Il calcolo su larga scala

Dall'analisi dei dati genetici all'analisi del web

Luca Pireddu

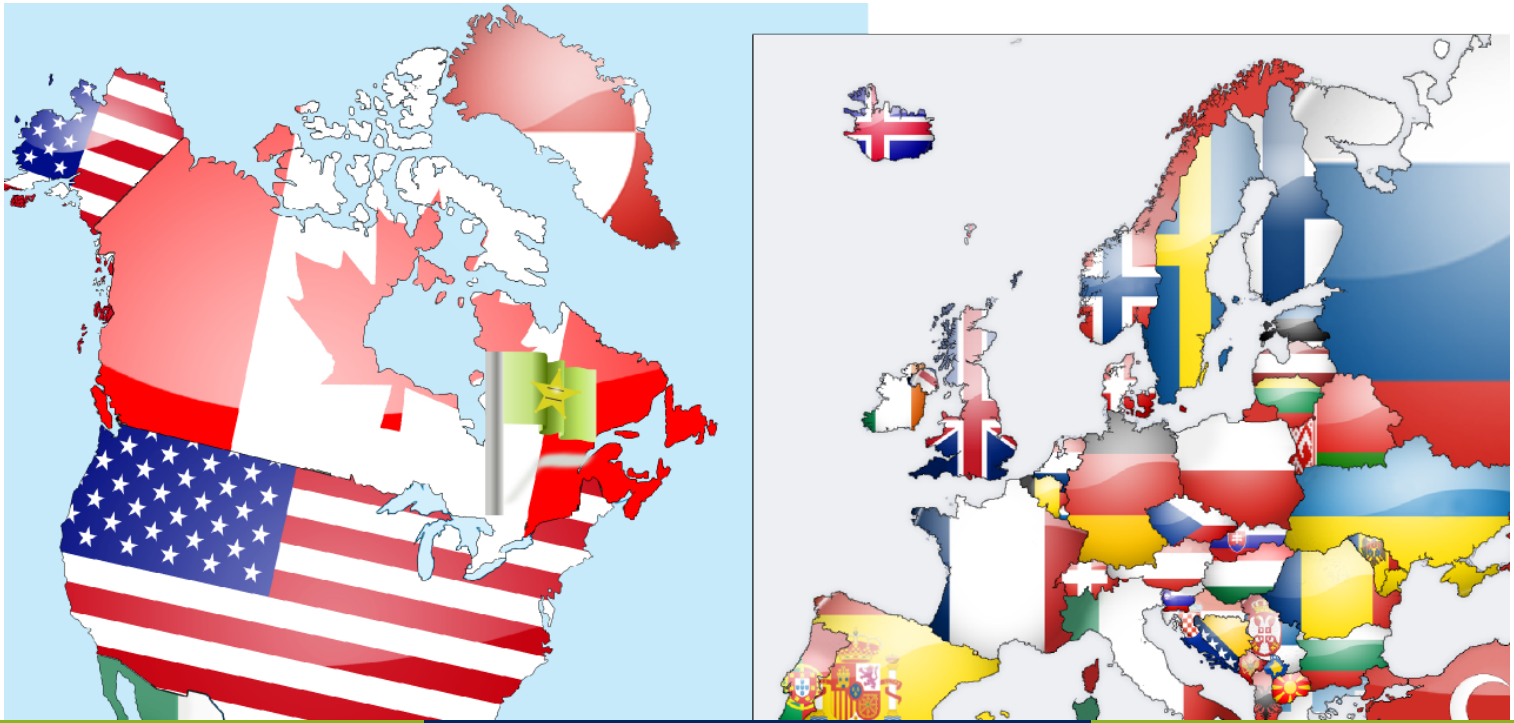
CRS4—Distributed Computing Group

October 13, 2011

Mi presento

Mi presento

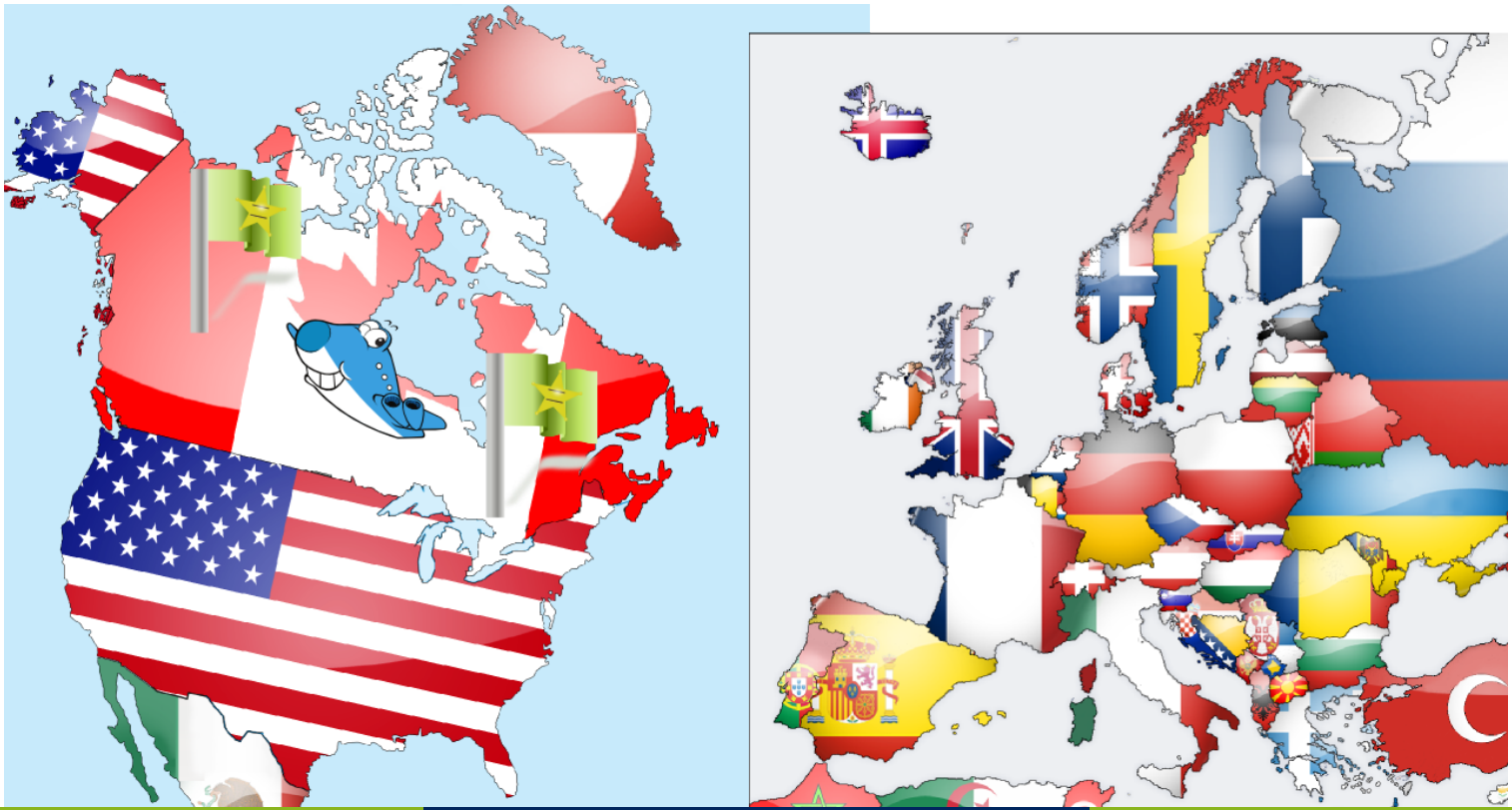
- Abito in Italia da 5 anni
- Sono nato e in gran parte cresciuto in Canada
- Sono un ricercatore al CRS4
- No, Pireddu non è un cognome tipico canadese; i miei genitori sono sardi al 100%!



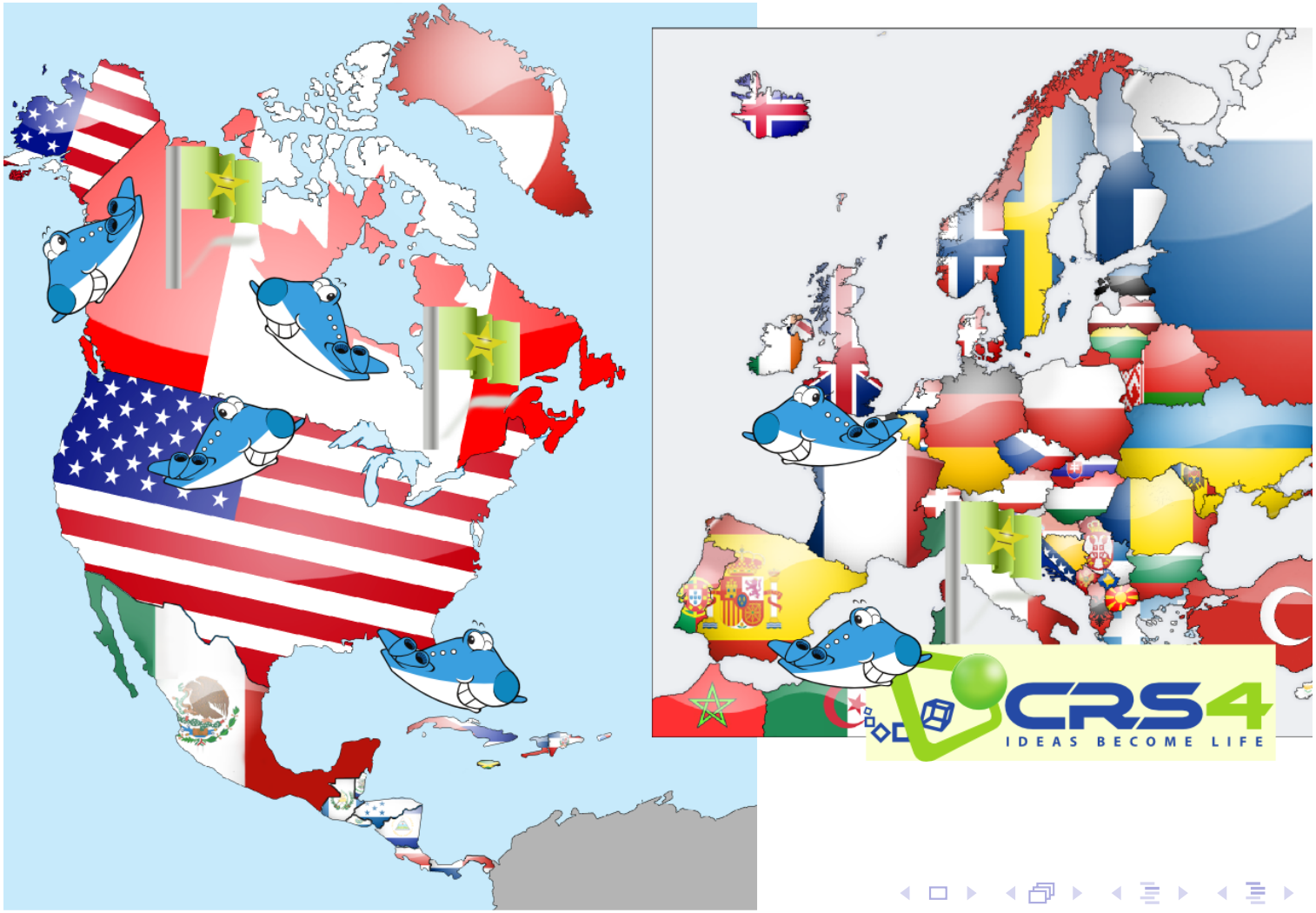
Mi presento

- Ho studiato informatica in Canada
  - Laurentian University
  - University of Alberta

Da li sono partito nella mia carriera e dopo un po' sono arrivato qui da voi



O, più precisamente, al CRS4



Cos'è il CRS4

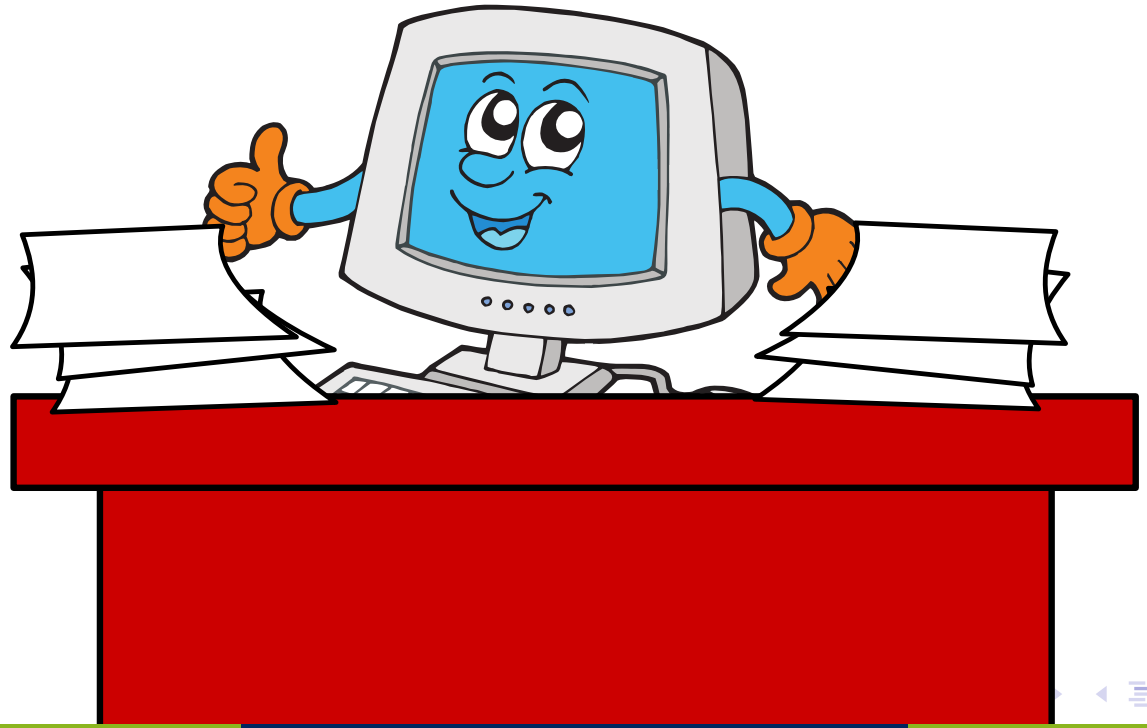
- Centro per la ricerca applicata
- Particolarmente orientato all'informatica e le alte tecnologie
- Fondato nel 1990, in Sardegna
- Alcuni primati:
  - ha realizzato il primo sito web italiano
  - ha contribuito a creare il primo quotidiano sul web in Europa
  - ha contribuito a creare uno dei primi internet provider in Italia
  - prima laboratorio per il sequenziamento del DNA ad alta velocità in Italia
- Oggi principalmente ricerca l'applicazione dell'informatica a:
  - la biologia e medicina
  - le fonti di energia
  - la visualizzazione dei dati
  - la tecnologia/internet mobile
- Inoltre, possiede uno dei maggiori centri di calcolo in Italia

Introduzione al Calcolo su larga scala

In principio. . .

In principio l'uomo creò un computer. . . e vide che era cosa buona!

- Con il computer si elaboravano dati più velocemente che a mano
- Fino a 15-20 anni fa in genere si elaboravano quantità modeste di dati
- Di solito 1 computer o mainframe era sufficiente
  - e in molte situazioni questo è ancora vero!



La crescita dei dati

- Da 15-20 anni c'è stata un'accelerazione importante della crescita della mole di dati da analizzare

Perché?

- In parte perché la tecnologia aumenta i limiti e propone nuove possibilità
- In parte perché abbiamo più fonti di dati o fonti più grandi
  - Internet! Viene letta interamente e ripetutamente
  - L'avanzamento della scienza (pensiamo al sequenziamento di genomi umani e al LHC al CERN)
  - Modelli meteorologici più dettagliati
  - Le varie carte fedeltà che ci danno i negozi
  - I dati raccolti dagli dispositivi informatici portabili (cellulari, sensori, lettori di codici a barre e RFID)
  - ecc.
- L'informazione è diventata un bene primario!

Ma quanti dati?

Il nostro PC a casa, se relativamente nuovo, potrà contenere circa:

- 20 giorni di video
- 50.000 foto

Un data center moderno può contenere dati a sufficienza per riempire:

- **un milione** dei nostri PC a casa
  - Almeno tutti i PC in Sardegna; forse di più!
- **decine di migliaia di anni** di video
  - Un documentario dalla fine dell'ultima era glaciale ad oggi!
- **decine di miliardi** di foto!
  - 1 foto ogni ha per l'intera superficie terrestre

Un computer?

Che fine farebbe un solo computer con tutti questi dati?

Un computer?

Che fine farebbe un solo computer con tutti questi dati?

Big Data

- Questo è il fenomeno detto **Big Data**
- Lavorare con più dati spesso porta benefici
  - ricerche più accurate
  - analisi più sensibili
- Però ha il prezzo della sua complessità
  - In particolare, queste quantità di dati non sono manipolabili con strumenti tradizionali
  - Come risolvere?

Lavoro di squadra

- Quel che uno non riesce a fare da solo lo si può fare in tanti!
- Facciamo lavorare molti computer sullo stesso problema
- Ognuno può contribuire un po' e con la somma dei contributi arrivare all'obbiettivo

Questo l'hanno fatto tante tante piccole termiti!

Calcolo Distribuito

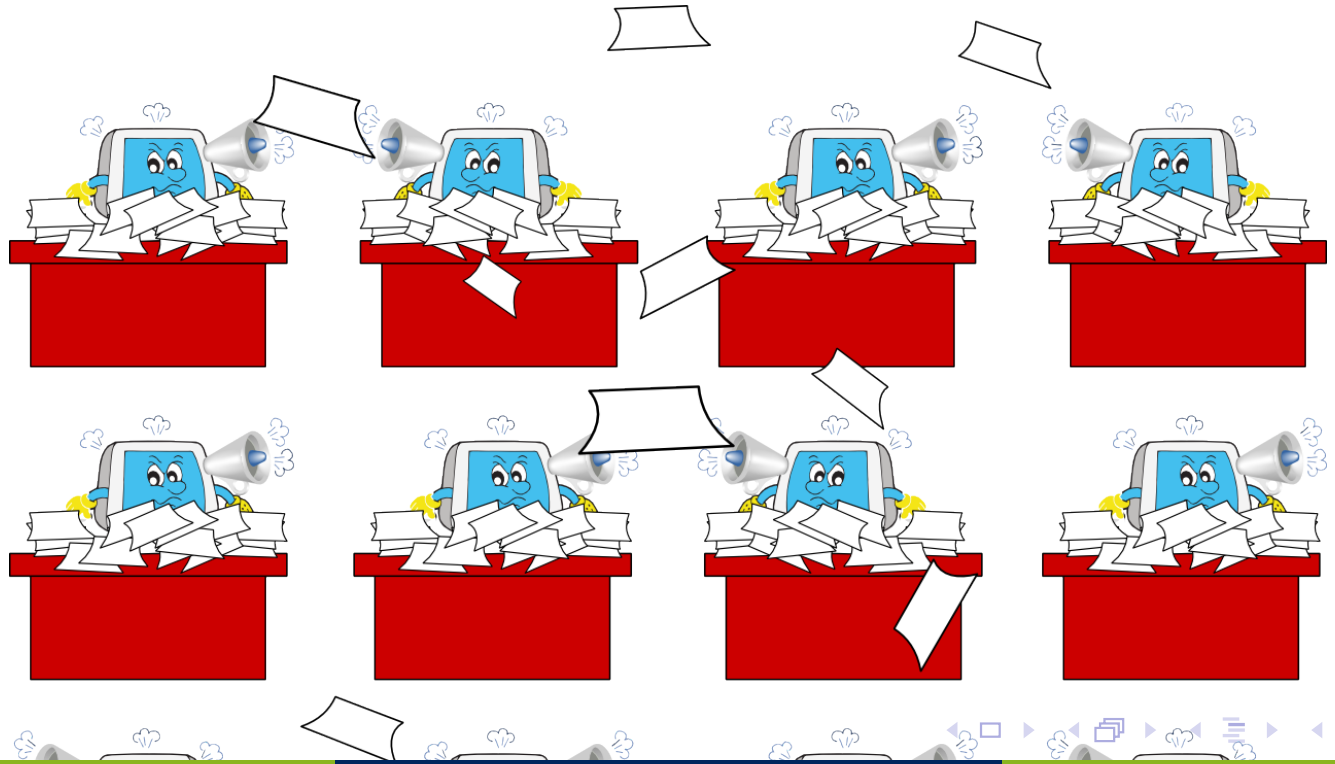
- Esiste un ramo di studio dedicato specificamente a come coordinare il lavoro di più computer

Calcolo Distribuito
in inglese,
Distributed Computing

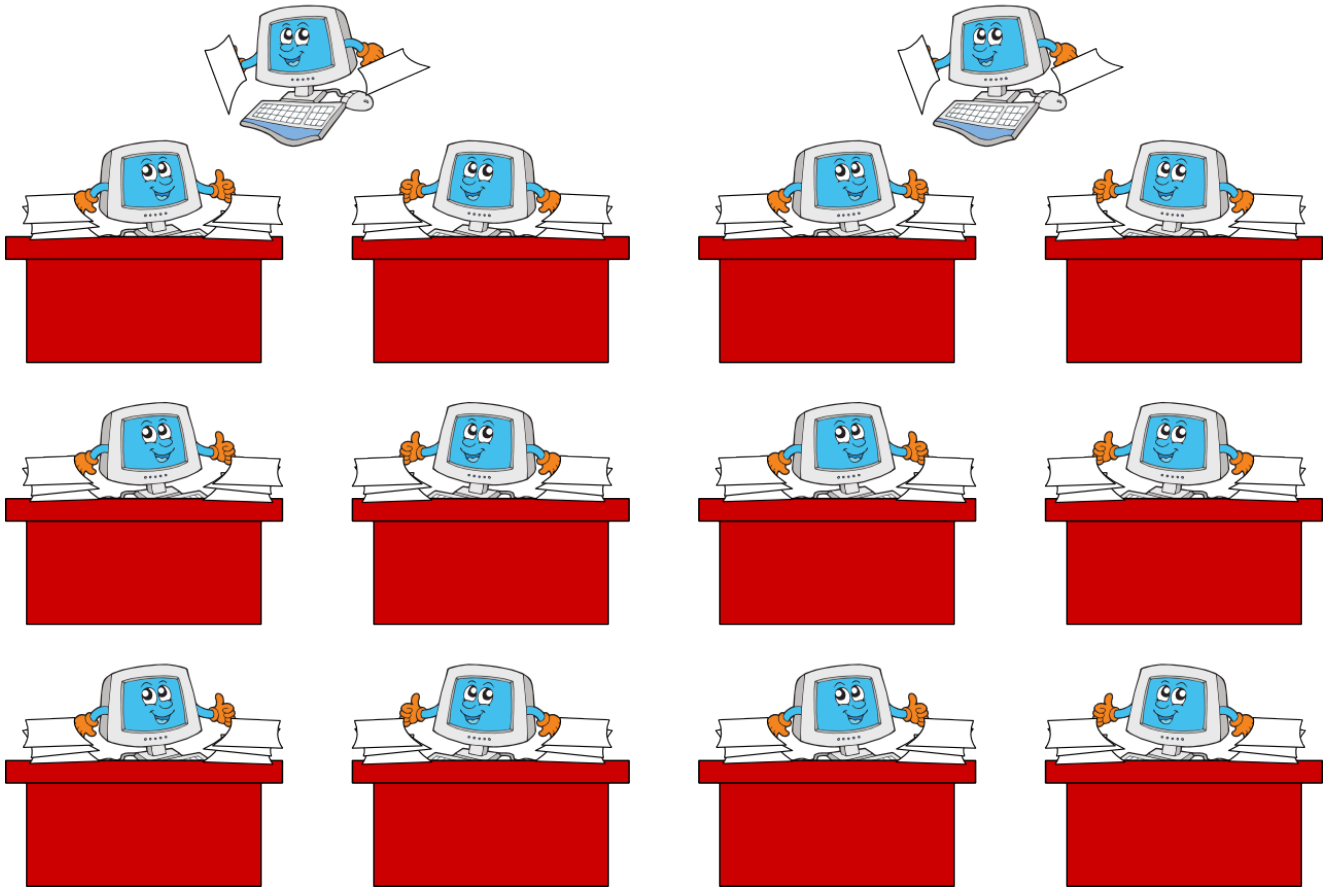
. . . avrete sentito che

troppi cuochi rovinano il brodo!

- Lavorare assieme non è banale
- È necessario essere organizzati, altrimenti. . .



- Se invece coordiniamo bene



Calcolo Distribuito

Quindi...

- Il calcolo distribuito è necessario per risolvere problemi di elaborazione di dati su larga scala.

Applicazioni del calcolo a larga scala

Applicazioni

- Sicuramente avete visto, e forse usate giornalmente, prodotti di questo genere di tecnologia
- Conoscete qualcuna di queste aziende?

Applicazioni

Google Oregon

Facebook Rutherford

Microsoft Dublin

Bank of NY Mellon

Anche il CRS4 ha il suo data center!

- 600 computer (≈4000 core)
- migliaia di dischi

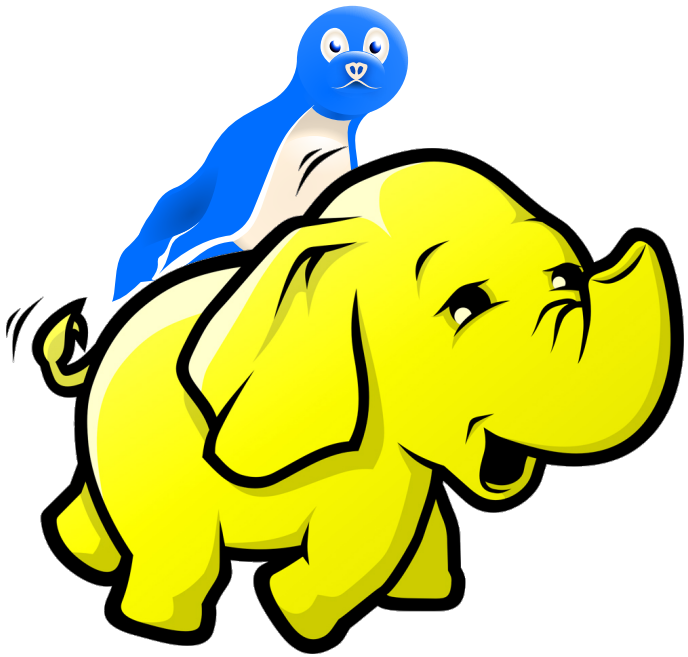
- Queste macchine sono condivise da tutti i gruppi del CRS4
- Vengono usate per vari compiti, tra i quali:

L'analisi del DNA

- Anche grazie al lavoro del mio gruppo!

Tra i nostri progetti c'è Seal

- In origine il nome venne da SEquence ALignment
- Ora è un gruppo di programmi fatti per analizzare i dati prodotti dal sequenziamento del DNA
  - Come quelli prodotti dalla Dr.ssa Maria Francesca Urru
- A differenza dei programmi tradizionali, Seal è creato per il *calcolo distribuito*



Tra i vantaggi del nostro approccio:

- Maggiore velocità
- Miglior utilizzo delle risorse computazionali
- Minor attenzione richiesta dagli operatori
  - Tenete conto: più computer avete maggiore è la probabilità che uno si rompa!
- Oltre che al CRS4, Seal è attualmente in prova anche in centri di sequenziamento in Finlandia e Olanda
  - Fatto interessante: anche loro stanno facendo degli studi genetici simili a quelli che stiamo conducendo in Sardegna

Conclusione

Conclusione

Riassumendo...

- I problemi di larga scala ("Big Data") non possono essere risolti con l'utilizzo di un unico computer
  - Tra i problemi di larga scala includiamo lo studio del DNA, della società su Internet, del meteo, dell'economia, ecc.
- Il calcolo distribuito consiste nel coordinare molti computer che lavorano assieme su un unico problema
- Il calcolo distribuito è necessario per risolvere i problemi di larga scala
- Usufruiamo dei risultati di calcoli di larga scala quotidianamente
- Al CRS4 lavoriamo anche su problemi di calcolo a larga scala con tecniche di calcolo distribuito, come l'analisi del DNA

Conclusione

Ringrazio. . .

- Massimo Mancini
- Greca Meloni
- Gli organizzatori della manifestazione
- Voi che mi ascoltate!

Grazie!

Conclusione

Ringrazio. . .

- Massimo Mancini
- Greca Meloni
- Gli organizzatori della manifestazione
- Voi che mi ascoltate!

Grazie!

Domande?