Exploring reproducibility and data parallelism in the egg project

R244 mini-project

Edmund Goodman

December 4, 2024

Three goals

1. Replicate paper[1] results
2. Profile compilation process
3. Explore opportunities for data parallelism

[1] Willsey et al., "egg".

Replicate paper results

What is result replication?

- “The main results of the paper have been obtained in a subsequent study by a person or team other than the authors, using, in part, artifacts provided by the author.”[a]

Why should we bother?

- Important aspect of the scientific process!
- Facilitates next two stages of the mini-project...

Figure 1: The current ACM artifact badges for the paper.

Figure 2: The goal, an ACM “Results Replicated” badge.

[a] *Artifact Review and Badging – Version 1.0 (not current).*

Profiling compilation process

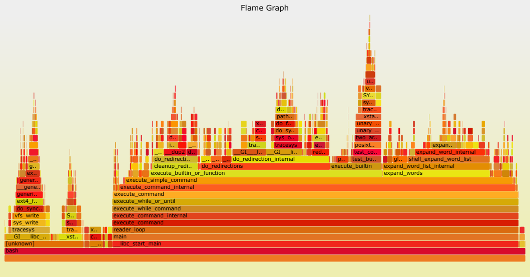

Figure 3: A flamegraph[a] of Rust code.

- What kernels dominate compilation time?
- How do these differ across workloads?
- How can these guide ideas for optimisations?

[a] Gregg, *brendangregg/FlameGraph*.

Fearless Parallelism in Rust

```cpp
double dot_product (
    int n, double* x, double* y
) {
    double result = 0.0;
#pragma omp parallel for reduction(+:result)
    for (int i = 0; i < n; i++) {
        result += x[i] * y[i];
    }
    return result;
}
```

(a) C++ OpenMP implementation

```rust
use rayon::prelude::*;

pub fn dot_product(
    lhs: &[f64], rhs: &[f64]
) -> f64 {
    lhs.par_iter()
        .zip(rhs.par_iter())
        .map((x, y) x * y)
        .sum()
}
```

(b) Rust `rayon` implementation

Figure 4: Parallel implementations of the dot product kernel

Explore opportunities for data parallelism

- `rayon`[2] to naïvely leverage simple data parallelism
  - Does this have any impact/give any speedup?
  - To what extent does it change viability?
- Further opportunities for data parallelism?
  - Any bottlenecks previously identified which could be reduced?
  - Opportunities for algorithmic modifications?

- Does this help justify scalability or just “kick the can down the road”?

[2] *rayon-rs/rayon*.

Exploring reproducibility and data parallelism in the egg project

R244 mini-project

Edmund Goodman

December 4, 2024

References i

[1] *Artifact Review and Badging – Version 1.0 (not current).* en. URL: `https://www.acm.org/publications/policies/artifact-review-badging` (visited on 12/04/2024).

[2] Brendan Gregg. *brendangregg/FlameGraph.* original-date: 2011-12-16T02:20:53Z. Dec. 2024. URL: `https://github.com/brendangregg/FlameGraph` (visited on 12/04/2024).

[3] *rayon-rs/rayon.* original-date: 2014-10-02T15:38:05Z. Dec. 2024. URL: `https://github.com/rayon-rs/rayon` (visited on 12/04/2024).

[4] Max Willsey, Chandrakana Nandi, Yisu Remy Wang, Oliver Flatt, Zachary Tatlock, and Pavel Panchekha. **“egg: Fast and extensible equality saturation”.** In: *Artifact for ”Fast and Extensible Equality Saturation”* 5.POPL (Jan. 2021), 23:1–23:29. DOI: `10.1145/3434304`. URL: `https://dl.acm.org/doi/10.1145/3434304` (visited on 11/22/2024).