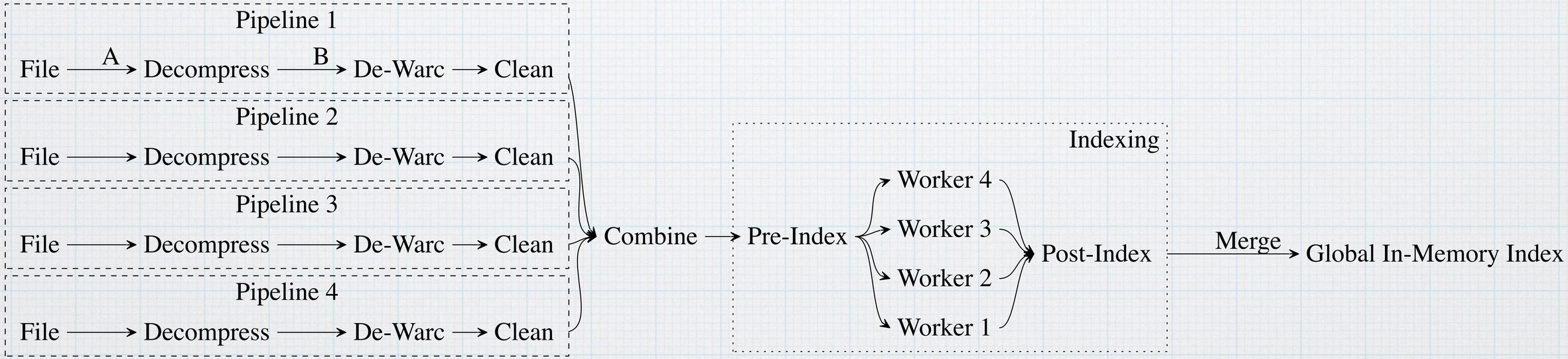
Improving Throughput of a Pipeline Model Indexer

Matt Crane, Andrew Trotman, & David Eyers

Introduction

* Indexing needs I/O and CPU
* Do them simultaneously speeds up indexing
* Pipeline-based & Map/Reduce alternatives

ATIRE's Pipeline



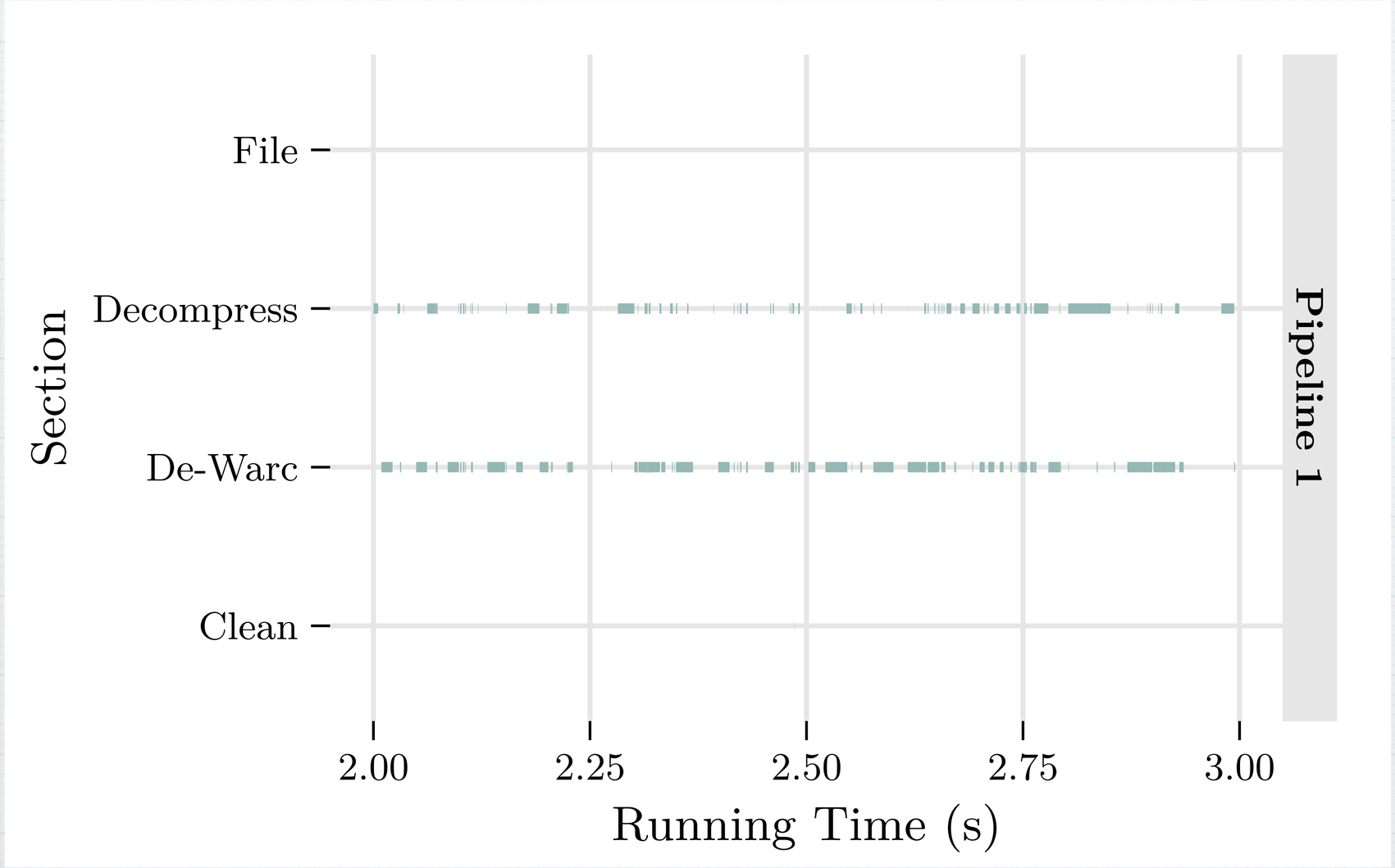
Collections

Collection	Documents	Size
.GOV2	25M	426GB
ClueWeb09 Cat. B	50M	1.5TB

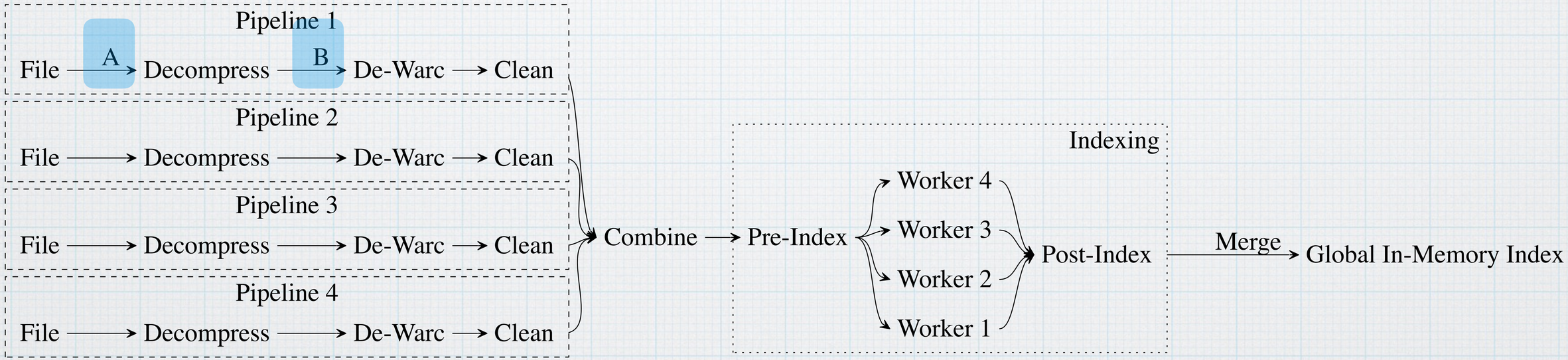
Baseline

System	.GOV2 Indexing Time (min)
ATIRE*	46
Lucene	85
MG4J	85
Galago	392
Indri	460
Terrier	484

Waiting


Section
File
Decompress
De-Warc
Clean
Pipeline 1
2.00
2.25
2.50
2.75
3.00
Running Time (s)

ATIRE's Pipeline

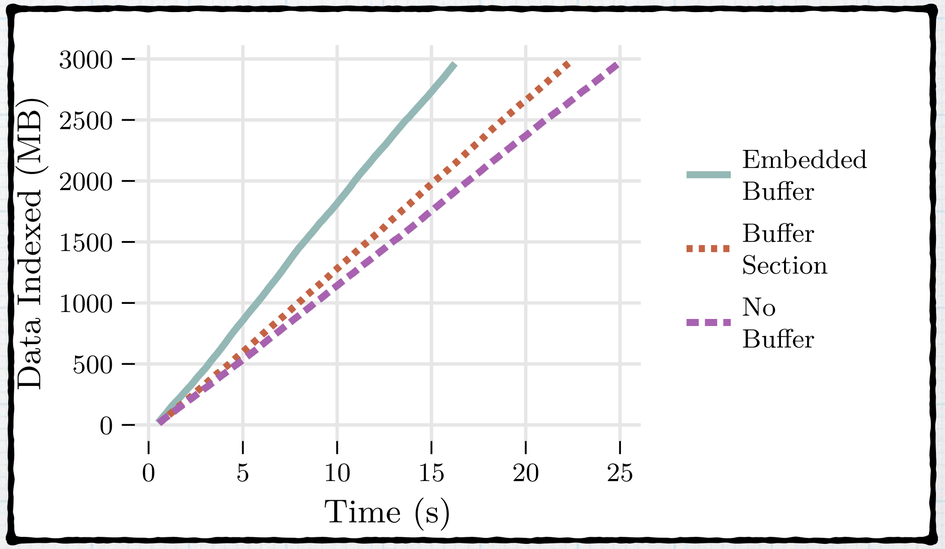


Buffer Position & Type

* As well as position there's type to consider:
  * Single
  * Double
  * Triple (can be emulated by two doubles)

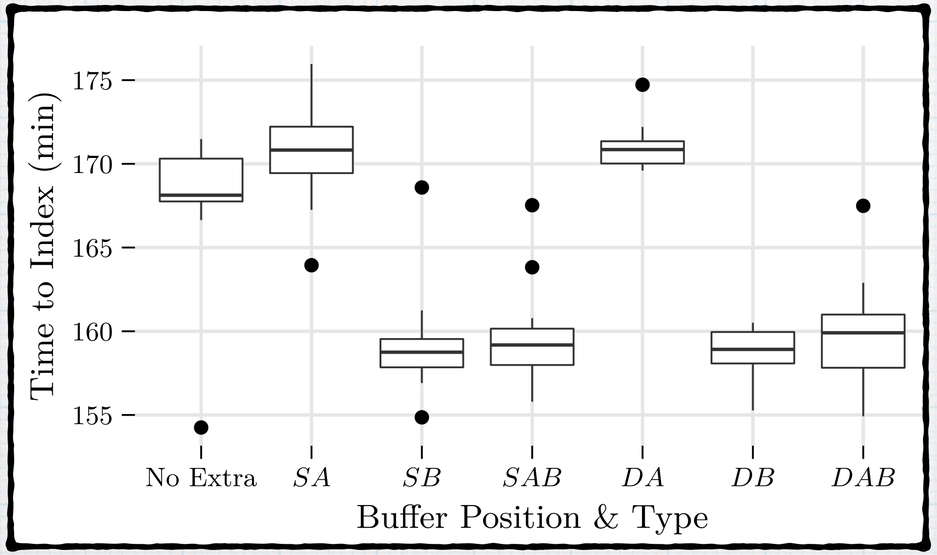
Buffering @ A

* 16MB single buffer between decompression and disk (A)
* Or: a 1MB buffer embedded within decompression section
* Or: none

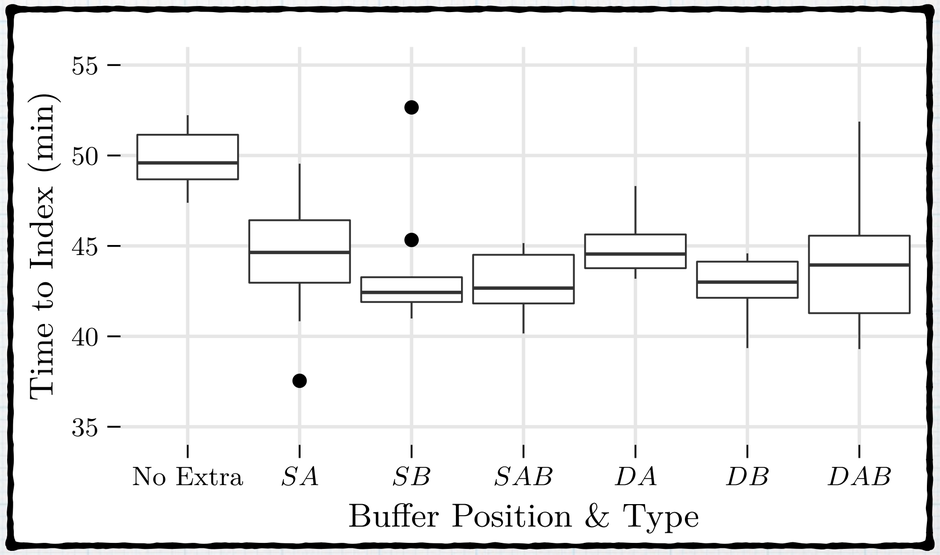


Buffer Position & Type

CW09B

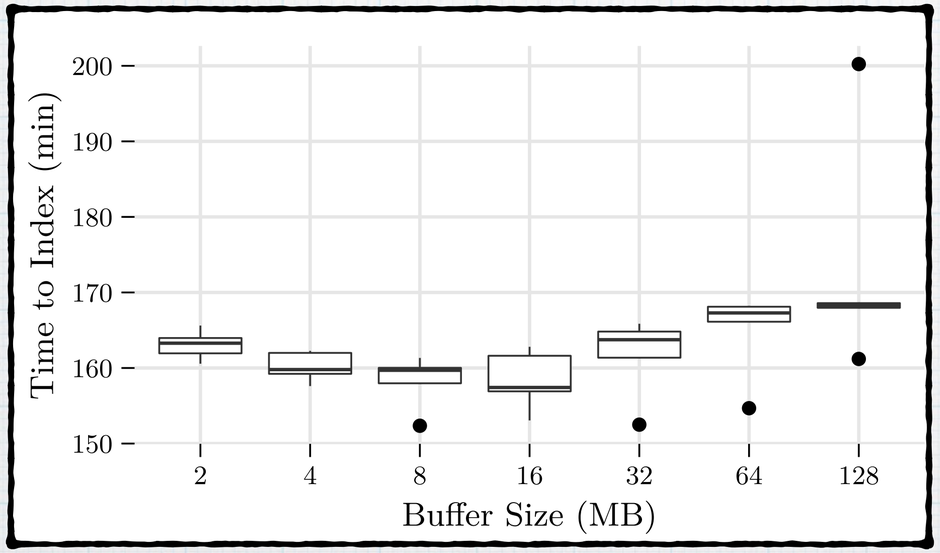


.GOV2

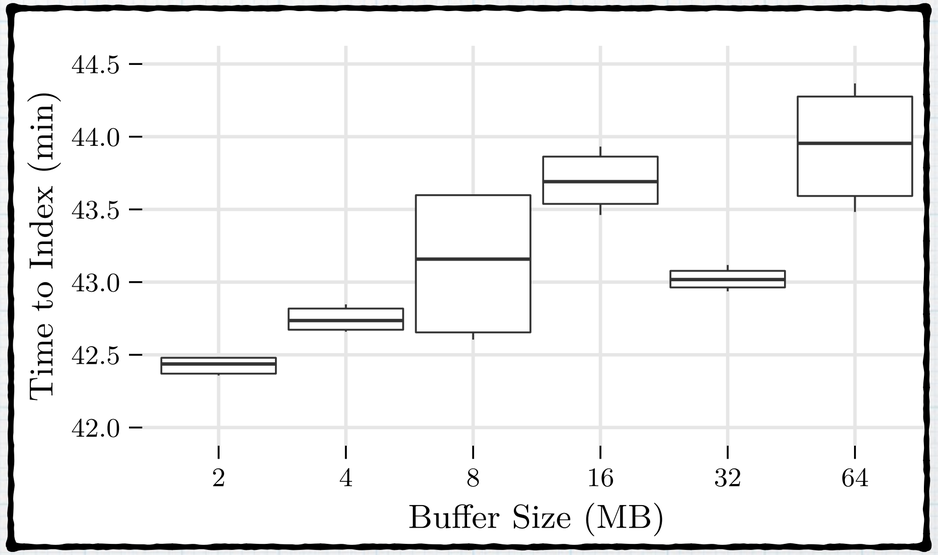


Buffer Size

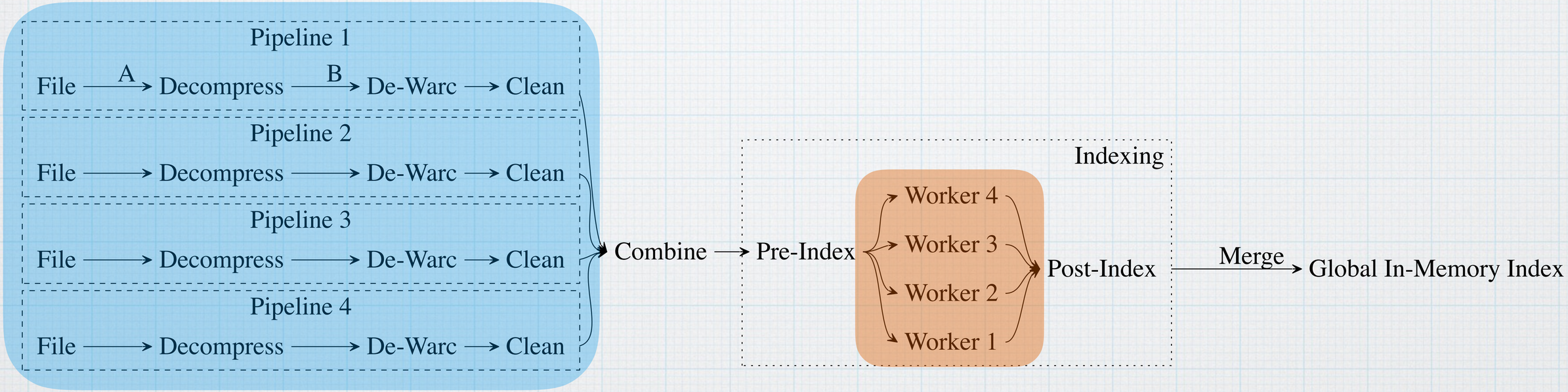
CW09B



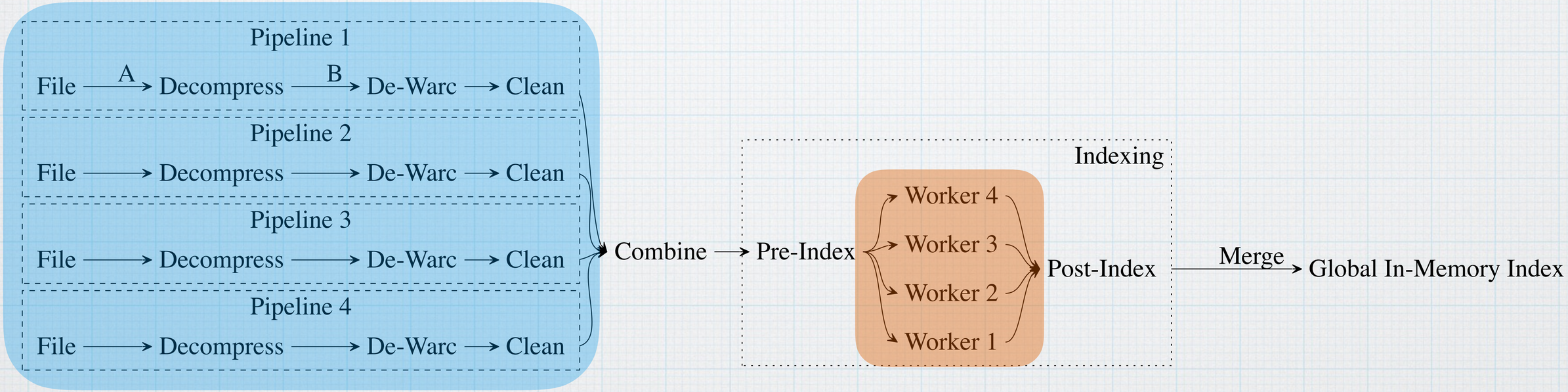
.GOV2



ATIRE's Pipeline

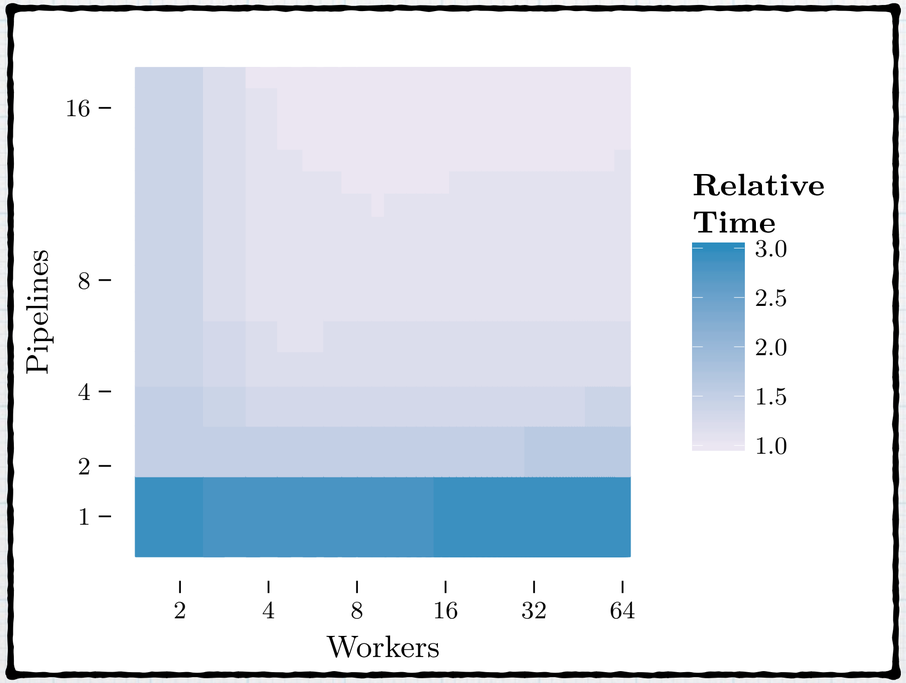


ATIRE's Pipeline

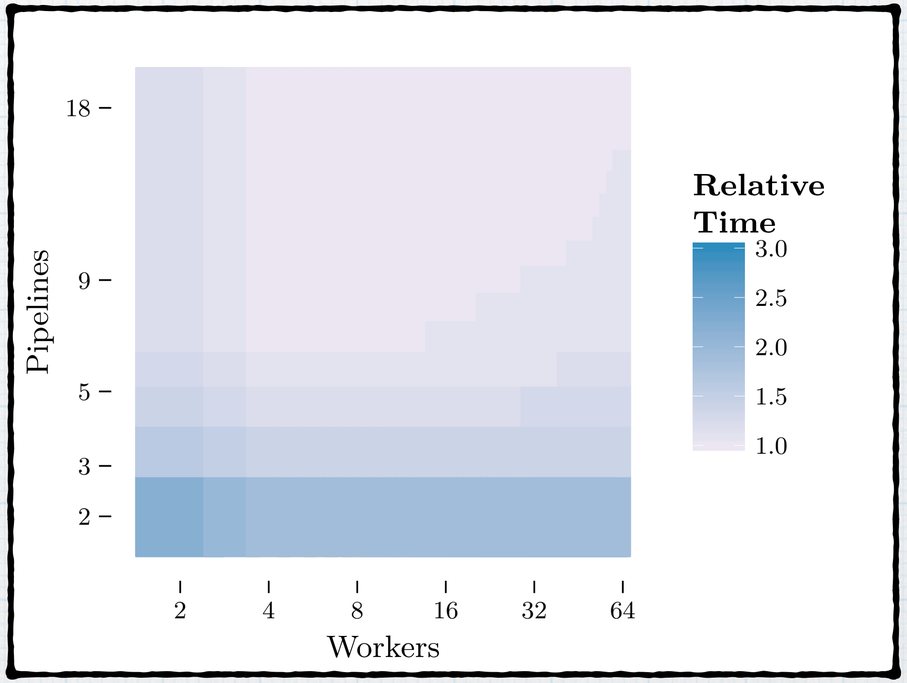


Workers & Pipelines

CW09B



.GOV2



Future Work

* Work stealing pipelines
* Distributed indexing

Questions?
// Comments