OTTIMIZZAZIONE DI ALGORITMI PER L'ANALISI DI LOG FILES SU SISTEMI BATCH

Relatore:

LORENZO RINALDI

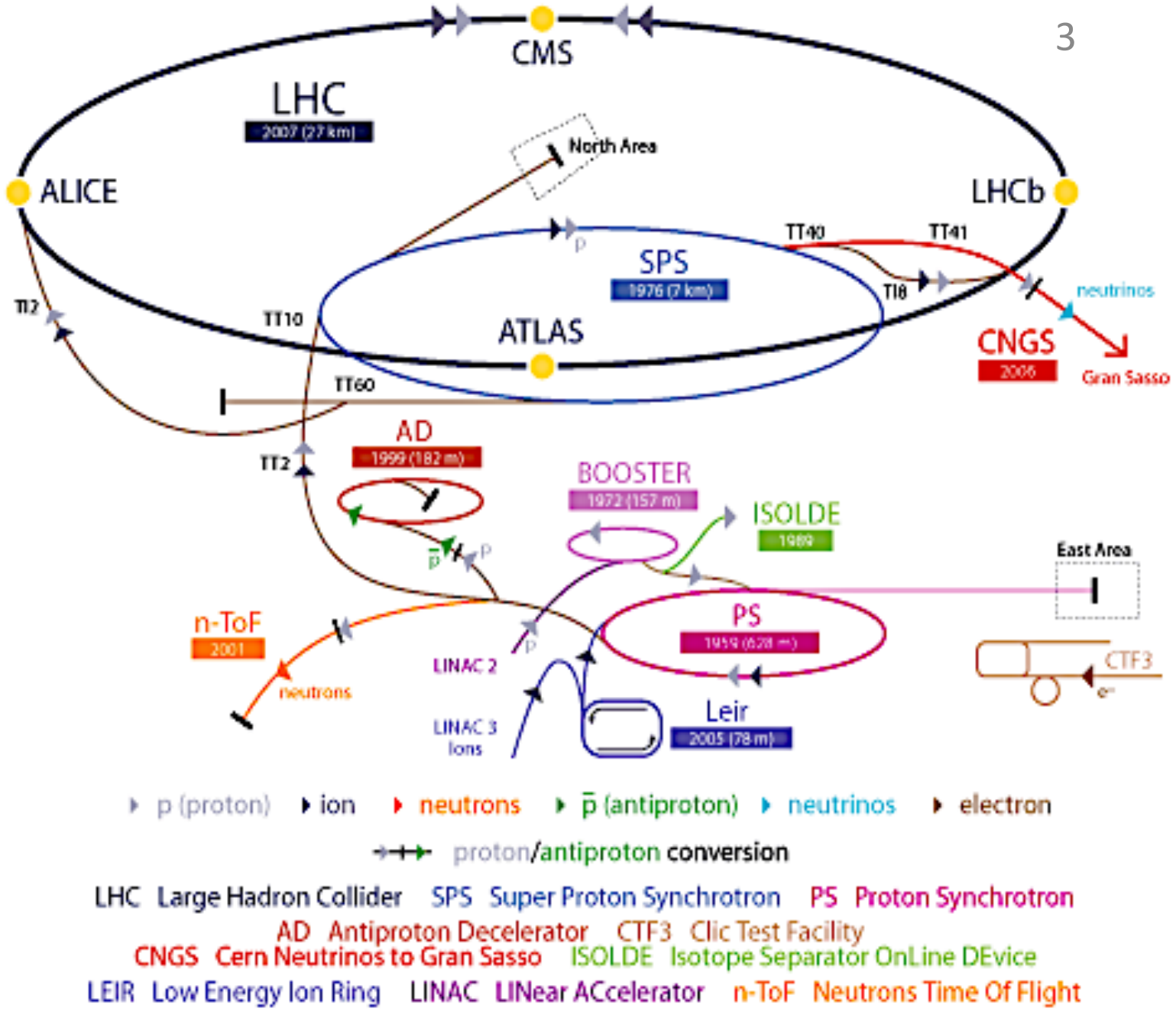
Candidato:

LUCIA GASPERINI

INDICE

CERN

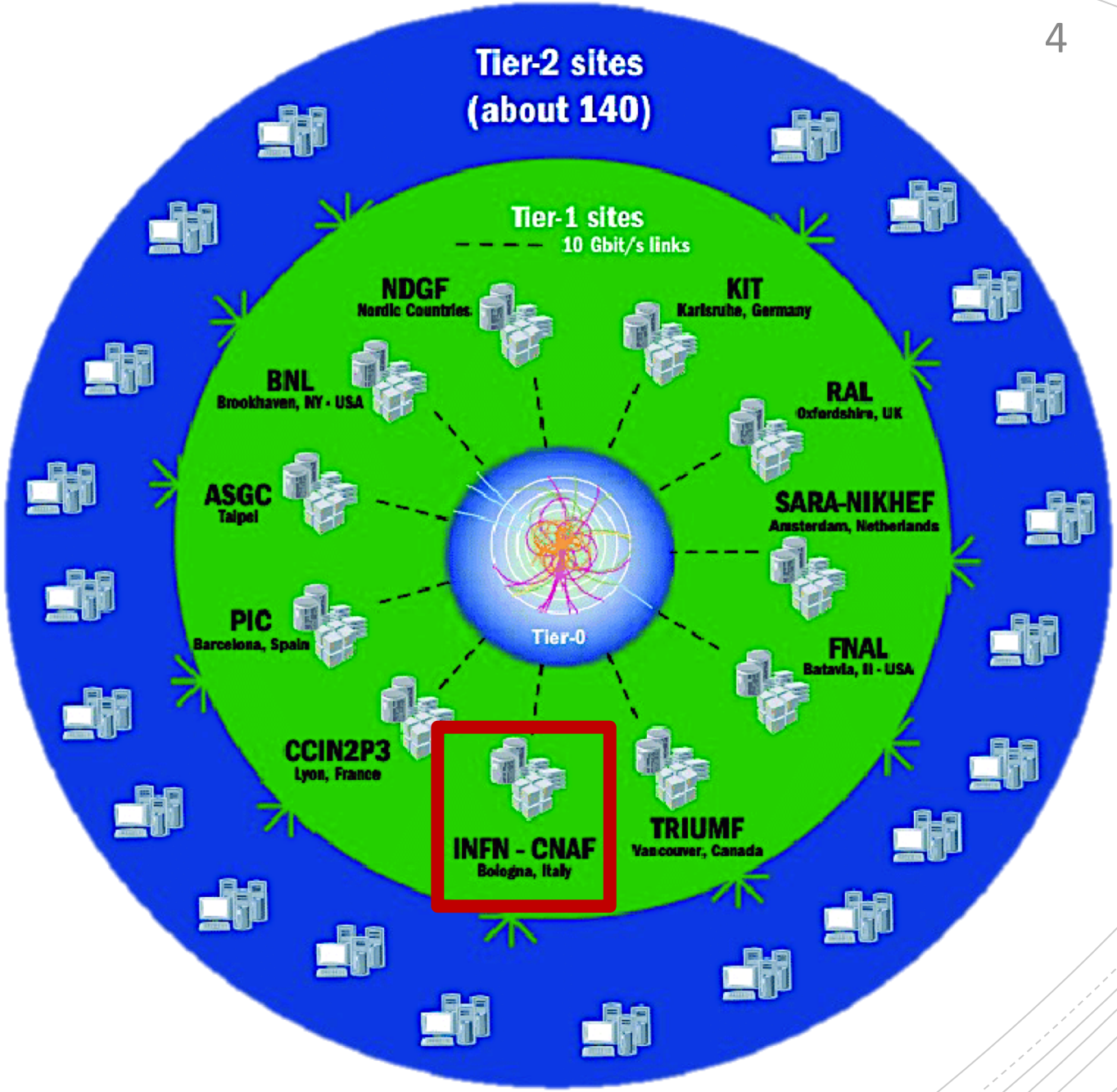
- ESPERIMENTI DI FISICA DELLE ALTE ENERGIE
- COSTRUZIONE, FUNZIONAMENTO, AGGIORNAMENTO DI **LHC**
  → COMPLESSO MACCHINE ACCELERATRICI



WLCG

WORLDWIDE LHC COMPUTING GRID

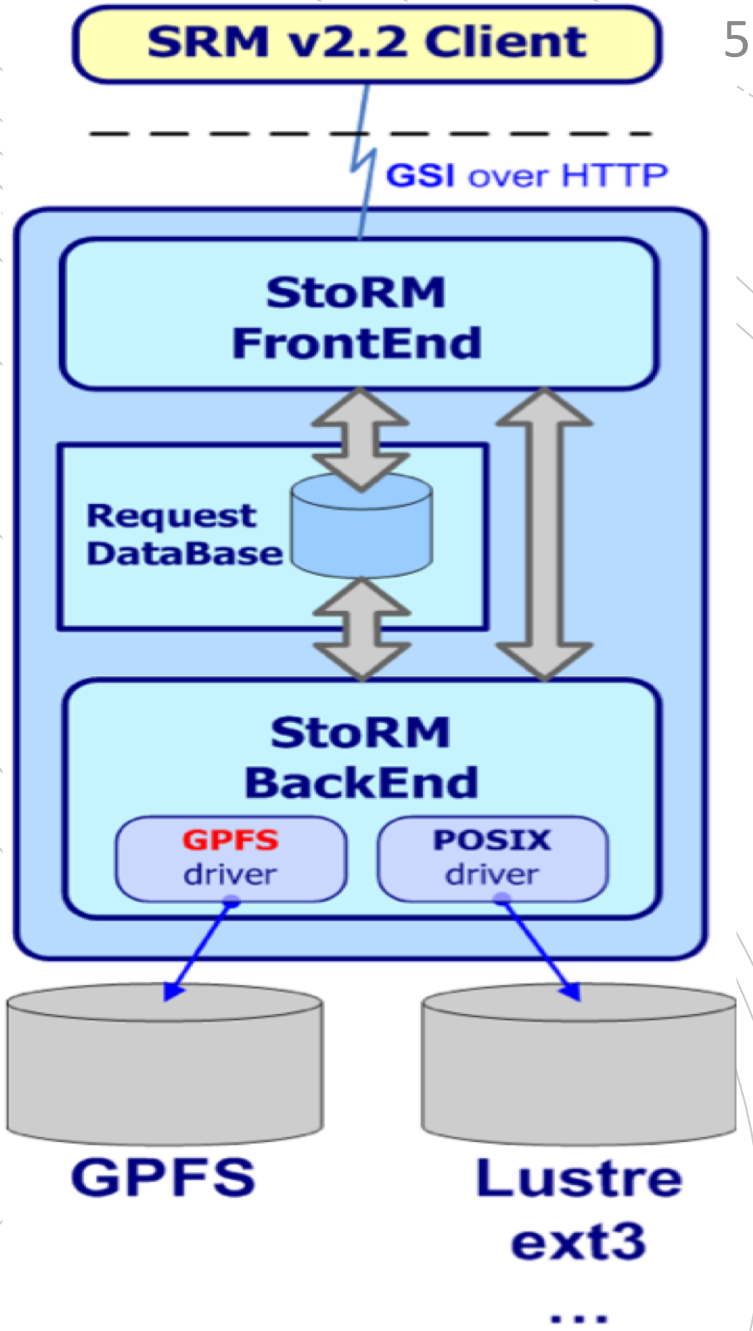
- COLLABORAZIONE GLOBALE COORDINATA DAL CERN
- RETE DI CALCOLO A **4 LIVELLI** (TIER) PER I DATI DEL LHC



StoRM

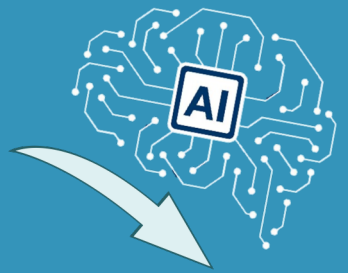
STORAGE RESOURCE MANAGER

- RISORSA DI ARCHIVIAZIONE
- SERVIZIO LEGGERO, SCALABILE, FLESSIBILE E AD ALTE PRESTAZIONI
- ALLOCAZIONE DINAMICA DELLO SPAZIO
- SCHEMA XML PER POSIZIONE FISICA
- ARCHITETTURA MULTISTRATO
- OPERAZIONI SALVATE SU LOG FILE



SCOPO DELLA TESI

LOG FILE DEL CNAF



PREVENZIONE PROBLEMI

INFN CNAF
Istituto Nazionale di Fisica Nucleare
Centro Nazionale per la Ricerca e lo Sviluppo nelle Tecnologie Informatiche e Telematiche

```
03/08 03:45:03.202 Thread 4 -  INFO [c74701c9-20ad-46db-acbb-96de84838c68]: process_request : Connection from
2001:1458:201:e3::100:6f4
03/08 03:45:03.275 Thread 4 -  INFO [c74701c9-20ad-46db-acbb-96de84838c68]: ns1__srmReleaseFiles : Request: Release files. IP:
2001:1458:201:e3::100:6f4. Client DN: /DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=ddmadmin/CN=531497/CN=Robot: ATLAS Data
Management. surl(s): srm://storm-
fe.cr.cnaf.infn.it/atlas/atlasdatadisk/rucio/data17_13TeV/99/2a/data17_13TeV.00338846.physics_Main.daq.RAW._lb0570._SFO-
6._0003.data. token: 78892824-8495-49b4-86e9-6db71bc110e9
03/08 03:45:03.287 Thread 4 -  INFO [c74701c9-20ad-46db-acbb-96de84838c68]: Result for request 'Release files' is
'SRM_SUCCESS'
03/08 03:45:03.694 Thread 35 -  INFO [68311b06-c66f-4d71-8a53-2e451c0df80b]: Request 'PTP status' from Client
IP='::ffff:131.154.194.217' Client DN='/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=atlpilo1/CN=614260/CN=Robot: ATLAS Pilot1'
# Requested token 'd3e7e258-9983-420a-bed3-c585bf90e928'
03/08 03:45:03.695 Thread 35 -  INFO [68311b06-c66f-4d71-8a53-2e451c0df80b]: Result for request 'PTP status' is 'SRM_SUCCESS'
03/08 03:45:04.406 Thread 35 -  INFO [68311b06-c66f-4d71-8a53-2e451c0df80b]: ns1__srmPutDone : Request: Put done. IP:
::ffff:131.154.194.217. Client DN: /DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=atlpilo1/CN=614260/CN=Robot: ATLAS Pilot1.
surl(s): srm://storm-
fe.cr.cnaf.infn.it/atlas/atlasscratchdisk/rucio/user/fkaya/da/5d/user.fkaya.364227.Sherpa_221_NNPDF30NNLO_Wenu_PTV1000_E_CMS.S
herpa_221.Rivet_ATLAS14_I1319490.Wenu_muChnnl.v01.log.20752208.000016.log.tgz.rucio.upload. token: d3e7e258-9983-420a-bed3-
c585bf90e928
03/08 03:45:04.420 Thread 35 -  INFO [68311b06-c66f-4d71-8a53-2e451c0df80b]: Result for request 'Put done' is 'SRM_SUCCESS'
03/08 03:45:04.423 Thread 35 -  INFO [68311b06-c66f-4d71-8a53-2e451c0df80b]: ns1__srmLs : Request: Ls. IP:
```

ARTIFICIAL INTELLIGENCE

MACHINE LEARNING

SUPERVISIONATO

Necessità di etichette

NON SUPERVISIONATO

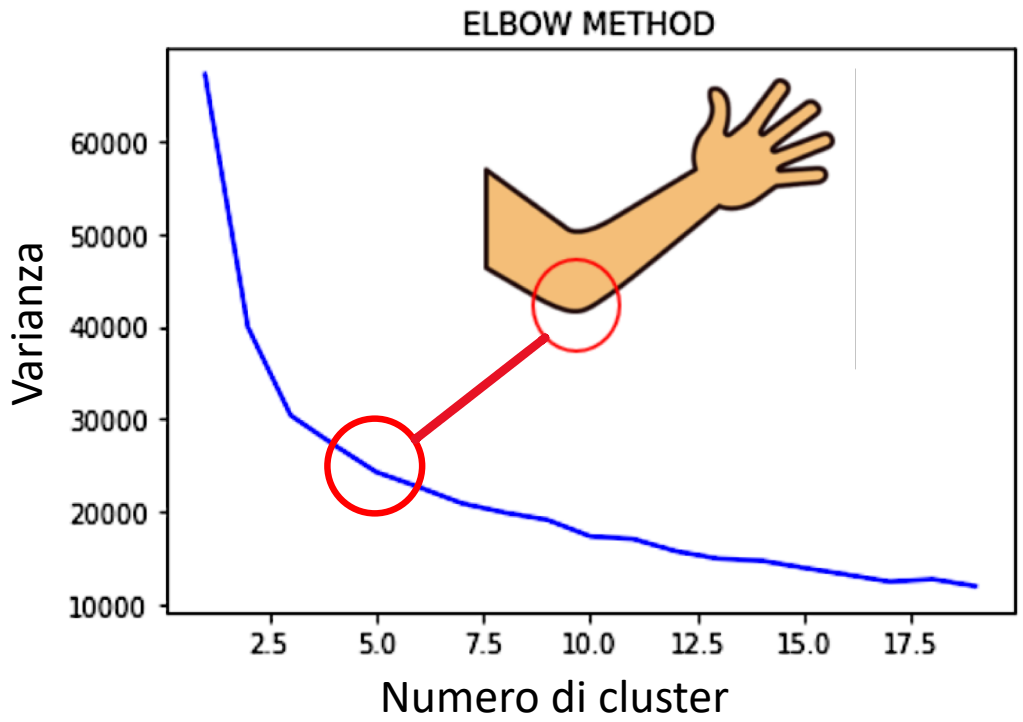
CLUSTERING

Raggruppa i dataset dai soli input

K-MEANS

Separa in n gruppi di uguale varianza per ridurre l'INERZIA

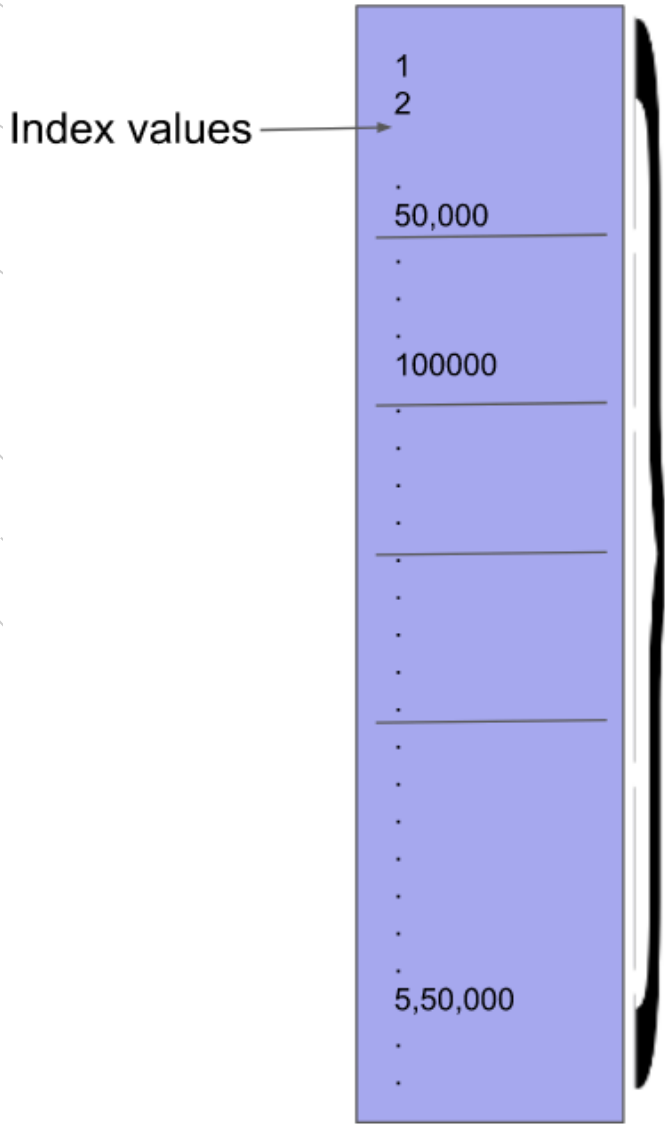
$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$



TECNOLOGIE UTILIZZATE

- LINGUAGGIO: PYTHON
- GUI: ANACONDA NAVIGATOR
- SCRITTURA CODICE: JUPYTER NOTEBOOK (applicazione web based di Anaconda)
- CLUSTERING: Scikit-learn

Pandas → dask



Pandas Dataframe

= dato tabulare bidimensionale, variabile in dimensione

Dask Dataframe

→ partizionato per riga e ogni blocco contiene un frame di dati Pandas

PUNTI PRINCIPALI DEL CODICE

PRE-PROCESSAMENTO

Suddivisione log in 5 colonne: timestamp, threadId, type, procId, message

MASKING

Rimozione dati sensibili
→ più workers e thread
(***for*** + ***delayed***)

VETTORIZZAZIONE

Estrazione del testo in una matrice
→ ***Hashing Vectorize***

ELBOW METHOD

K-MEANS

PRIMA

```
03/08 03:45:03.202 Thread 4 -  INFO [c74701c9-20ad-46db-acbb-96de84838c68]: process_request : Connection from
2001:1458:201:e3::100:6f4
03/08 03:45:03.275 Thread 4 -  INFO [c74701c9-20ad-46db-acbb-96de84838c68]: ns1__srmReleaseFiles : Request: Release files. IP:
2001:1458:201:e3::100:6f4. Client DN: /DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=ddmadmin/CN=531497/CN=Robot: ATLAS Data
Management. surl(s): srm://storm-
fe.cr.cnaf.infn.it/atlas/atlasdatadisk/rucio/data17_13TeV/99/2a/data17_13TeV.00338846.physics_Main.daq.RAW._lb0570._SFO-
6._0003.data. token: 78892824-8495-49b4-86e9-6db71bc110e9
03/08 03:45:03.287 Thread 4 -  INFO [c74701c9-20ad-46db-acbb-96de84838c68]: Result for request 'Release files' is
'SRM_SUCCESS'
03/08 03:45:03.694 Thread 35 -  INFO [68311b06-c66f-4d71-8a53-2e451c0df80b]: Request 'PTP status' from Client
IP='::ffff:131.154.194.217' Client DN='/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=atlpilo1/CN=614260/CN=Robot: ATLAS Pilot1'
# Requested token 'd3e7e258-9983-420a-bed3-c585bf90e928'
03/08 03:45:03.695 Thread 35 -  INFO [68311b06-c66f-4d71-8a53-2e451c0df80b]: Result for request 'PTP status' is 'SRM_SUCCESS'
03/08 03:45:04.406 Thread 35 -  INFO [68311b06-c66f-4d71-8a53-2e451c0df80b]: ns1__srmPutDone : Request: Put done. IP:
::ffff:131.154.194.217. Client DN: /DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=atlpilo1/CN=614260/CN=Robot: ATLAS Pilot1.
surl(s): srm://storm-
fe.cr.cnaf.infn.it/atlas/atlasscratchdisk/rucio/user/fkaya/da/5d/user.fkaya.364227.Sherpa_221_NNPDF30NNLO_Wenu_PTV1000_E_CMS.S
herpa_221.Rivet_ATLAS14_I1319490.Wenu_muChnnl.v01.log.20752208.000016.log.tgz.rucio.upload. token: d3e7e258-9983-420a-bed3-
c585bf90e928
03/08 03:45:04.420 Thread 35 -  INFO [68311b06-c66f-4d71-8a53-2e451c0df80b]: Result for request 'Put done' is 'SRM_SUCCESS'
03/08 03:45:04.423 Thread 35 -  INFO [68311b06-c66f-4d71-8a53-2e451c0df80b]: ns1__srmLs : Request: Ls. IP:
```

DOPO PRE-PROCESSAMENTO

	ProcId	timestamp	threadID	\	type	msg
0	c74701c9-20ad-46db-acbb-96de84838c68	03/08 03:45:03.20	4		INFO	process_request : Connection from 2001:1458:2...
1	c74701c9-20ad-46db-acbb-96de84838c68	03/08 03:45:03.27	4		INFO	ns1__srmReleaseFiles : Request: Release files...
2	c74701c9-20ad-46db-acbb-96de84838c68	03/08 03:45:03.28	4		INFO	Result for request 'Release files' is 'SRM_SU...
3	68311b06-c66f-4d71-8a53-2e451c0df80b	03/08 03:45:03.69	35		INFO	Request 'PTP status' from Client IP='::ffff:1...
4	68311b06-c66f-4d71-8a53-2e451c0df80b	03/08 03:45:03.69	35		INFO	Result for request 'PTP status' is 'SRM_SUCCESS'
...	...	...	...		...	...
99995	fd3ea1c6-542e-40a2-811a-6b1df26ef143	03/08 05:09:22.89	44		INFO	Request 'PTG status' from Client IP='2620:0:2...
99996	fd3ea1c6-542e-40a2-811a-6b1df26ef143	03/08 05:09:22.89	44		INFO	Result for request 'PTG status' is 'SRM_SUCCESS'
99997	6aa3e8e5-f080-4f10-984a-d991c26844f1	03/08 05:09:23.03	14		INFO	Request 'PTG status' from Client IP='2620:0:2...
99998	6aa3e8e5-f080-4f10-984a-d991c26844f1	03/08 05:09:23.03	14		INFO	Result for request 'PTG status' is 'SRM_SUCCESS'
99999	d46dc2d1-4069-4b19-a7b1-1862db176f9f	03/08 05:09:23.26	63		INFO	process_request : Connection from 2001:1458:2...

DOPO MASKING

process_request : Connection from IPmask process_request : Connection from IPmask ns1__srmLs : Request: Ls. IP: IPmask. Client DN: /DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=ddmadmin/CN=531497/CN=Robot: ATLAS Data Management Result for request 'Ls' is 'SRM_SUCCESS' ns1__srmLs : Request: Ls. IP: IPmask. Client DN: /DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=ddmadmin/CN=531497/CN=Robot: ATLAS Data Management Result for request 'Ls' is 'SRM_SUCCESS' Request 'PTG' from Client IP='IPmask' Client DN='/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=ddmadmin/CN=531497/CN=Robot: ATLAS Data Management' # Requested '1' SURL(s): 'srm://storm-fe.cr.cnaf.infn.it/atlas/atlasdatadisk/rucio/tests/8f/02/step14.2666.78200.recon.ESD.63309.28199' Result for request 'PTG' is 'SRM_REQUEST_QUEUED'. # Produced request token: 'f3e46cba-d0c6-4ee2-949c-f819e2e17b53' Request 'PTG status' from Client IP='IPmask' Client DN='/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=ddmadmin/CN=531497/CN=Robot: ATLAS Data Management' # Requested token 'f3e46cba-d0c6-4ee2-949c-f819e2e17b53' Result for request 'PTG status' is 'SRM_REQUEST_QUEUED' Request 'PTG status' from Client IP='IPmask' Client DN='/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=ddmadmin/CN=531497/CN=Robot: ATLAS Data Management' # Requested token 'f3e46cba-d0c6-4ee2-949c-f819e2e17b53' Result for request 'PTG status' is 'SRM_REQUEST_QUEUED' Request 'PTG status' from Client IP='IPmask' Client DN='/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=ddmadmin/CN=531497/CN=Robot: ATLAS Data Management' # Requested token 'f3e46cba-d0c6-4ee2-949c-f819e2e17b53' Result for request 'PTG status' is 'SRM_SUCCESS'

TEMPO DI SPEED - UP

$$SPEED\_UP = \frac{T\ SINGOLO\ PROCESSO}{T\ MOLTI\ PROCESSI}$$

1.06 → **1.3**

100.000 RIGHE

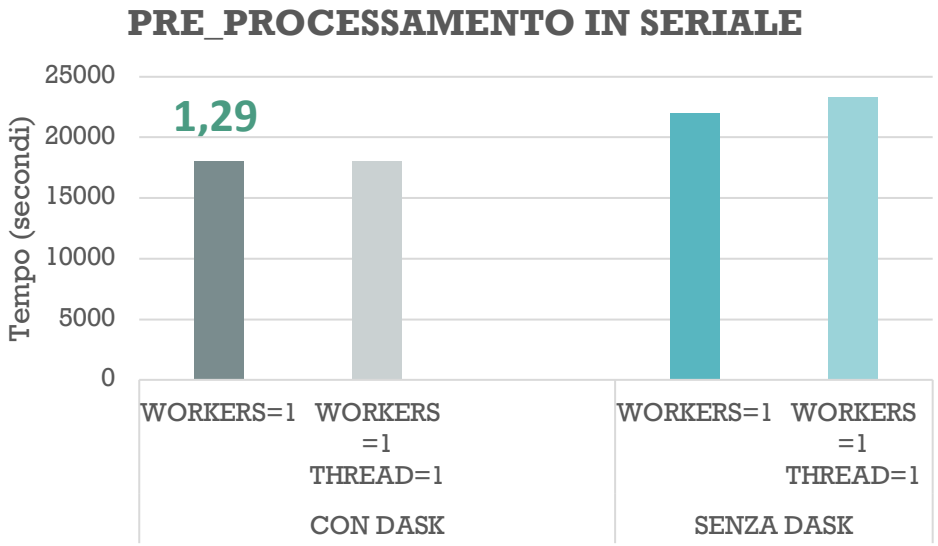
1.000.000 RIGHE

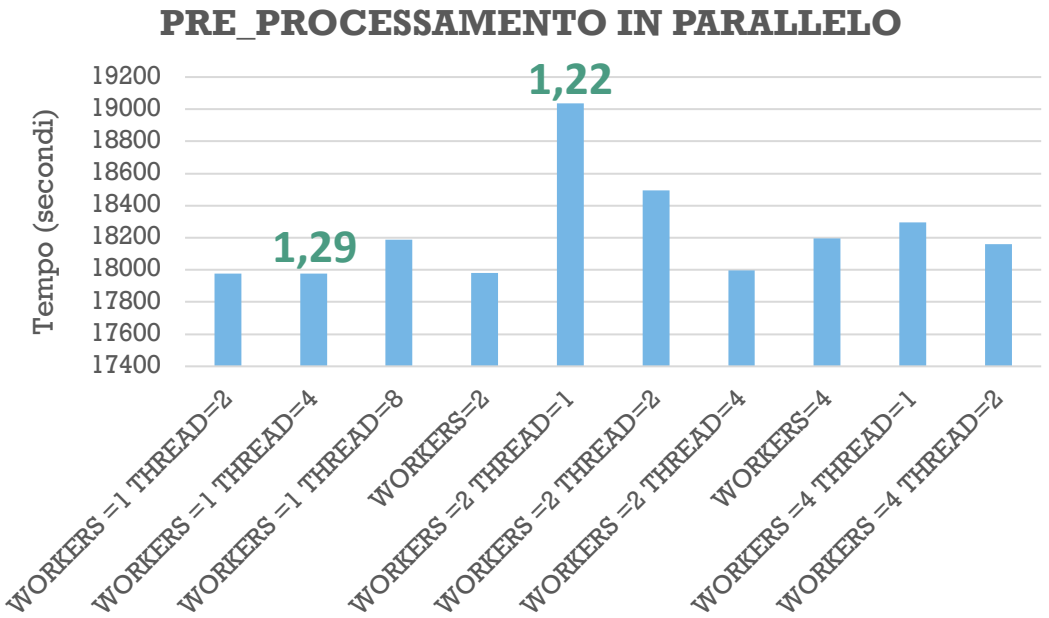
Configurazioni studiate

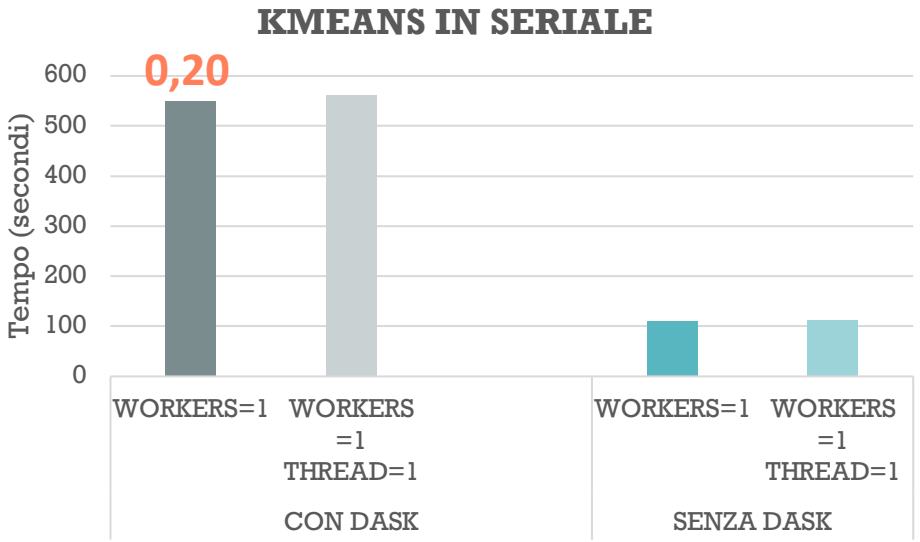
WORKERS =1 THREAD=2
WORKERS =1 THREAD=4
WORKERS =1 THREAD=8
WORKERS=2
WORKERS =2 THREAD=1
WORKERS =2 THREAD=2
WORKERS =2 THREAD=4
WORKERS=4
WORKERS =4 THREAD=1
WORKERS =4 THREAD=2

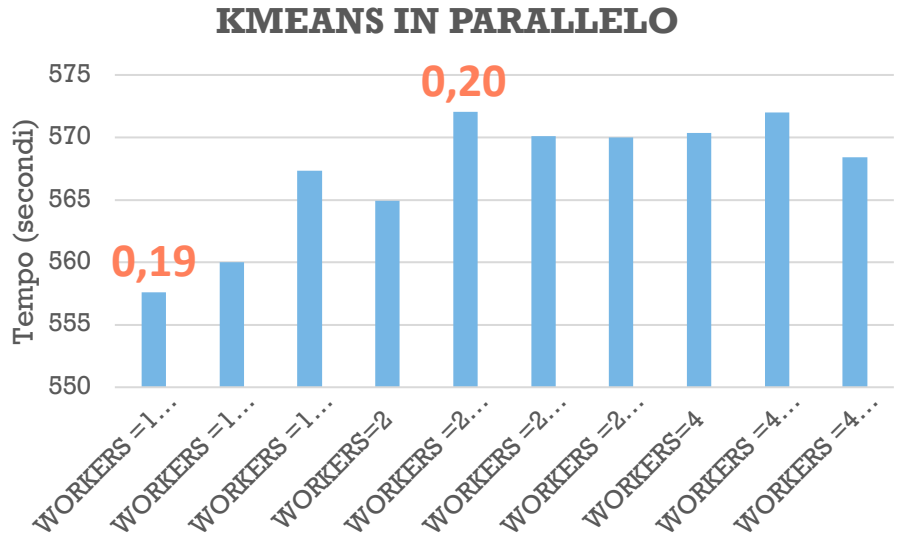
Dispositivo usato: Mac Mini con chip Apple M1, 8 core (4 di prestazioni e 4 di efficienza), 8 GB di Ram, sistema operativo MacOS Monterey

RISULTATI

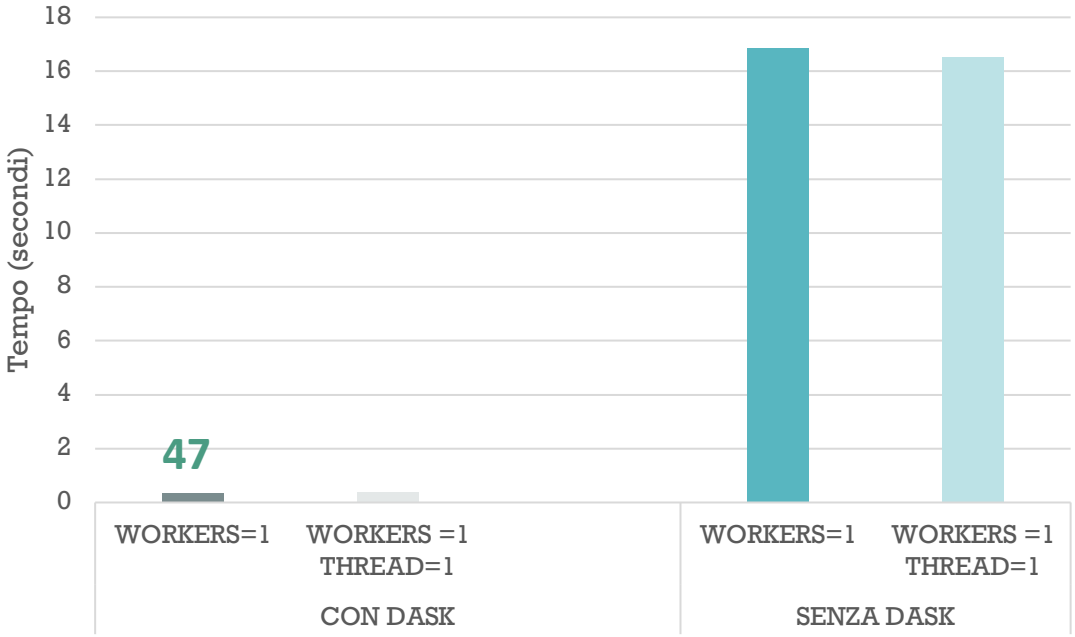

PRE_PROCESSAMENTO IN SERIALE
Tempo (secondi)
25000
20000
15000
10000
5000
0
1,29
WORKERS=1
WORKERS =1 THREAD=1
CON DASK
WORKERS=1
WORKERS =1 THREAD=1
SENZA DASK


PRE_PROCESSAMENTO IN PARALLELO
Tempo (secondi)
19200
19000
18800
18600
18400
18200
18000
17800
17600
17400
1,22
1,29
WORKERS =1 THREAD=2
WORKERS =1 THREAD=4
WORKERS =1 THREAD=8
WORKERS=2
WORKERS =2 THREAD=1
WORKERS =2 THREAD=2
WORKERS =2 THREAD=4
WORKERS=4
WORKERS =4 THREAD=1
WORKERS =4 THREAD=2

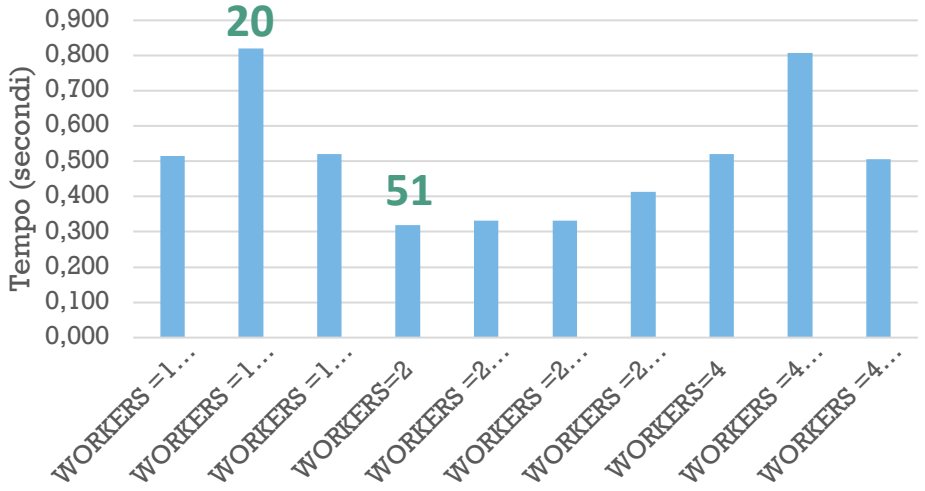

KMEANS IN SERIALE
Tempo (secondi)
600
500
400
300
200
100
0
0,20
WORKERS=1
WORKERS =1 THREAD=1
CON DASK
WORKERS=1
WORKERS =1 THREAD=1
SENZA DASK


KMEANS IN PARALLELO
Tempo (secondi)
575
570
565
560
555
550
0,20
0,19
WORKERS =1…
WORKERS =1…
WORKERS =1…
WORKERS=2
WORKERS =2…
WORKERS =2…
WORKERS =2…
WORKERS=4
WORKERS =4…
WORKERS =4…

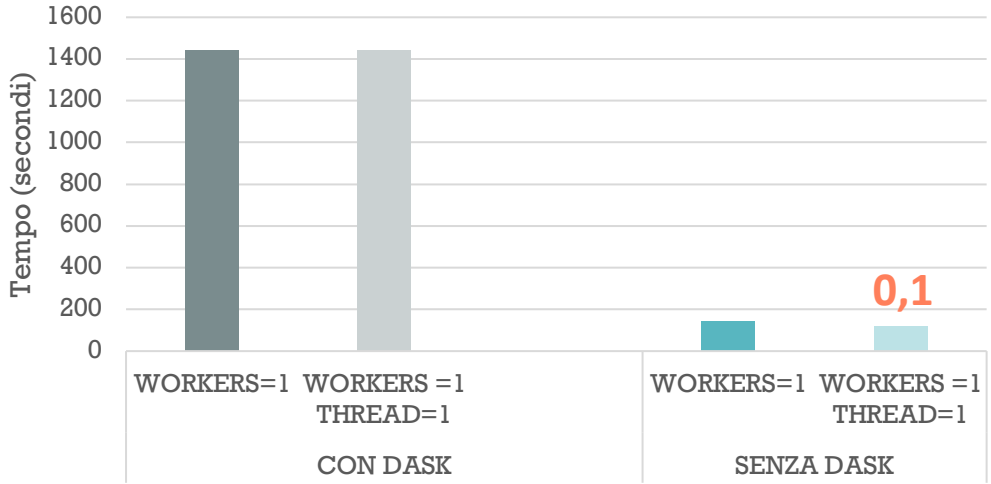
VETTORIZZAZIONE IN SERIALE



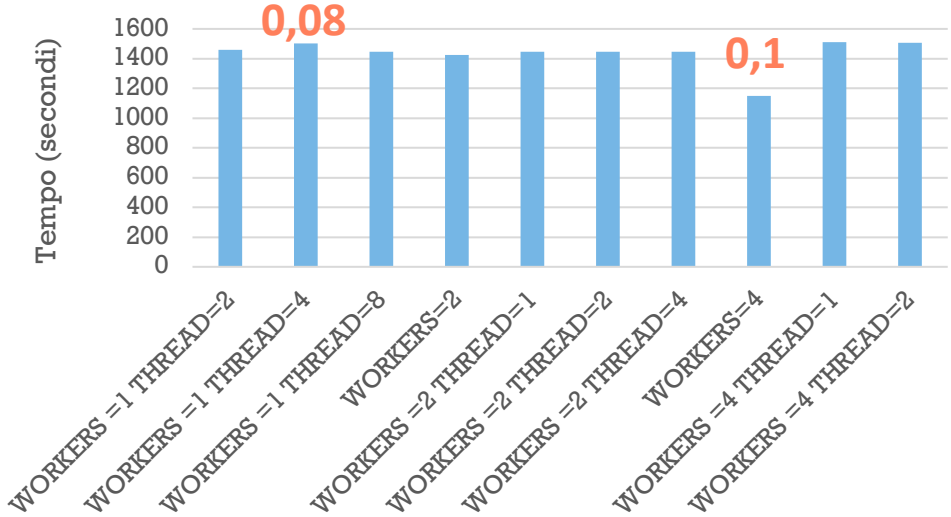
VETTORIZZAZIONE IN PARALLELO



ELBOW METHOD IN SERIALE



ELBOW METHOD IN PARALLELO



SVILUPPI FUTURI e CONLUSIONI

GRAZIE PER L'ATTENZIONE

SLIDE DI BACK-UP

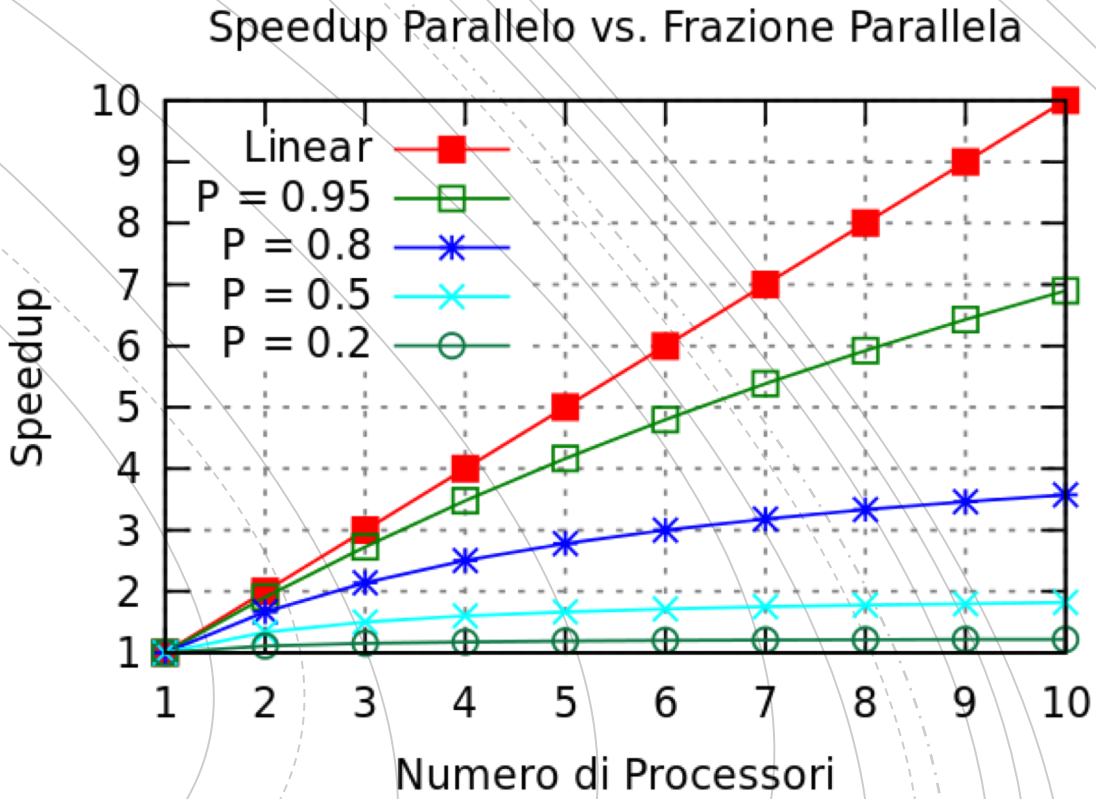
	PRE_PROCESSAMENTO	KMEANS	VETTORIZZAZIONE	ELBOW METHOD	CONTEGGIO NUMEROSITÁ CLUSTER
SPEED_UP	1,294379419	0,200675343	32,10108868	0,082401712	0,34462451
	1,294528163	0,199813616	20,17386125	0,080066638	0,309820752
	1,279429082	0,197241065	31,73241317	0,08293434	0,372280582
	1,294267189	0,198077094	51,86551627	0,084401232	0,355132737
	1,222372186	0,195623331	49,94924515	0,08308884	0,364479556
	1,258125029	0,196289392	49,99074578	0,08308884	0,353650424
	1,292946715	0,196322757	39,99199248	0,083085511	0,34451156
	1,278793078	0,196197105	31,73241317	0,104698363	0,345078662
	1,271875863	0,195635577	20,49090202	0,079439216	0,349318745
	1,281382648	0,196860836	32,72639293	0,079693404	0,343909664
	da 1,22 1,29	**0,2**	**da 20 a 50**	**da 0,08 a 0,1**	**da 0,31 a 0,37**

LEGGE DI AMDAHL

- Usata per trovare il **miglioramento atteso massimo** in una architettura di calcolatori o in un sistema informatico quando vengono migliorate solo alcune parti del sistema

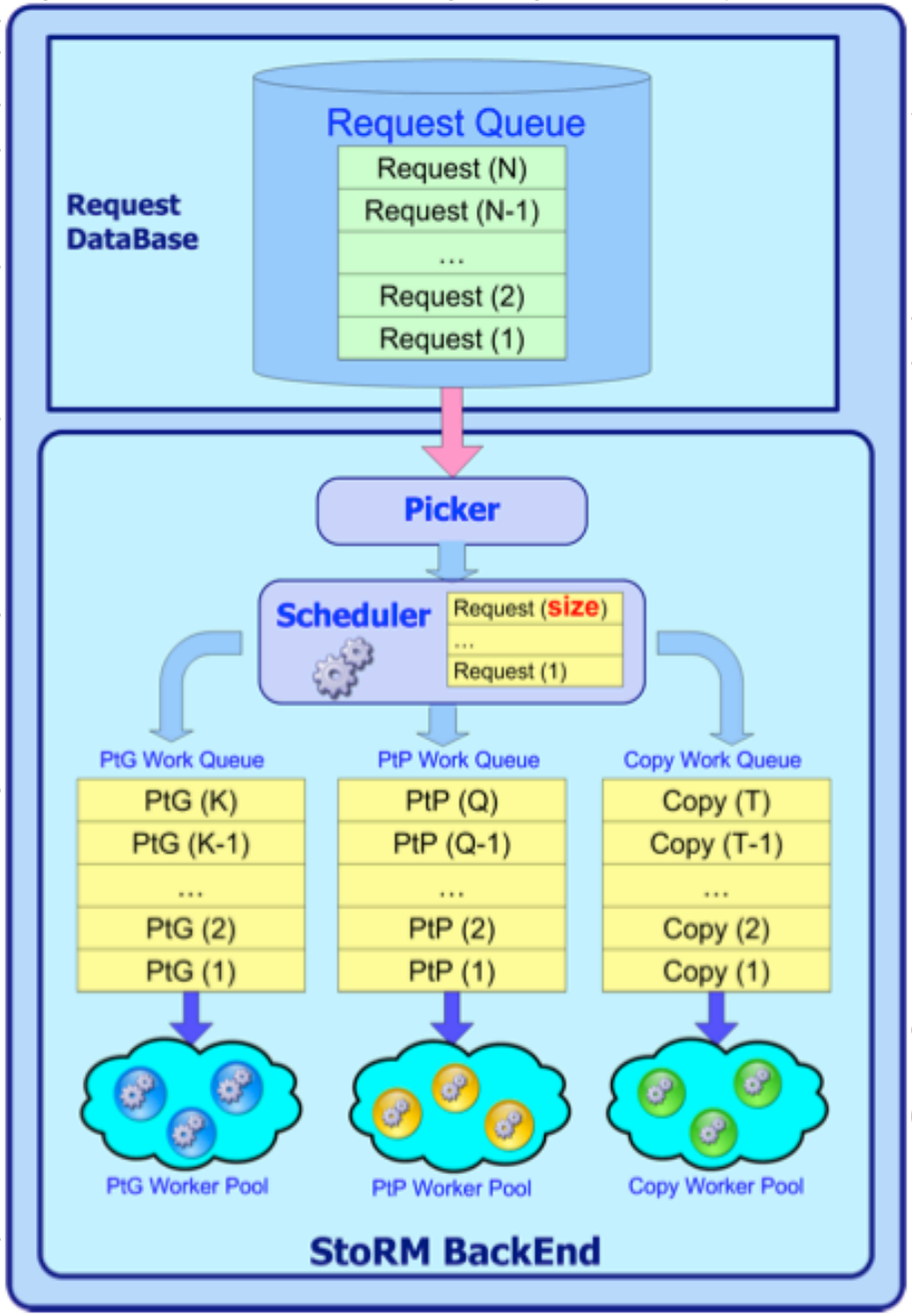
$$\frac{1}{(1-F)+\frac{F}{N}}$$

F = frazione di un calcolo che è parallelizzabile (cioè che può beneficiare dal parallelismo)
(1 – *F*) = frazione che non può essere parallelizzata
N = processori



RICHIESTE SRM DAL DB

Il componente **Picker** recupera la quantità specificata di nuove richieste SRM dal database a ogni intervallo di tempo e le inoltra a uno Scheduler. Lo **Scheduler** si occupa di inoltrare la richiesta al thread corretto come una nuova attività da eseguire. Lo stato della richiesta viene aggiornato nel Database con tutte le informazioni relative ai risultati della richiesta, errori e altri dati. Questi dati sono accessibili dal Frontend per rispondere a una richiesta srmStatusOf*.



SICUREZZA DI StoRM

1. L'utente fa una richiesta con il suo proxy (si spera con estensioni VOMS)
2. StoRM verifica se l'utente può eseguire l'operazione richiesta sulla risorsa richiesta
3. StoRM chiede la mappatura dell'utente al servizio LCMAPS
4. StoRM applica un vero ACL (Access Control List) sul file e sulle directory richieste
5. I lavori in esecuzione per conto dell'utente possono eseguire un accesso diretto ai dati.