

# Distributed Text Index Construction

Patrick Dinklage, Johannes Fischer, Florian Kurpicz

Department of Computer Science

Chair 1: Algorithm Engineering

Algorithmic Foundations and Education in Computer Science

# Distributed Text Index Construction

Patrick Dinklage, Johannes Fischer, Florian Kurpicz

Department of Computer Science

Chair 1: Algorithm Engineering

Algorithmic Foundations and Education in Computer Science

# Distributed Text Index Construction

Patrick Dinklage, Johannes Fischer, Florian Kurpicz

Department of Computer Science

Chair 1: Algorithm Engineering

Algorithmic Foundations and Education in Computer Science

# Distributed Text Index Construction

Patrick Dinklage, Johannes Fischer, Florian Kurpicz

Department of Computer Science

Chair 1: Algorithm Engineering

Algorithmic Foundations and Education in Computer Science

# DISTRIBUTED (COMPUTING)

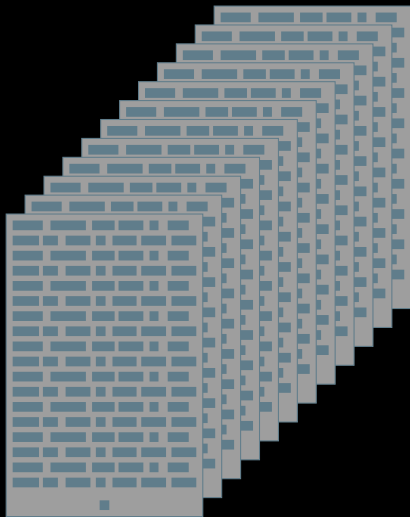


[by courtesy of ido.tu-dortmund.de]

## In Practice

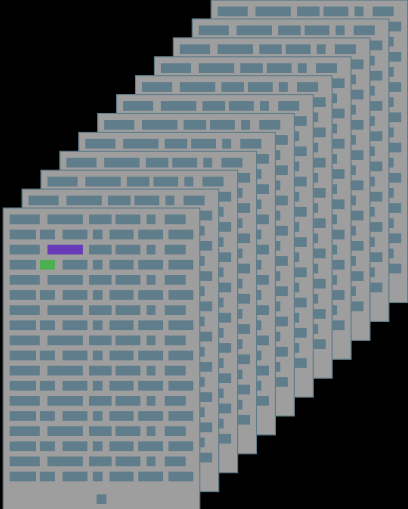
- ▶ Synchronization is expensive?
- ▶ Load-balancing is important?
- ▶ Communication & memory efficiency?

# TEXT INDEX

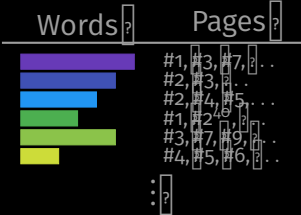


~ 100 GiB ?

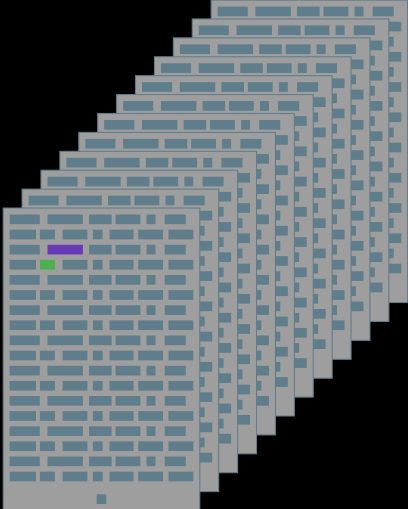
# TEXT INDEX



~ 100 GiB



# TEXT INDEX



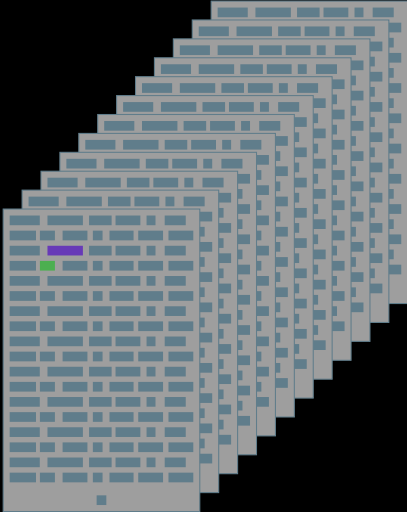
~ 100 GiB

Words	Pages
██████████	#1, #3, #7, ? ..
██████████	#2, #3, ? ..
██████████	#2, #4, #5 ..
██████████	#1, #2, ? ..
██████████	#3, #7, #9, ? ..
██████████	#4, #5, #6, ? ..
...	...
?	?





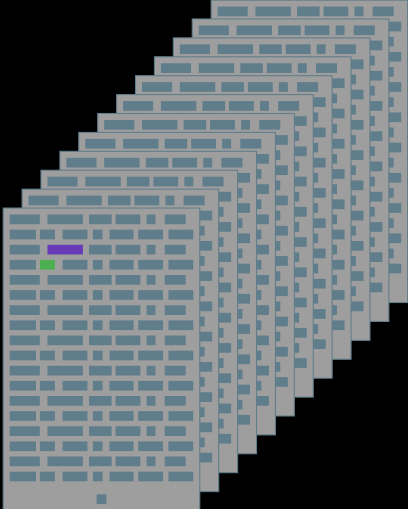
# TEXT INDEX



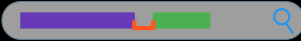
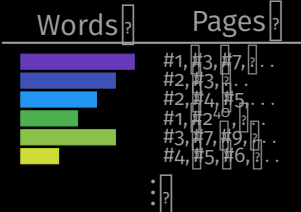
~100 GiB



# TEXT INDEX



~100 GiB

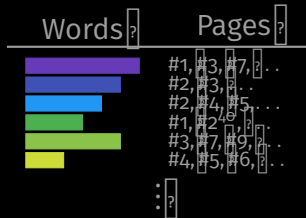


... but what if?

# TEXT INDEX

GAATGCCAGTCAGCATTAAGGCCCGG  
GGAGAGCTCAGGGCAGGTCACGTTGGG  
AACTCGCATAGTGAGGGTTATCGTCAG  
ACATGTTGTTGGGCTCTTCACTCGCA  
CCGACACGAACCTCAGTTAGTTTCGTA  
CCTACATCCTACCAGAGGTCGCAGGTC  
TGTGCCCCGGTGGTGAGAAGGAGAGCT  
TGGGATTTTCGTATTTGCAGATGGCCT  
CTCGTCAGTACTTTCAGAATAACCTCA  
CATGGCCTGCACGGCAAATGGCTCTT  
GACGCTTATAATGGACTTCGACATGTT  
AACTCGCATAGTGAGGGTTATCGTCAG  
ACATGTTGTTGGGCTCTTCACTCGCA  
CCGACACGAACCTCAGTTAGTTTCGTA  
TGTGCCCCGGTGGTGAGAAGGAGAGCT

~100 GiB

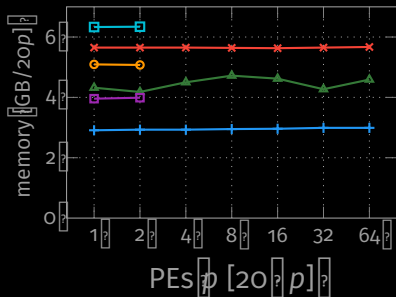
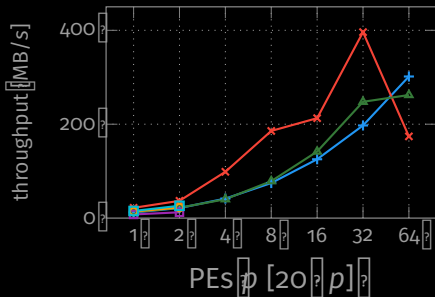


... but what if?



# (EVALUATING THE) CONSTRUCTION

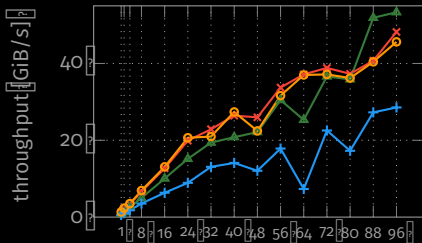
- ▶ Evaluated suffix array construction on the LiD03 cluster
- ▶ Weak scaling experiments (only DNA) with 90MB per PE



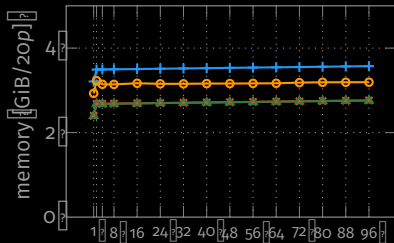
+ dDivSufSort   
 ▲ dPD   
 × PSAC [F&A<sub>15</sub>]   
 □ / ○ / □ DC3/7/13 [B<sub>18</sub>]

# (EVALUATING THE) CONSTRUCTION

- ▶ Evaluated wavelet tree construction on the LIGO3 cluster
- ▶ Weak scaling experiments (only DNA) with 53 MB per PE



PEs [20 p]

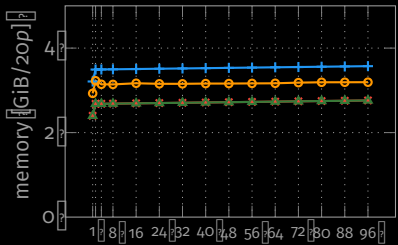
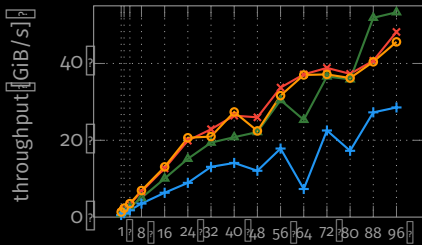


PEs [20 p]

—×— bsort —▲— dd —+— dplit —○— dynbsort

# (EVALUATING THE) CONSTRUCTION

- ▶ Evaluated wavelet tree construction on the LIGO3 cluster
- ▶ Weak scaling experiments (only DNA) with 53 MB per PE



PEs [p [20 p]]

—×— bsort —▲— dd —+— dplit —○— dynbsort

Thank You