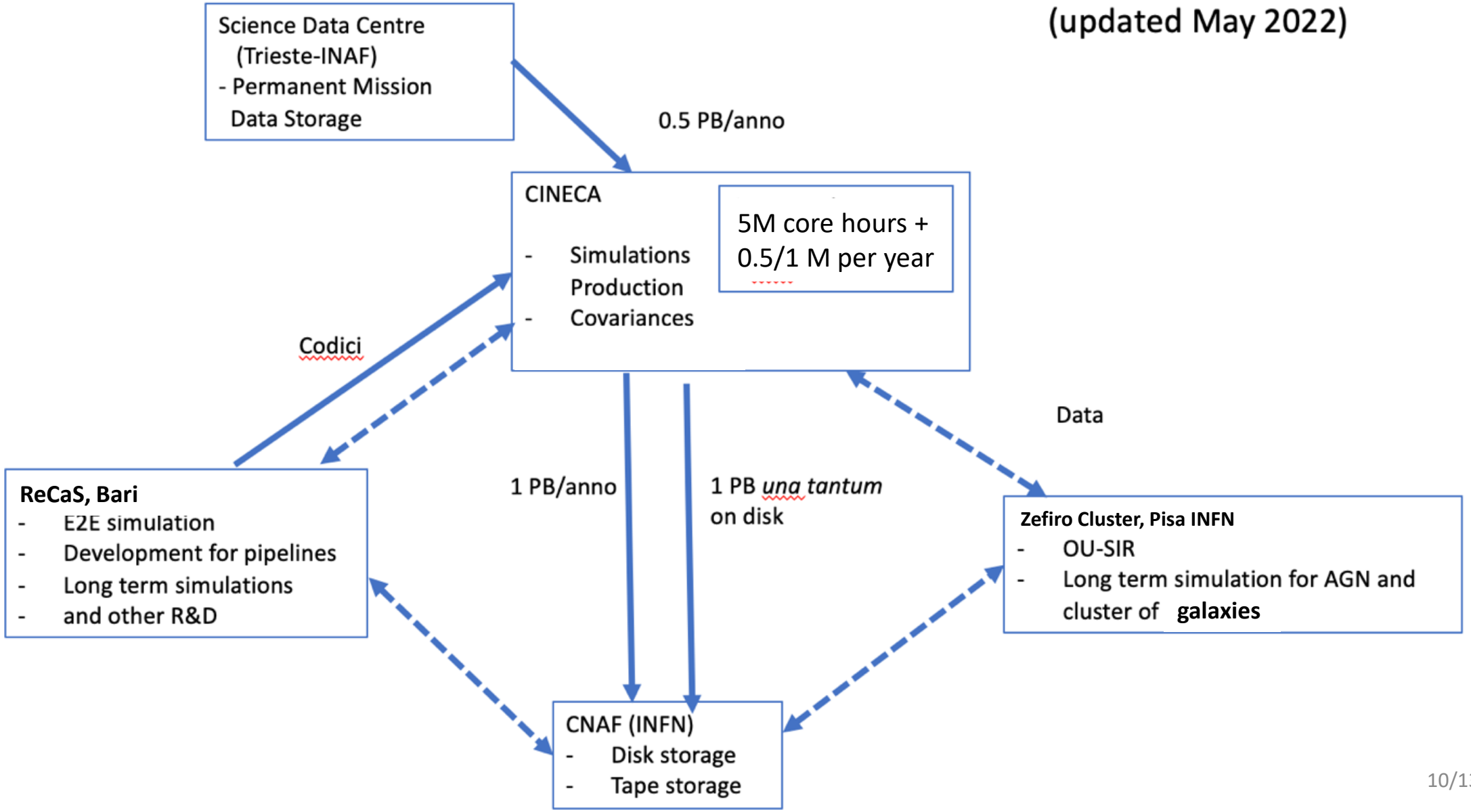
- Science Ground Segment (SGS) ha la responsabilità di:
  - archiviazione e diffusione pubblica dei dati raccolti da Euclid
  - sviluppo e run della pipeline per l'estrazione dei prodotti per la scienza
- Risorse di calcolo per l'attività SGS:
  - Enorme spazio disco per achiviazione dei dati di tutta la missione ~ 100PB
  - Modeste richieste per numero e performance CPU
  - Alta richiesta di risorse per la memoria RAM: 8GB/core
  - Al momento no GPU anche se si è iniziato a discutere di un loro possibile utilizzo

Il calcolo nell'esperimento Euclid

- Lo sfruttamento scientifico dei prodotti Euclid è demandata ai singoli Science Working Group (SWG) che, al momento, devono sfruttare risorse di calcolo (proprie) esterne a SGS per queste.
- Come gruppo INFN in Euclid ci siamo organizzati per l'analisi scientifica di una parte dei dati di Euclid, in particolare dai dati che derivano dalla spettroscopia, dove abbiamo già ricercatori INFN che hanno partecipato allo sviluppo hardware per la spettroscopia (in NISP).
- I dati di Euclid sono immagini, e il loro studio, dal punto di vista della cosmologia, è sostanzialmente un'analisi delle correlazioni in queste immagini confrontato ai modelli cosmologici simulati. Queste considerazioni vincolano fortemente l'hardware necessario per l'analisi di questi dati:
  - Simulazioni N-body (i.e. particelle di materia su griglia) – puro HPC - solo CINECA
  - Analisi con statistiche riassuntive (correlation function / spectrum) – alta richiesta di memoria RAM – HPC, ma per molte analisi i c.d. «fat node» possono essere sufficienti
  - Di solito i codici sono di tipo ibrido (MPI+openmp) - la memoria disponibile per nodo è il principale collo di bottiglia di questi codici
  - Pochi codici di riferimento nella comunità cosmologica, per la quasi totalità le analisi si svolgono con codici «homebrew», questo significa spesso scarsa ottimizzazione, inoltre scarso utilizzo di tecnologia recente (e.g. GPU) in quanto la comunità non ha forti aggregazioni nel calcolo (i.e. pochi ricercatori per software e spesso poco esperti in programmazione, no ingegneri software)

Euclid – INFN Computing Model

(updated May 2022)



Euclid-INFN Computing Model

- Anche nel caso di Euclid-INFN molti codici per le analisi scientifiche sono in fase di sviluppo
- Abbiamo scelto di sviluppare questi codici in «containers», con i seguenti vantaggi:
  - Anche il consorzio Euclid usa container per la pipeline dell'esperimento
  - Ci permette di sviluppare codice in soluzioni cloud con molta semplicità
  - Anche CINECA permette il run di container (singularity)
- Un cluster di dimensioni medie/piccole ci permette di fare il debug del codice e una prima fase di ottimizzazione (data handling optimization and profiling)
  - Per alcune analisi speciali, o per step intermedi di analisi più grandi, può anche essere direttamente utile per la scienza
- CINECA per le analisi che richiedono più risorse
- L'idea è di trasferire il codice il più possibile minimizzando il trasferimento di dati da un sistema computazionale ad un altro (anche in questo caso i container semplificano questo passaggio)
- Dove possibile l'utilizzo di tape serve come storage a lungo termine dei risultati delle analisi o delle simulazioni prodotte

Prospettive future per il calcolo Euclid-INFN

- Abbiamo bisogno di poter contare su un sistema HPC stabile per i prossimi 8/10 anni su cui sviluppare ed ottimizzare i nostri codici
  - Oltre ad essere l'unico modo per sviluppare decentemente codici, un sistema stabile e sufficientemente potente permette di costruire nuove collaborazioni aumentando il ritorno scientifico dell'analisi dati di un esperimento
- Con Leonardo il CINECA passerà ad un sistema pesantemente basato su GPU
  - Proprio perché molti codici in Euclid-INFN sono ancora in fase di sviluppo stiamo valutando la possibilità di utilizzare GPU per quanto possibile
  - Lo sviluppo e l'ottimizzazione in GPU è molto più difficile, l'utilizzo di nuovi linguaggi è una possibile strada per semplificare lo sviluppo? (es. Julia language)
- Sviluppo di algoritmi più efficienti per hardware disponibile. Nonostante si sia passati a sistemi manycore da molti anni, molte librerie e algoritmi sono ancora ottimizzati per sistemi a pochi core molto veloci. Esplorare nuovi algoritmi può essere un buon modo per ottimizzare le analisi su hardware moderno (es. Wavelets su GPU)
- Il calcolo scientifico dovrebbe essere un settore strategico per la ricerca. Come Euclid-INFN cerchiamo di trovare sinergie con INAF e ASI per aumentare le potenzialità di calcolo HPC in Euclid

Grazie per l’attenzione