CINECA_A2 e LHC

(Nomi nell’ultima slide!)

MARCONI

Model: Lenovo NeXtScale
Architecture: Intel OmniPath Cluster
Internal Network: Intel OmniPath
Disk Space: 17PB (raw) of local storage
Total Peak Performance: ~ 20 PFlop/s
UserGuide

MARCONI - A2 (KNL)

Nodes: 3.600
Processors: 1 x 68-cores Intel Xeon Phi7250 (KnightLandings) at 1.4 GHz
Cores: 68 cores/node, 244.800 cores in total
RAM: 16+96 GB/node
Peak Performance: ~11 PFlop/s
Quick startup guide

MARCONI - A3

Nodes: 1.512 + 792 + 912
Processors: 2 x 24-cores Intel Xeon 8160 (SkyLake) at 2.10 GHz
Cores: 48 cores/node, 72.576 + 38.016 cores in total
RAM: 196 GB/node
Peak Performance: ~8 PFlop/s
Quick startup guide

- L'INFN ha gia' parte del Tier-1 su A1 (qui non e' neppure piu' elencata, e' una vecchia partizione Broadwell ora dismessa) - ~ 280 kHS06 - 0.5 MEur, attivo da 1 anno, altri 2 da contratto
- L'INFN ha una frazione di A2 per il calcolo teorico (1.5 MEur) - stessa tempistica

- Nessuna sperimentazione / accordi fatti invece per provare a usare A2 (Intel KNL) con LHC
  - Interessante per usi futuri, architettura x86_64 "peculiare"
  - Interessante per stabilire **come girare LHC su macchine standard CINECA (la partizione A1 e' installata ad hoc)**
  - Interessante per capire i limiti del CINECA su utenti / accessi / rete / installazioni particolari

Idea di applicare a un Grant PRACE specifico per LHC/INFN su HPC

PRACE 19th Call for Proposals for Project Access

- **Opening date**: 05 March 2019
- **Closing date**: 30 April 2019, 10:00 CEST
- **Applicants' reply to scientific reviews**: Mid-July 2019
- **Submission of Progress Reports** (Multi-year Projects) / **Final Reports** (Single-year Projects): 05 June 2019 @ 10:00 CEST
- **Communication of allocation decision**: End of September 2019
- **Allocation period for awarded proposals**: 01 October 2019 – 30 September 2020
- **Type of Access**: Single-year Project Access and Multi-year Project Access

Industry Access Pilot: Call 19 offers Principal Investigators from industry the possibility to apply for Single-year access to a special Industry Track which prioritises 10% of the total resources available (see Eligibility Criteria on page 11).

Pilot Phase from the European ICEI project (Fenix Research Infrastructure)
This call includes the opportunity to benefit from resources by the European ICEI project, contributed as a Pilot Phase.

System	Architecture	Site (Country)	Core Hours (node hours)	Minimum request (core hours)
HAWK(*)	HPE System	GCS@HLRS (DE)	460 million	35 million
Joliot Curie – AMD(*)	BULL Sequana X1000/XH2000	GENCI@CEA (FR)	343 million (2.7 million)	15 million
Joliot Curie – SKL	BULL Sequana X1000	GENCI@CEA (FR)	132 million (2.7 million)	15 million
Joliot Curie – KNL	BULL Sequana X1000	GENCI@CEA (FR)	94 million (1.5 million)	15 million
JUWELS	BULL Sequana X1000	GCS@JSC (DE)	70 million (1.5 million)	35 million
Marconi-Broadwell(*)	Lenovo System	CINECA (IT)	36 million (1 million)	15 million
Marconi-KNL(*)	Lenovo System	CINECA (IT)	408 million (6 million)	30 million
MareNostrum	Lenovo System	BSC (ES)	240 million (5 million)	30 million
Piz Daint	Cray XC50 System	ETH Zurich / CSCS (CH)	544 million (8 million)	68 million *Use of GPUs*

- Abbiamo partecipato ad una Call simile a questa, scaduta a ottobre 2018
- Le call PRACE sono atipiche per noi: devi aver dimostrato di saper fare le cose che vuoi fare, su sistemi analoghi; no R&D
- Noi: da maggio 2018, ottenute 30000 ore su KNL per fare i test

⟶ **Dear Dr. Tommaso Boccali, we would like to thank you for submitting a proposal to the 18th Call for PRACE Project Access…. We are happy to inform you that your proposal is amongst those being awarded resources. The amount of resources allocated to your proposal (number 2018194658) is 30.000.000 core hours on Marconi – KNL, for a period of 12 months.**

Cosa pensiamo di farci?

- Full proposal qui
- Presi **3 use cases di fisica** LHCb, ATLAS e CMS (ALICE → troppa poca RAM su queste macchine…), valutato l'impatto delle ore di calcolo sull'analisi finale (errore sistematico)
- Nella pratica, soprattutto Monte Carlo simulation, ma l'idea e' di testare anche workflow completi (da Generatori a analysis format)

Valutazione tecnica ok, valutazione di fisica 18/18 (grazie al lavoro e all'inventiva degli esperimenti…)

Ma come si procede tecnicamente?

Caratteristiche di un nodo A2

- KNL a 68(*4) cores(Threads) → fino a 272 x86_64 threads
  - Notare che il grant costa a “core fisici”, quindi 1 ora = 68 coreh (e non 272)
- Ogni thread ~ ¼ come potenza di un thread HTon Xeon
- 96 GB di RAM (< 0.5 GB/thread -- standard LHC e’ > 2 GB/thread)
- No external network connectivity
- No CVMFS, no virtualizzazione
- Politica utenti chiusa (una persona fisica ←→ un account)
- Linux based, release “vicina” a openSuse?
- Nodi accessibili solamente via Slurm, da alcuni login nodes
- Nessun disco locale utente, tutto GPFS mounted (Omnipath)

Cosi’ non va ….. Costanti meetings per cercare di trovare soluzioni / compromessi. Non solo per ora, ma anche per i sistemi futuri!

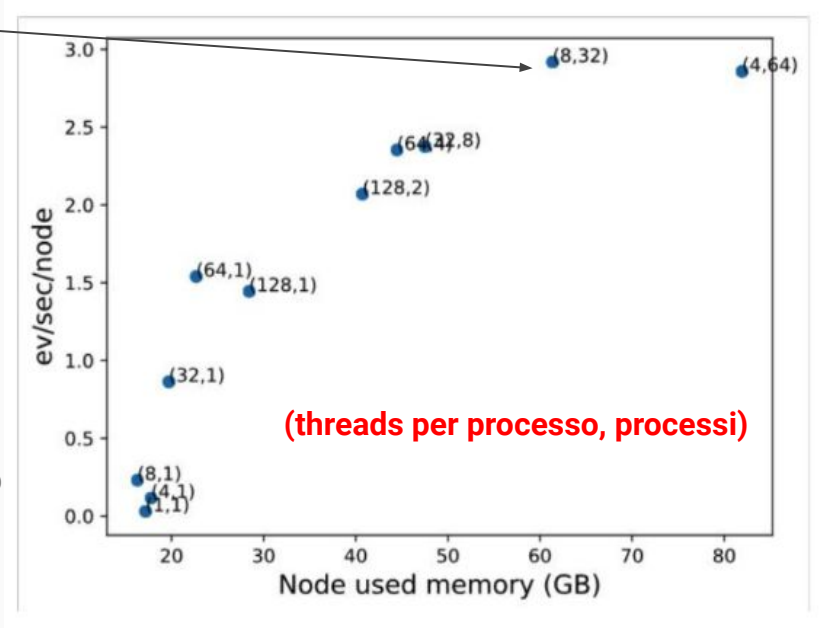
Cambiamenti realizzati / decisi / realizzati

- **Ok** per pool accounts per esperimento / grant
- **Ok** per CVMFS
  - Con squid CINECA
- **Ok** per singularity
- **Ok** per apertura rete esterna, **ma solo verso CERN, CNAF e FNAL** (solo outgoing connections) - nella pratica non vogliono fare fanout ma accettano una “lista di reti”
- **Ok** per installare Condor su un “login node dedicato” (che vede WAN e Slurm) - noi forse preferivamo che stesse fisicamente a CNAF, ma pare che “vedere Slurm” voglia dire esportare filesystems etc ...

Questo ha permesso di fare i tests e dimostrare le capability come necessario per applicare al grant; notare che sono modifiche system wide, valide anche per A3...

Qualche esempio dal grant proposal

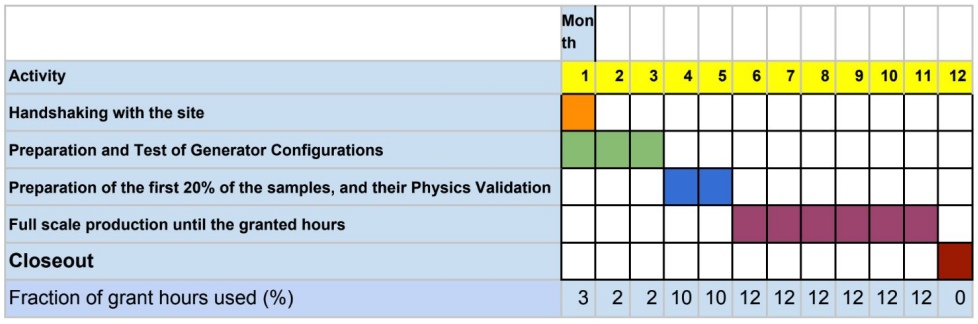
- **CMS Simulation**: di solito giriamo a 8 thread, ma cosi' si e' troppo vicini ai 96 GB di RAM (anche considerando un +50% per la generazione di eventi + pesanti)
- Pero' possiamo riconfigurare i processi, in modo da utilizzare 256 threads con un numero variabile di thread e processi
- Se si passa da (32,8) a (8,32) ~40% di guadagno in RAM, con ~15% performance inferiori
  - **Il messaggio e' che siamo tunabili, questo e' piaciuto molto a CINECA**
- Anthe ATLAS e LHCb hanno positivamente testato i loro workflow di base (functionality test)



E ora?

- **Manca la messa in produzione come estensione elastica del CNAF**
- Il lavoro era gia' cominciato l'estate scorsa con l'installazione di un CreamCE/Slurm (SDP); poi abbandonato per 2 motivi
  - Meglio provare direttamente Condor/Slurm, Cream CE e' comunque in decommissioning; inoltre Condor e' piu' "wan friendly" di Cream
  - Per il grant non serviva dimostrare questa parte, e si e' aspettata allocazione prima di perderci tempo
- Adesso: grant partito 1 Aprile, 12 mesi. Pero' da gantt chart di proposal, i primi 5 mesi erano comunque di startup
- → non siamo in ritardo mostruoso

3.1 Gantt Chart



	Month											
Activity	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Handshaking with the site												
Preparation and Test of Generator Configurations												
Preparation of the first 20% of the samples, and their Physics Validation												
Full scale production until the granted hours												
Closeout												
Fraction of grant hours used (%)	3	2	2	10	10	12	12	12	12	12	12	0

Cosa serve adesso?

- Ricreare account finali e settare fairshare relativi etc (== tante mail)
- Installazione CondorCE/Slurm: SDP sul pezzo!
  - Creata una installazione simile @ CNAF per accumulare esperienza
  - Testare config CINECA (al momento attivita' lenta perche' non abbiamo accesso root al login node, anche qui via mail → da cambiare)
- Come far vedere ai WN piu' di CNAF/CERN/FNAL:
  - Almeno per CMS (ma dovrebbe funzionare per tutti) brillantemente risolto da DS, DC, DC usando un Xrootd proxy caching server al CNAF che fa fanout. Si vede tutta la federazione, nessuna limitazione vera
- La rete non e' banale (→ SZ):
  - L'idea iniziale era configurare l'accesso al xrootd CNAF via la rete dedicata Infinera.
  - Pare non possibile per problema hardware (== non hanno la bretella che colleghi i due switch) -- per il momento si va avanti con la GPN GARR, il che puo' voler dire che i workflows + pesanti non gireranno
  - Se non altro, hanno detto che in effetti a questa connessione non avevano pensato, e che nel prossimo tender di rete (prossima macchina?) metteranno ab initio questa possibilita'
- Testare i config da girare, e mettere in produzione la soluzione lato esperimento

Chi ci ha lavorato

- Lato Sottomissione jobs, **Stefano dal Pra** e' come al solito indispensabile
- Per la rete, **Stefano Zani** come consulente del CINECA
- Per la soluzione dell'xrootd proxy cachato al CNAF, **Daniele Cesini, Daniele Spiga e Diego Ciangottini**
- Per il test di esperimento
  - **Io** per CMS (e ora **Daniele Bonacorsi** entra nel giro)
  - ATLAS **Alessandro de Salvo**
  - Per LHCb **Concezio Bozzi, Anna Lupato, Alessio Gianelle** e l'oriundo **Andrea Valassi**
- **Luca dell'Agnello e Gaetano** come supporto politico