Implementazione, creazione e ottimizzazione di una pipeline per l'analisi biofisica su cluster a basso consumo energetico

Daniele Dall'Olio
Relatore: Dott. Enrico Giampieri
Correlatori: Prof. Gastone Castellani Ing. Andrea Ferraro

ALMA MATER STUDIORUM · UNIVERSITÁ DI BOLOGNA

22 Settembre 2017

Problema

- Costo medio elevato
- Consumo energetico elevato
- Spese per il raffreddamento elevate

Conseguenze

- Minor accessibilità
- Poche unità acquistabili
- Ridotta scalabilità e flessibilità per aggiornare l'hardware dei server

Tecnologia di calcolo low power

Vantaggi

- Costo delle singole unità basso
- Consumo elettrico inferiore
- Flessibilità nell'acquisto di nuovi hardware

Svantaggi

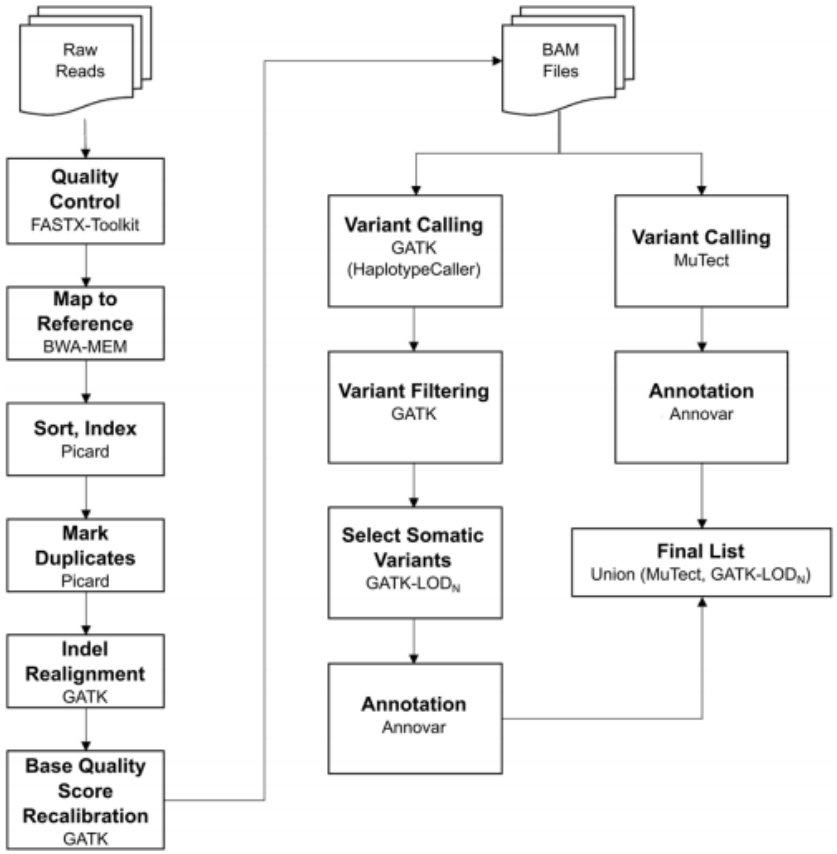
- Potenza inferiore
- Cache ridotta
- Numero inferiori di core

Obiettivo della tesi

Ottenere risultati con i nodi low power comparabili a quelli ottenuti con i nodi tradizionali.

GATK-LODn

Requisiti molto elevati in termini di potenza di calcolo, di occupazione di memoria e di spazio d'archiviazione.



Next Generation Sequencing

- Comprende le nuove tecniche per il sequenziamento del DNA.
- Tecniche più rapide e meno costose dei metodi precedenti.
- Shotgun Sequencing.

Gli algoritmi di analisi sui dati di NGS sono basati anche sulla teoria dei network.

Studio sulle varianti

- Confronto con il genoma di riferimento.
- Individuazione delle mutazioni somatiche sia nel paziente che nel tumore.
- Ricerca di quelle mutazioni che sono presenti solo nel tumore.
- Confronto tumori simili tra diversi soggetti.

Lavoro svolto per la tesi

- **Reimplementazione di una parte di GATK-LOD$_n$ nel tool Snakemake**
- Scritti i file di configurazione per il programma
- Scritto uno script per l'estrazioni di subset
- Scritto uno script per raggruppare i dati
- Creato l'installer
- Creato uno script per l'esecuzione delle simulazioni
- Effettuate le simulazioni
- Analisi dei dati

Regole indipendenti dal paziente

Indicizzazione del genoma di riferimento per:

- BWA
- Picard
- Samtools(e GATK)

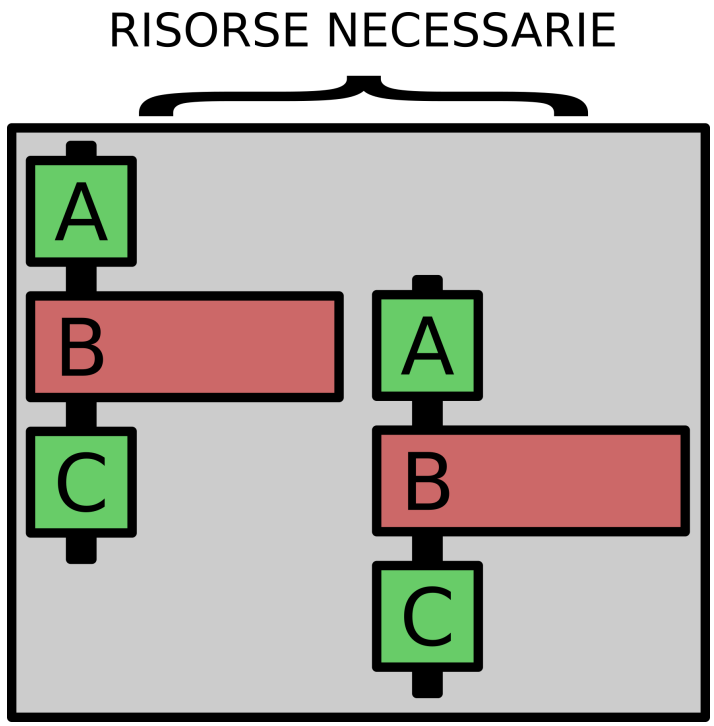
Regole dipendenti dal paziente

- **Mapping**: mappatura delle sequenze del paziente sul riferimento.
- **Sort Picard**: riordinamento dei file.
- **Mark Duplicates**: identificazione dei duplicati.
- **Build BAM**: indicizza i file per velocizzare l'analisi.
- **Realigner**: determina gli intervalli che necessitano del riallineamento.

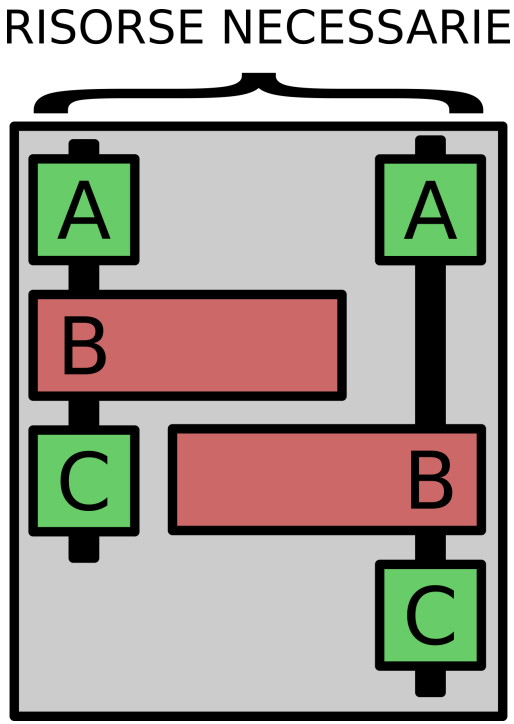
Snakemake

- Sistema di gestione dei flussi di lavoro
- Pochi requisiti
- Scritto in Python e modellato su Make
- Gestione specifica delle risorse
- Permette di eseguire su cluster

Procedura tradizionale



Procedura ricercata



Analisi effettuate

- Tempo di esecuzione
- Memoria utilizzata

Simulazioni effettuate

numero di letture	dimensione su disco
1×10^5	2x 28.4 MB
1×10^6	2x 284.9 MB
3×10^6	2x 854.9 MB
9×10^6	2x 2.6 GB
4.5×10^7	2x 12.8 GB

Tabella: Stima della dimensione dei subset in relazione al numero di letture. L'ultimo valore si riferisce all'intero paziente.

Nodo	*CPU*	*Memory*	*Storage*	*Costo**	*Consumo**
xeond	1x Xeon D-1540	16 *GB*	8 *TB*(*HDD*)	€1000	60 *W*
avoton	1x Atom C2750	16 *GB*	5 *TB*(*HDD*)	€600	30 *W*
*n*3700	1x Pentium N3700	8 *GB*	0.5 *TB*(*SSD*)	€130	8 *W*
*bio*8	2x Xeon E5-2620v4	128 *GB*	2 *TB*(*HDD*)	€10000	180 *W*

* I valori di costo e consumo energetico sono stimati.

Tabella: Caratteristiche dei nodi.

CPU	*Microarchitecture*(*Platform*)/*litho*	*Freq*(*GHz*)	*Cores*	*Cache*	*TDP*
Xeon D-1540	*Broadwell*/14*nm*	2.0(2.60)	8(16)	12 *MB*	45 *W*
Atom C2750	*Silvermont*(*Avoton*)/22*nm*	2.40(2.60)	8	4 *MB*	25 *W*
Pentium N3700	*Airmont*(*Braswell*)/14*nm*	1.60(2.40)	4	2 *MB*	6 *W*
Xeon E5-2620v4	*Broadwell* − *EP*/14*nm*	2.10(3.00)	8(16)	20 *MB*	85 *W*

Tabella: Caratteristiche delle CPU.

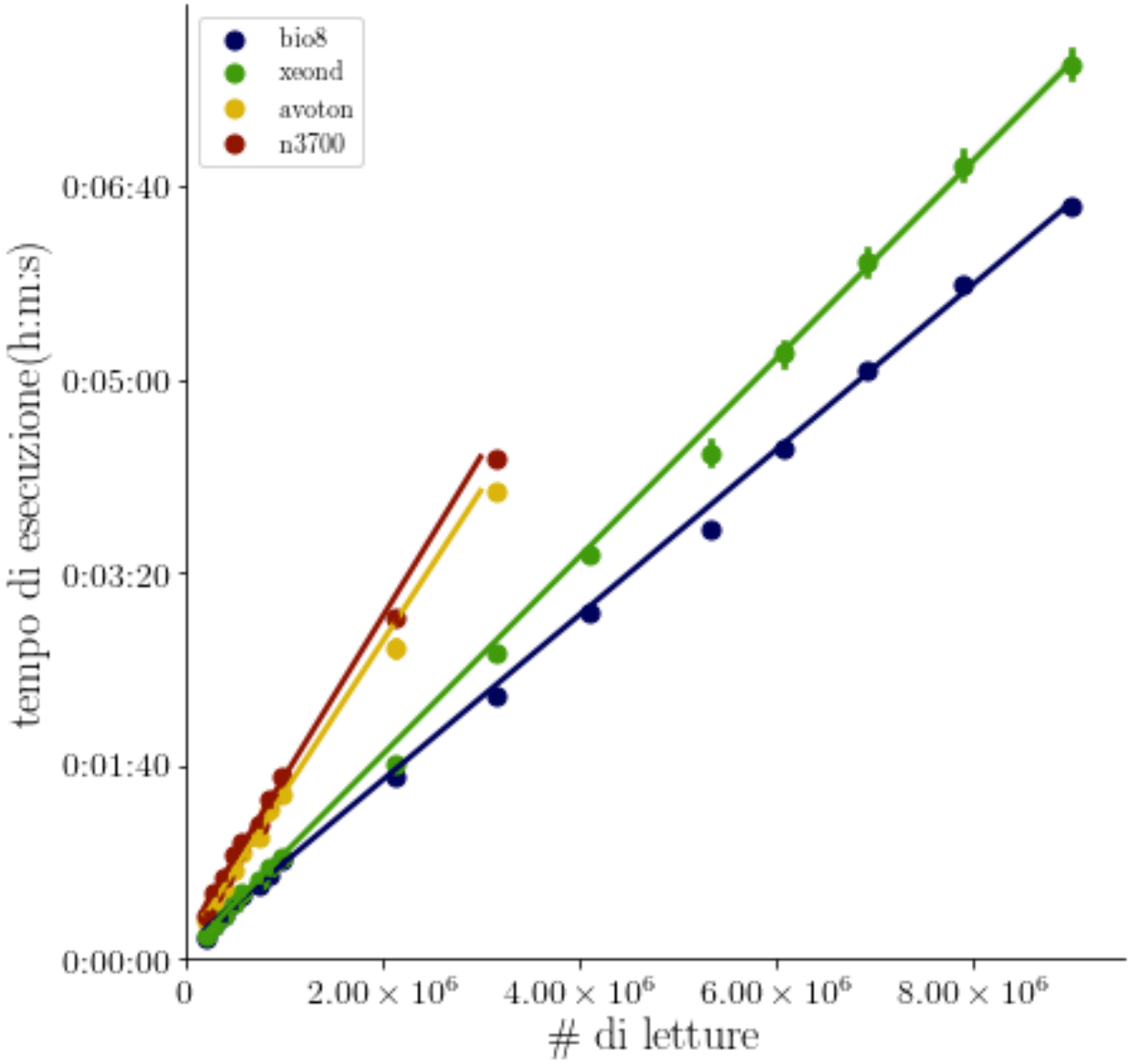


Figura: Tempi per Sort Picard.

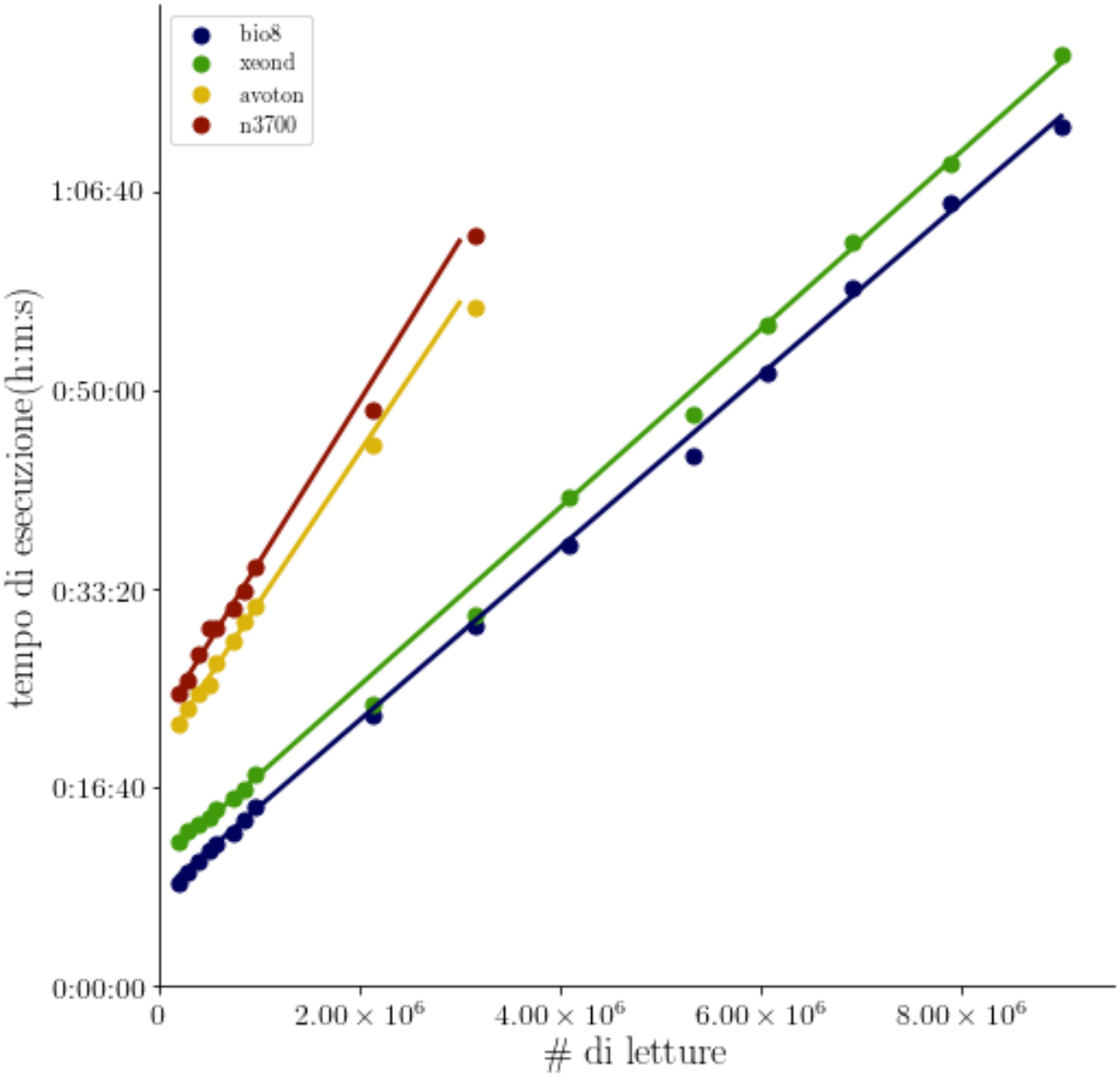


Figura: Tempi complessivi.

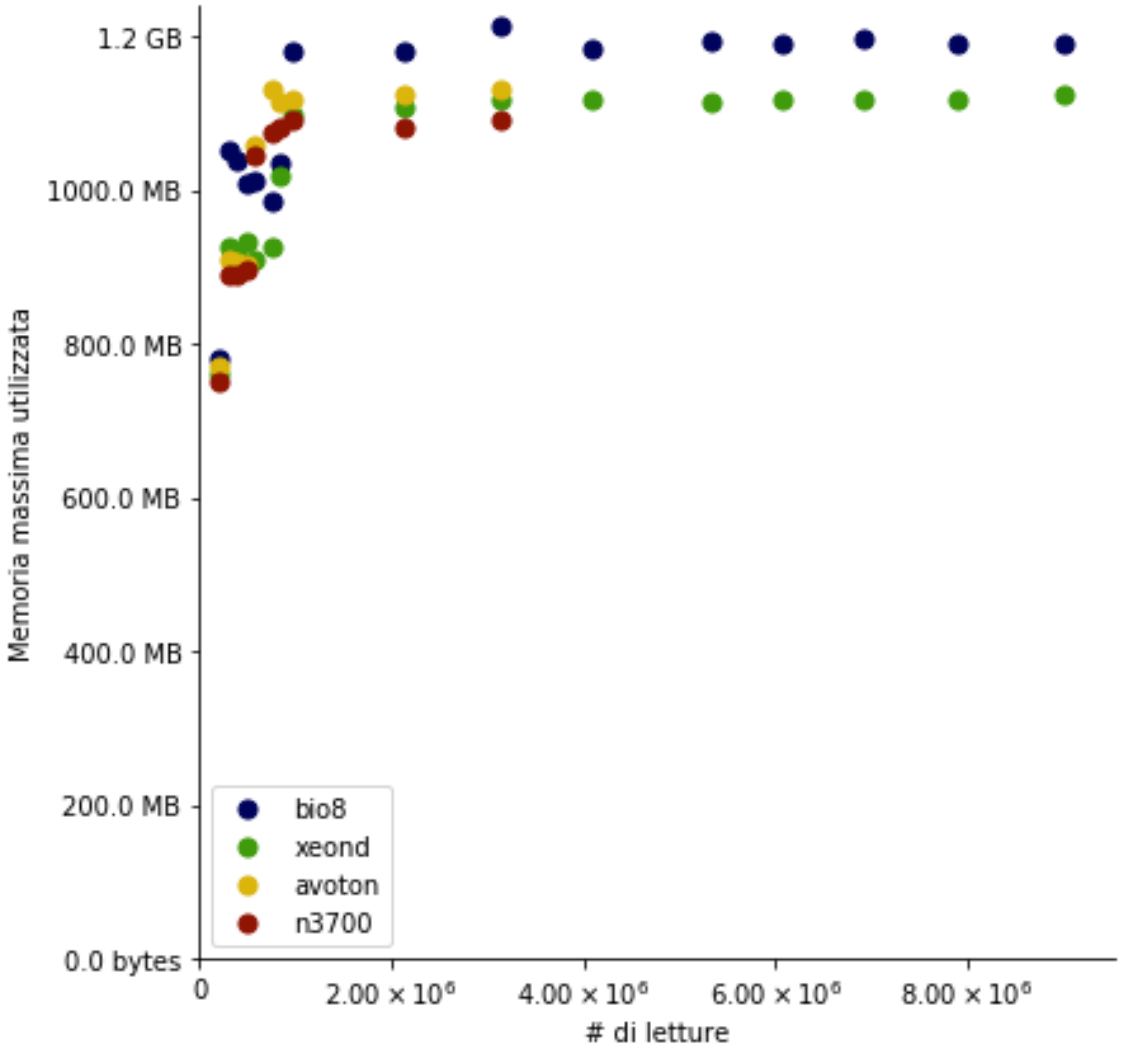


Figura: Sort Picard.

Tempo di esecuzione

- avoton e n3700 impiegano il doppio del tempo
- xeond è comparabile a bio8 consumando un terzo dell'energia e costando 10 volte di meno

Memoria utilizzata

- Saturazione
- Valori di saturazioni consistenti
- Sempre inferiore al massimo di memoria accessibile

Conclusione

In base a questi risultati questa pipeline di calcolo bioinformatico sembra essere realisticamente eseguibile anche su nodi a bassa potenza senza una perdita considerevole di prestazioni.

Sviluppo futuro

- Simulazioni a core multipli sui singoli nodi
- Completamento della pipeline
- Simulazioni su cluster

Una volta terminati questi passi intendiamo pubblicarli.