The Data Deluge

The Data Deluge

The DNA Deluge

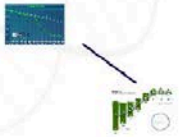
A Parallel Computing Approach

Brendan Lawlor

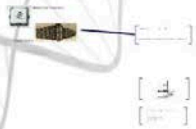
simplicity

ngilico

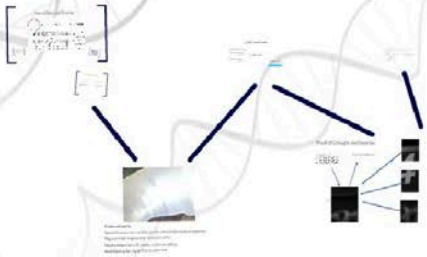
1. What is the deluge?



2. What is parallel computing?



3. How to harness parallel power?



The DNA Deluge

A Parallel Computing Approach

Brendan Lawlor

simplicity

nsilico

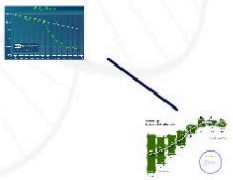
The Data Deluge

The DNA Deluge

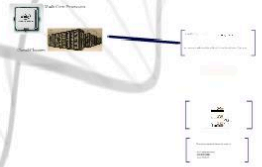
A Parallel Computing Approach

Brendan Lawlor

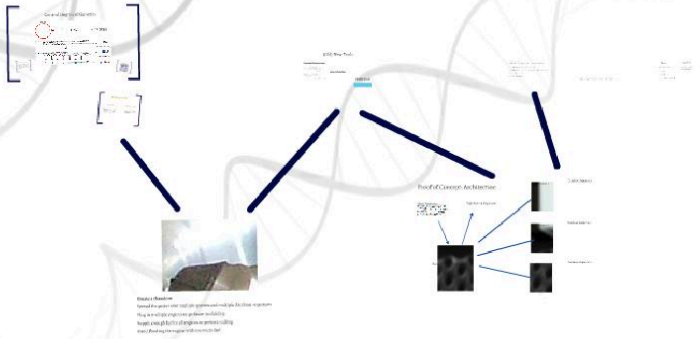
1. What is the deluge?



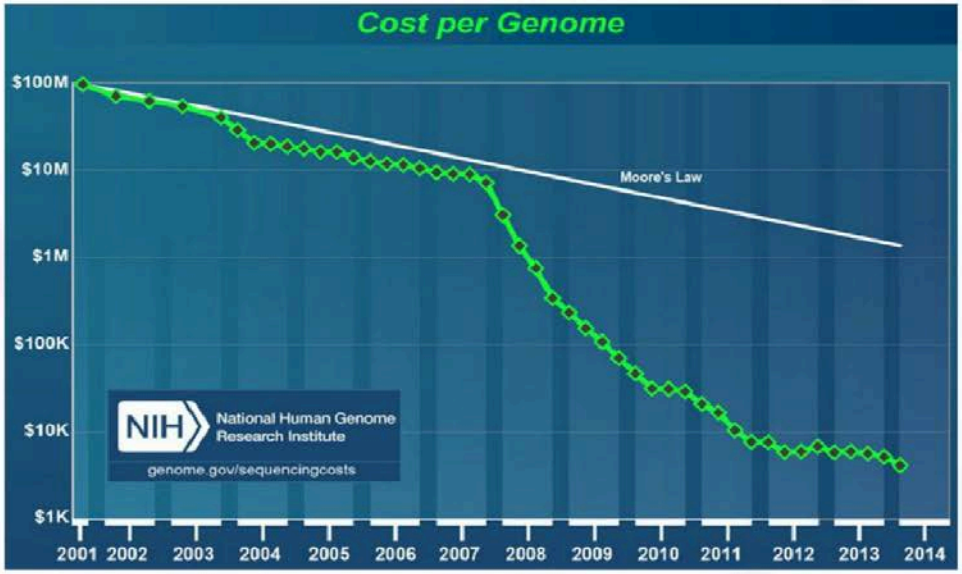
2. What is parallel computing?



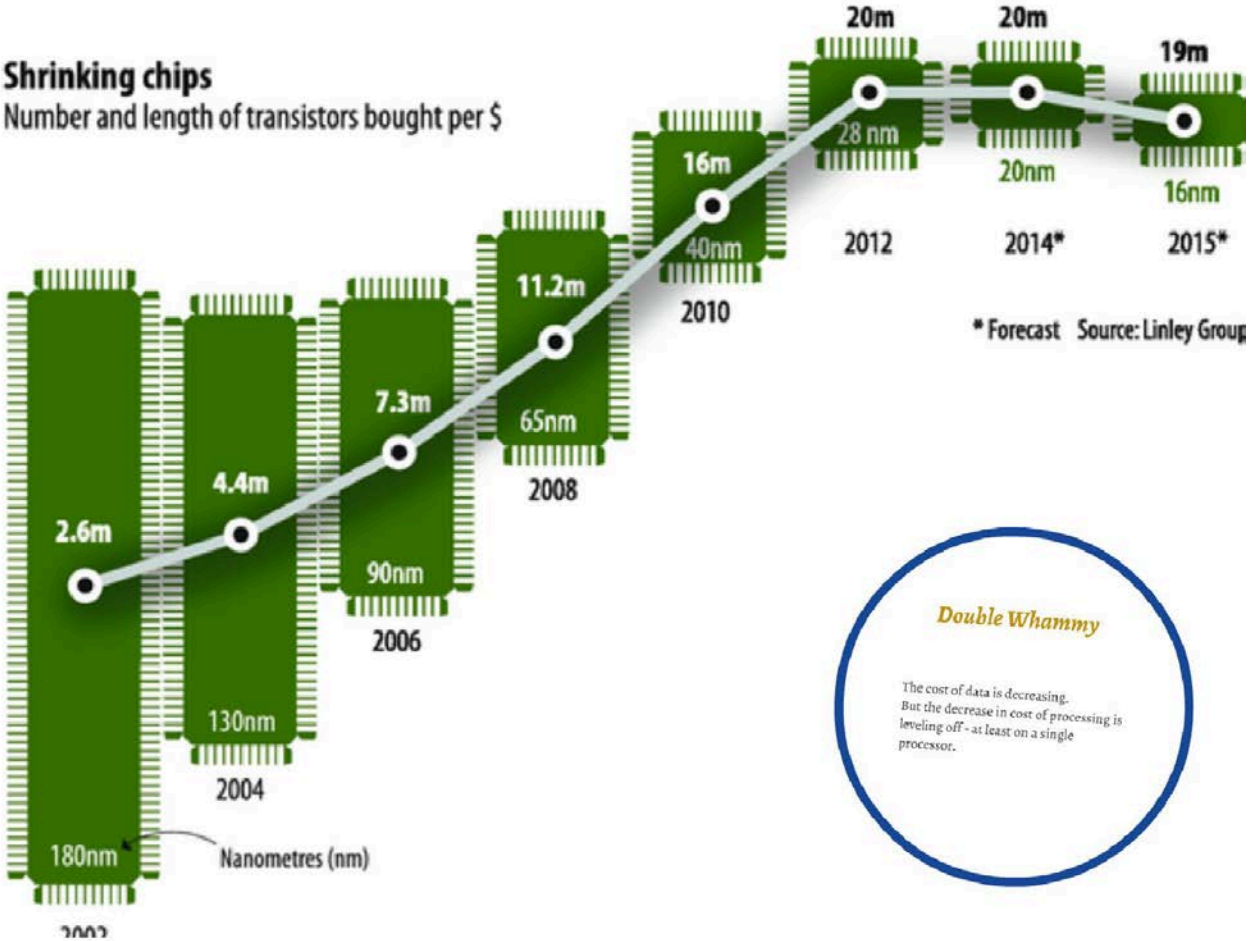
3. How to harness parallel power?



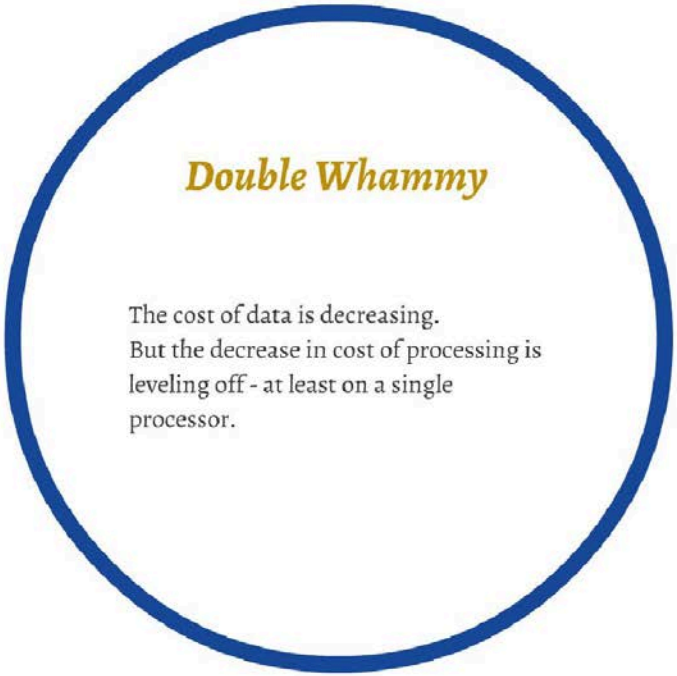
The Data Deluge



The Data Deluge

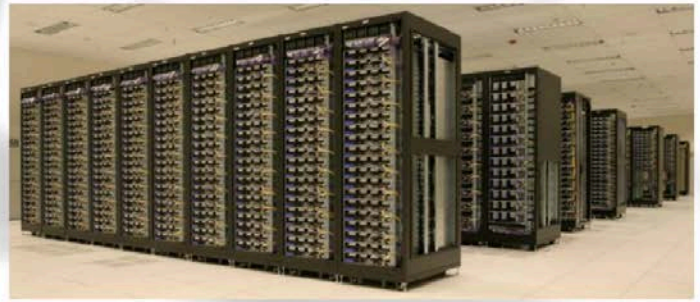


The Data Deluge



The Data Deluge

Multi Core Processors

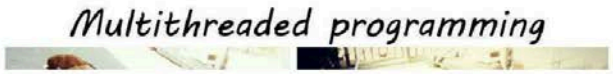


Cloud/Clusters

The Data Deluge

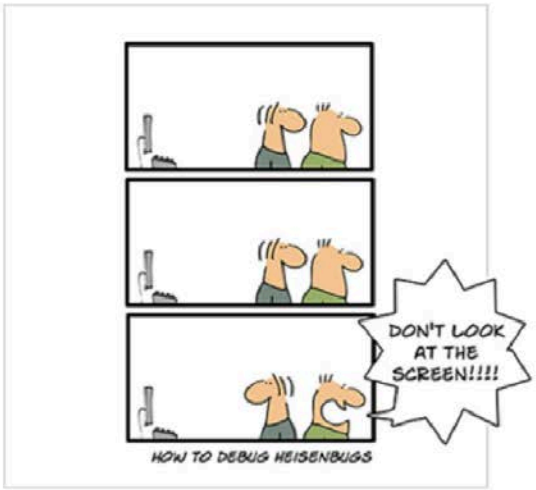
Amdahl's Law: $$S(n) = \frac{T(1)}{T(n)} = \frac{T(1)}{T(1)\left(B + \frac{1}{n}(1-B)\right)} = \frac{1}{B + \frac{1}{n}(1-B)}$$

or... you can't make something faster than its slowest (= serialized) part.



The Data Deluge

The Data Deluge



The Data Deluge

We are going to need tools that are fit for purpose.

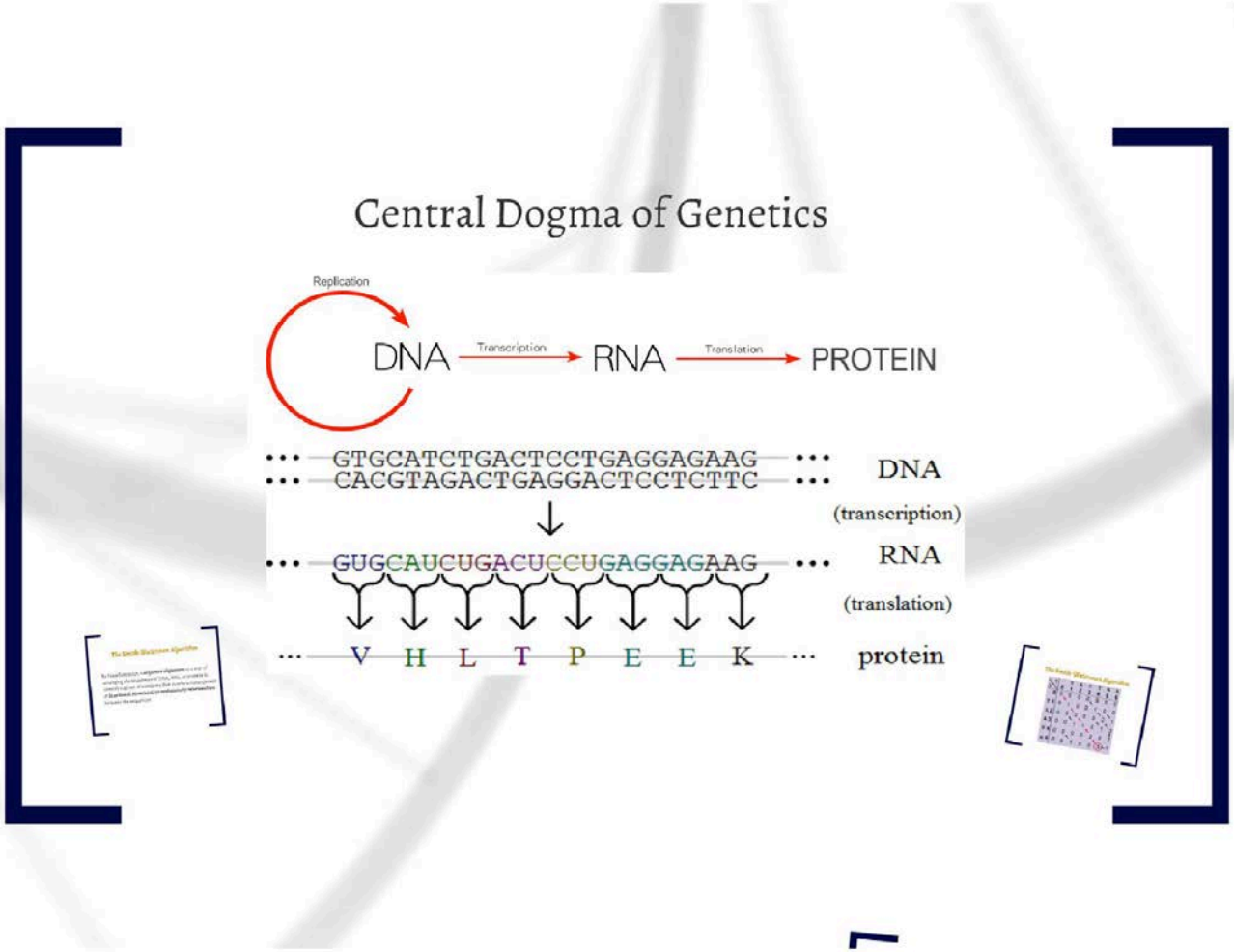
Functional Programming
Actor Architecture
Reactive Streams

The Data Deluge

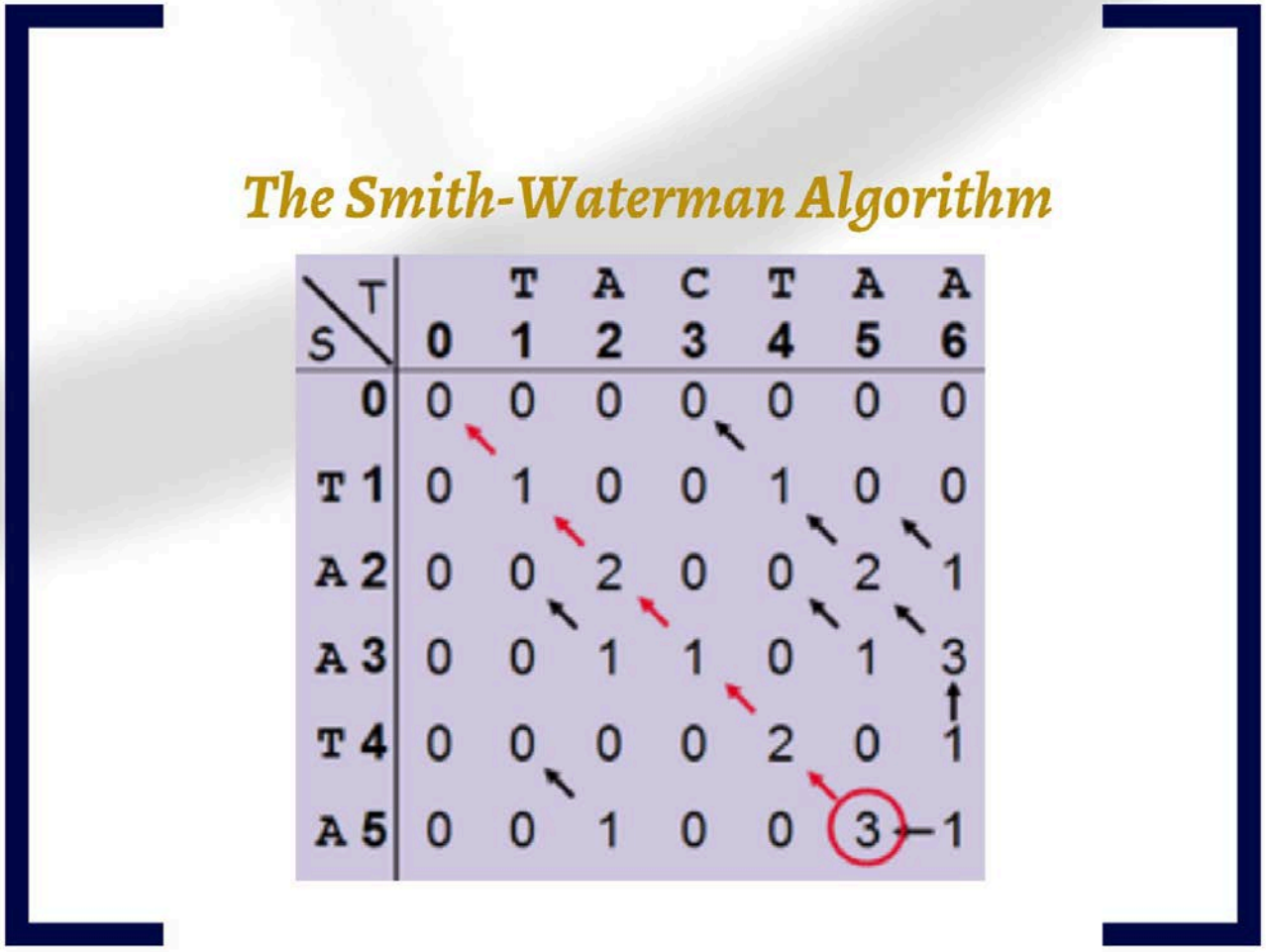
The Smith-Waterman Algorithm

In bioinformatics, a **sequence alignment** is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of **functional, structural, or evolutionary relationships** between the sequences.

The Data Deluge



The Data Deluge



The Data Deluge

The Data Deluge

Data-Dependent

The SW algorithm is an example of this kind: **some cells in the matrix require the results of others**. It uses intricate loops and intimate knowledge of the hardware to drive as much data as possible through a CPU in a single clock-cycle. It achieves SIMD using Intrinsics.

The Data Deluge

"Embarrassingly Parallel"

Running many query sequences against many database sequences is an example of this kind. **Each run of the S-W algorithm is independent of those that go before and after.**

The Data Deluge

The Data Deluge

Create a chassis to:

Spread the power over multiple queries and multiple database sequences

Plug in multiple engines to go faster (*scalability*)

Supply enough fuel to all engines to prevent stalling

Avoid flooding the engine with too much fuel

The Data Deluge

(Old) New Tools

Functional Programming

"The essence of functional programming is to concentrate on **transformations** of **immutable values** rather than stepwise modifications of mutable values."

- Martin Odersky, inventor of Scala programming language

In other words, with FP we can write code in an intrinsically parallel manner and let the compiler and run-time schedule on the parallel infrastructure.

- Paraphrasing Robert Harper, Professor of Computer Science at Carnegie Mellon University

Actor Architecture

Reactive Streams

The Data Deluge

Functional Programming

"The essence of functional programming is to concentrate on **transformations** of **immutable values** rather than stepwise modifications of mutable values."

- Martin Odersky, inventor of Scala programming language

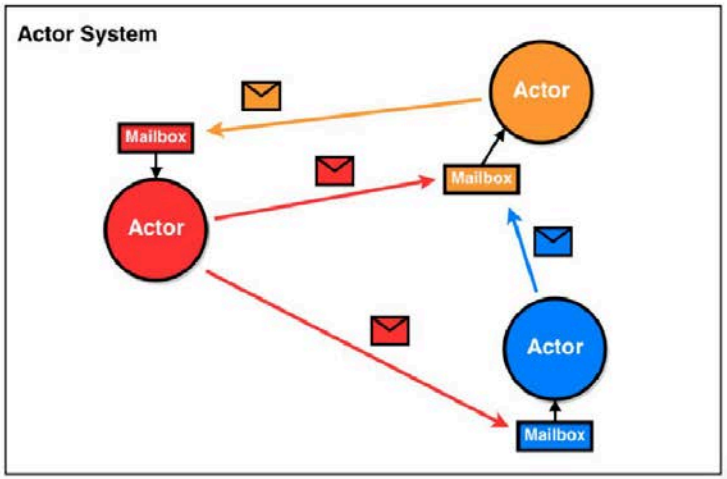
In other words, with FP we can write code in an intrinsically parallel manner and let the compiler and run-time schedule on the parallel infrastructure.

- Paraphrasing Robert Harper, Professor of Computer Science at Carnegie Mellon University

The Data Deluge

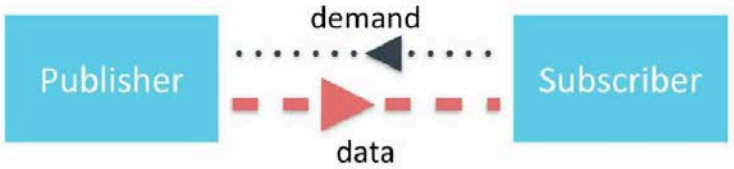
Actor Architecture

er Science

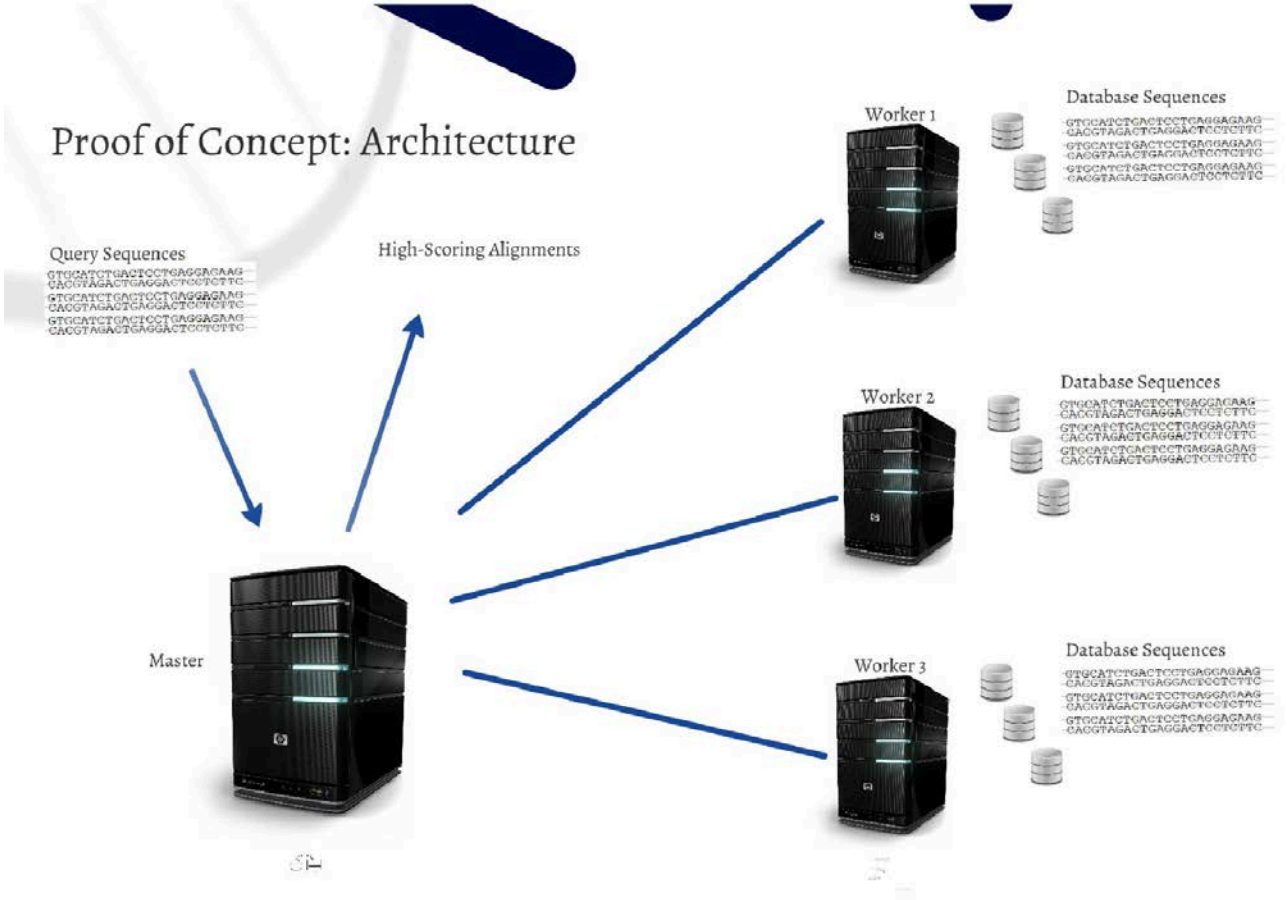


The Data Deluge

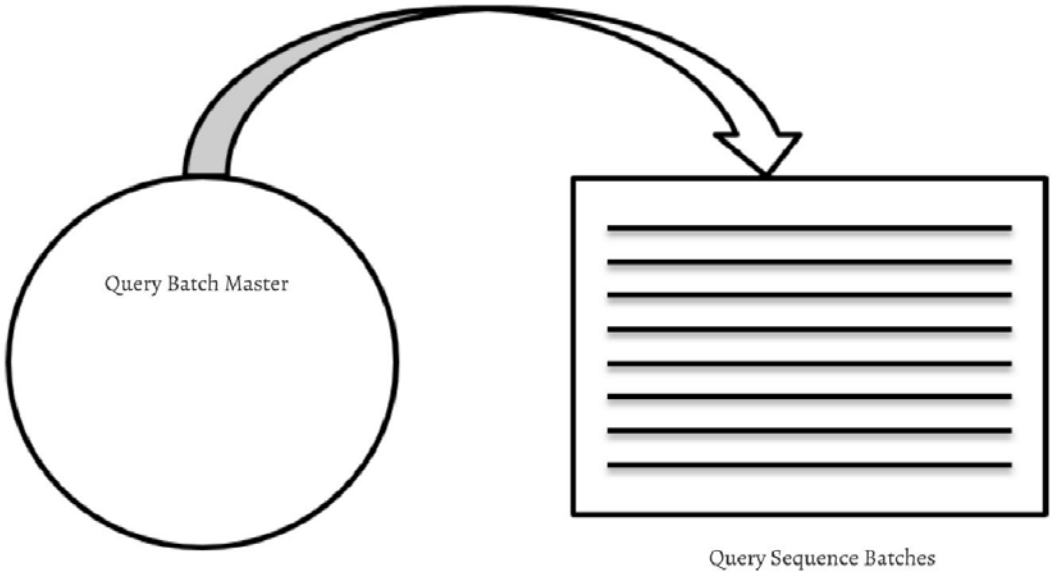
Reactive Streams



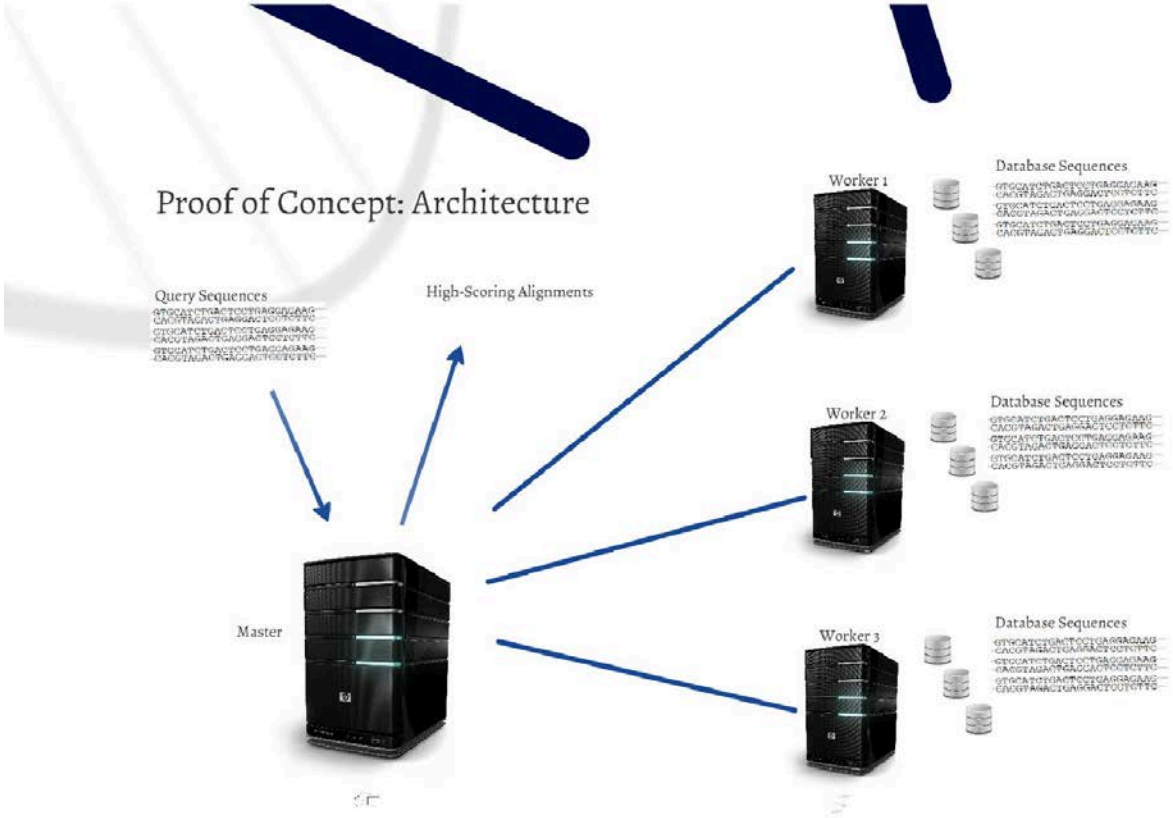
The Data Deluge



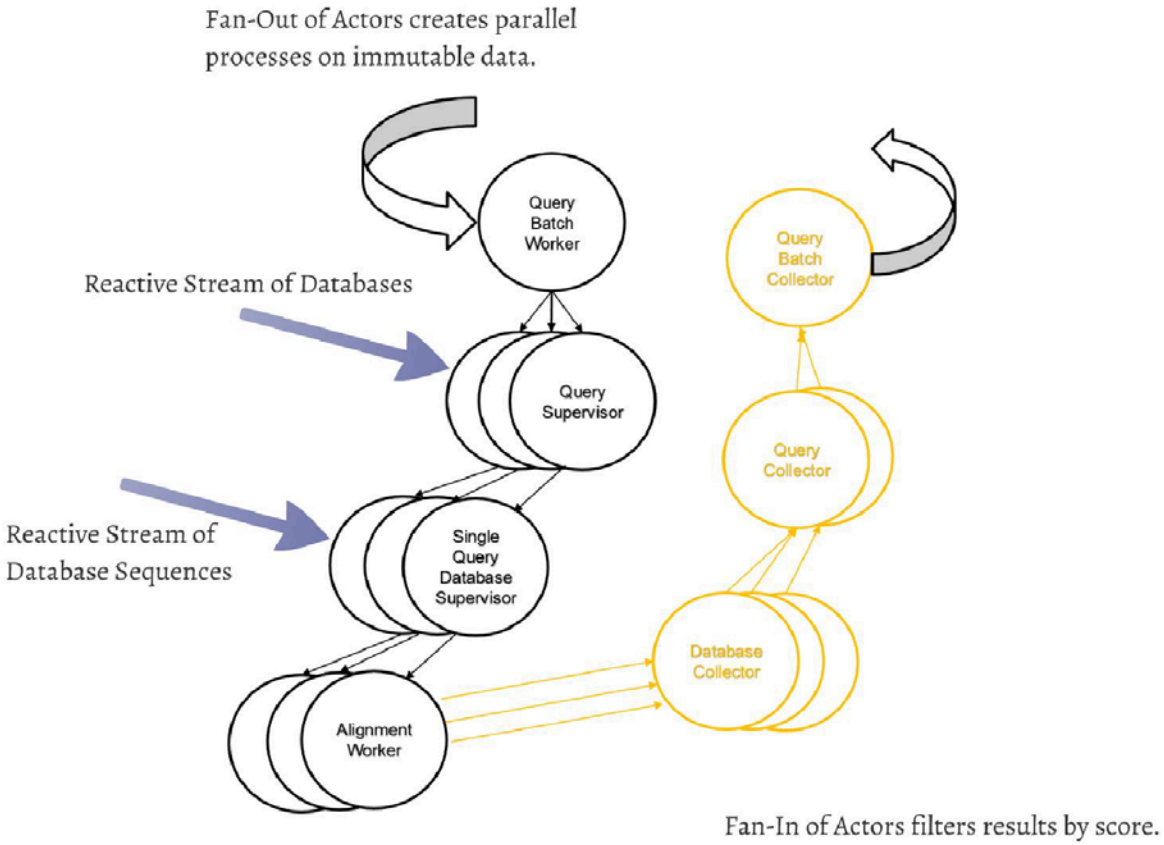
The Data Deluge



The Data Deluge



The Data Deluge



The Data Deluge

Proof of Concept: Results

Results on a single Node: Linux box with an i7-4770 3.4GHz processor(quad-core)

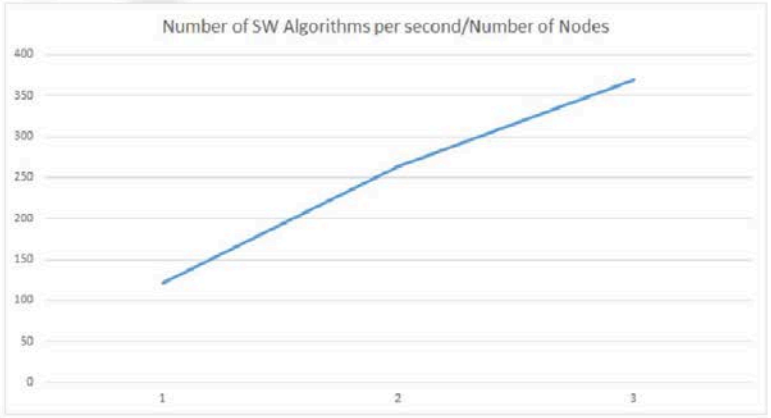
200 S-W invocations (or about 1.8 billion "cells") per second

Using a single-core baseline, this indicates ~ 15% over overhead (I/O & message passing).

~ 1000 lines of (readable!) Scala code.

What about multiple Linux boxes?

The Data Deluge



Multiple cheap computers, with the same cheap software, multiplies the performance.

The Data Deluge

What's next?

Optimization of Proof of Concept

Test on larger scale

Publish results

Consider a fully stream-centric solution.

(And So What!?)

Just one of many examples:
Clinical Microbiology

The Data Deluge

The DNA Deluge

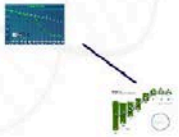
A Parallel Computing Approach

Brendan Lawlor

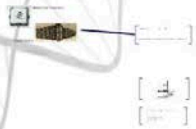
simplicity

ngilico

1. What is the deluge?



2. What is parallel computing?



3. How to harness parallel power?