Distributed Text Index Construction

Patrick Dinklage Johannes Fischer *Florian Kurpicz*

Department of Computer Science
Chair 11: Algorithm Engineering
Algorithmic Foundations and Education in Computer Science

Distributed Text Index Construction

Patrick Dinklage Johannes Fischer *Florian Kurpicz*

Department of Computer Science
Chair 11: Algorithm Engineering
Algorithmic Foundations and Education in Computer Science

Distributed Text Index Construction

Patrick Dinklage Johannes Fischer *Florian Kurpicz*

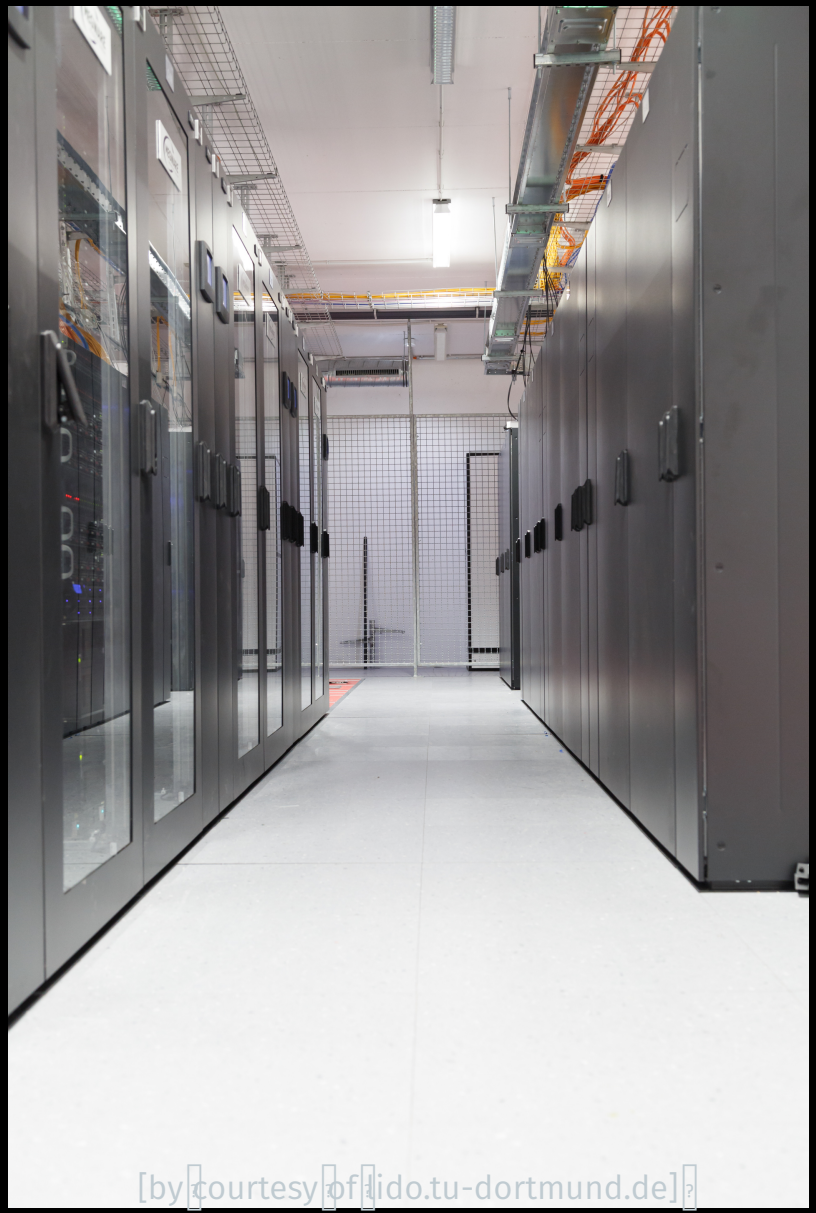
Department of Computer Science
Chair 11: Algorithm Engineering
Algorithmic Foundations and Education in Computer Science

Distributed Text Index Construction

Patrick Dinklage Johannes Fischer *Florian Kurpicz*

Department of Computer Science
Chair 11: Algorithm Engineering
Algorithmic Foundations and Education in Computer Science

DISTRIBUTED (COMPUTING)

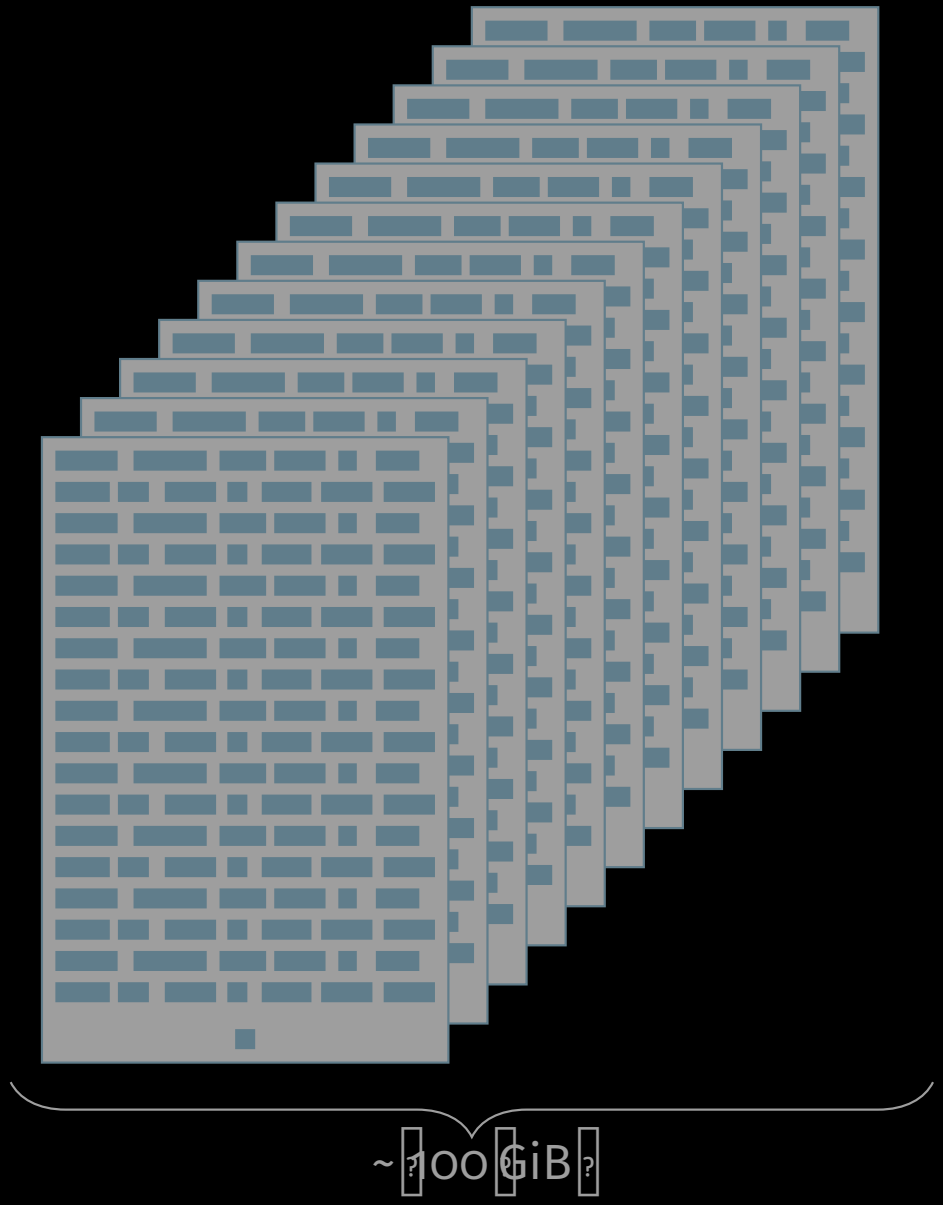


[by courtesy of ido.tu-dortmund.de]

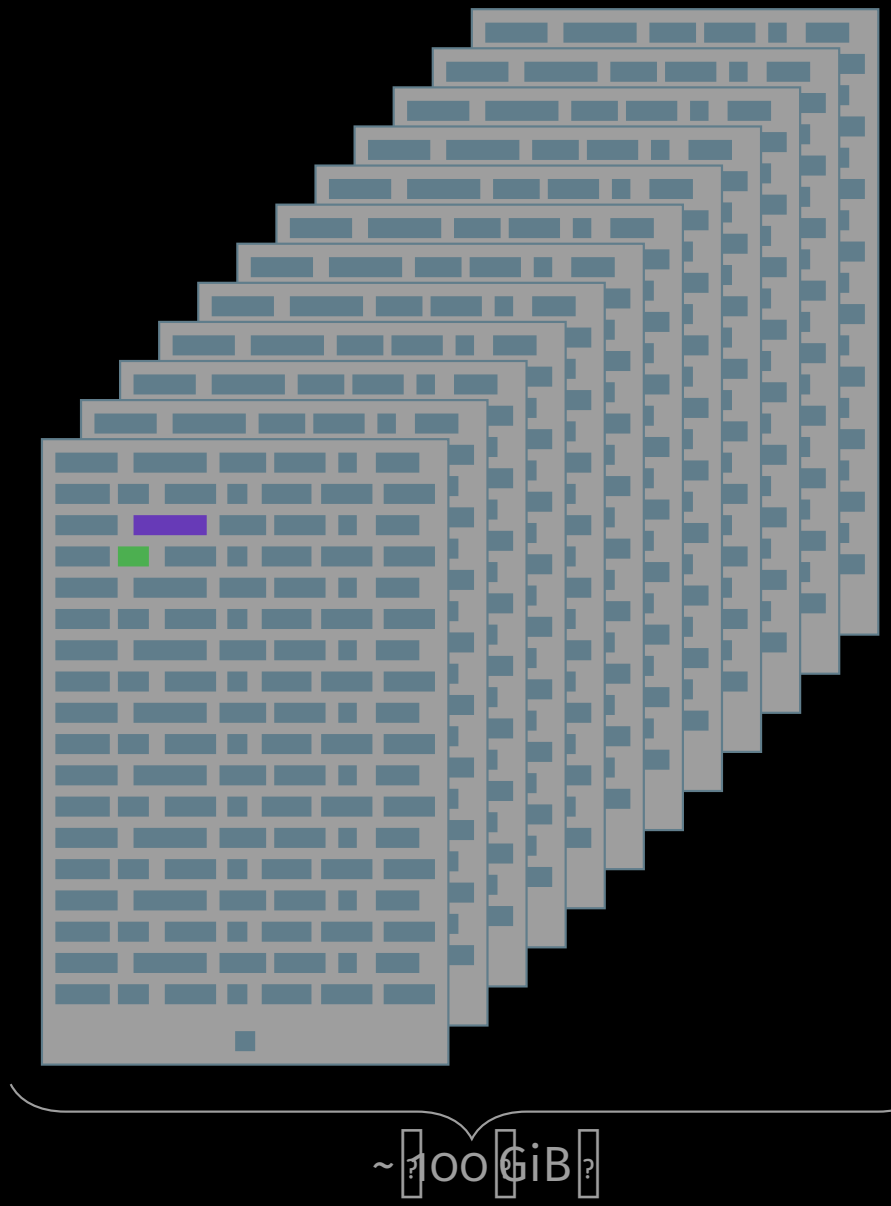
In Practice

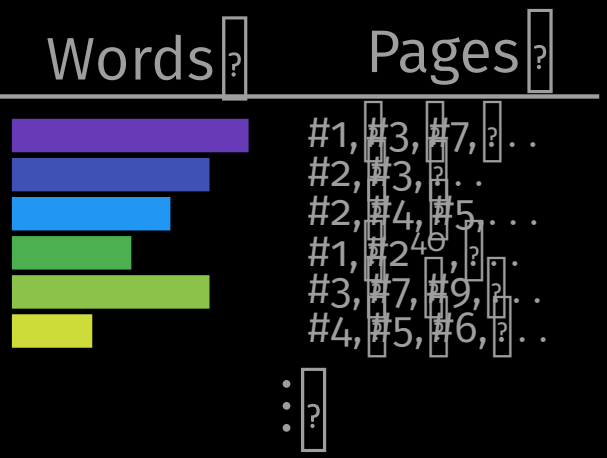
- Synchronization is expensive
- Load-balancing is important
- Communication & memory efficiency

TEXT INDEX

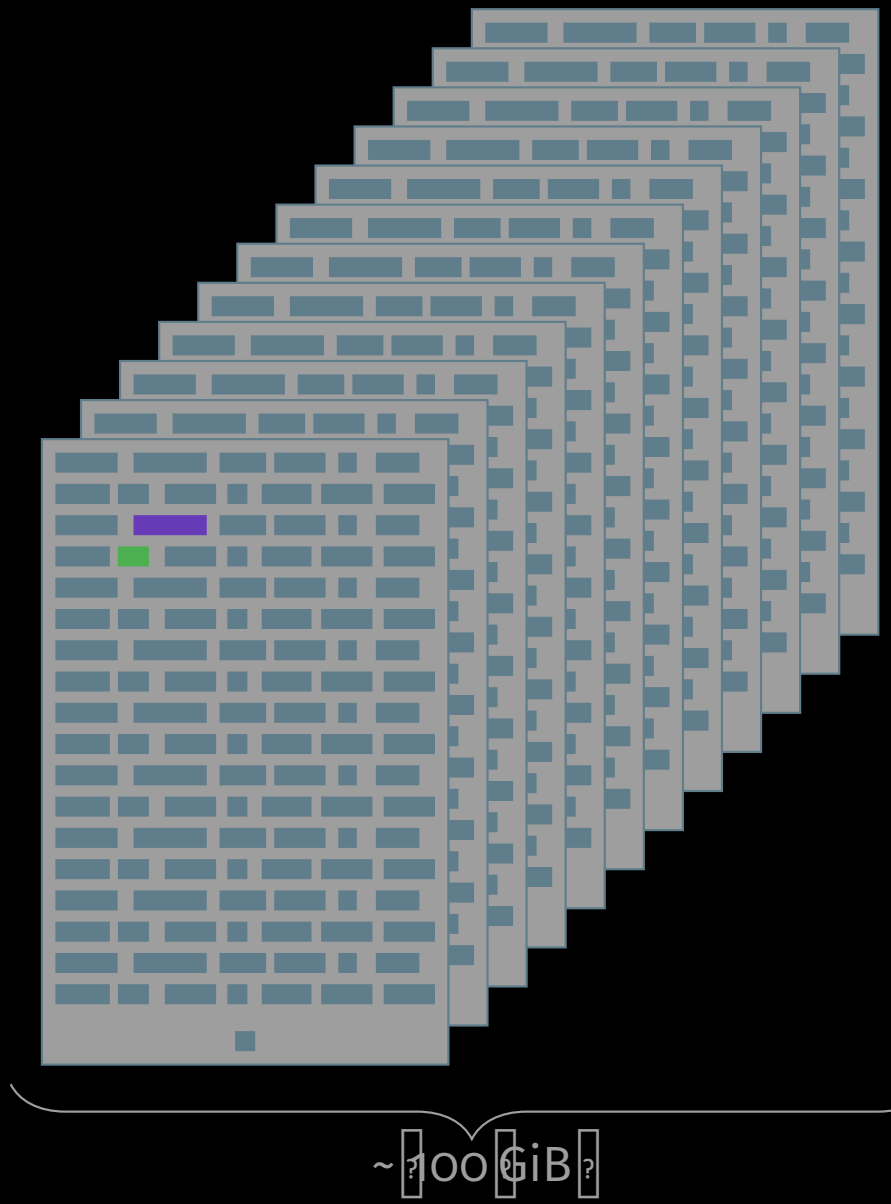


TEXT INDEX

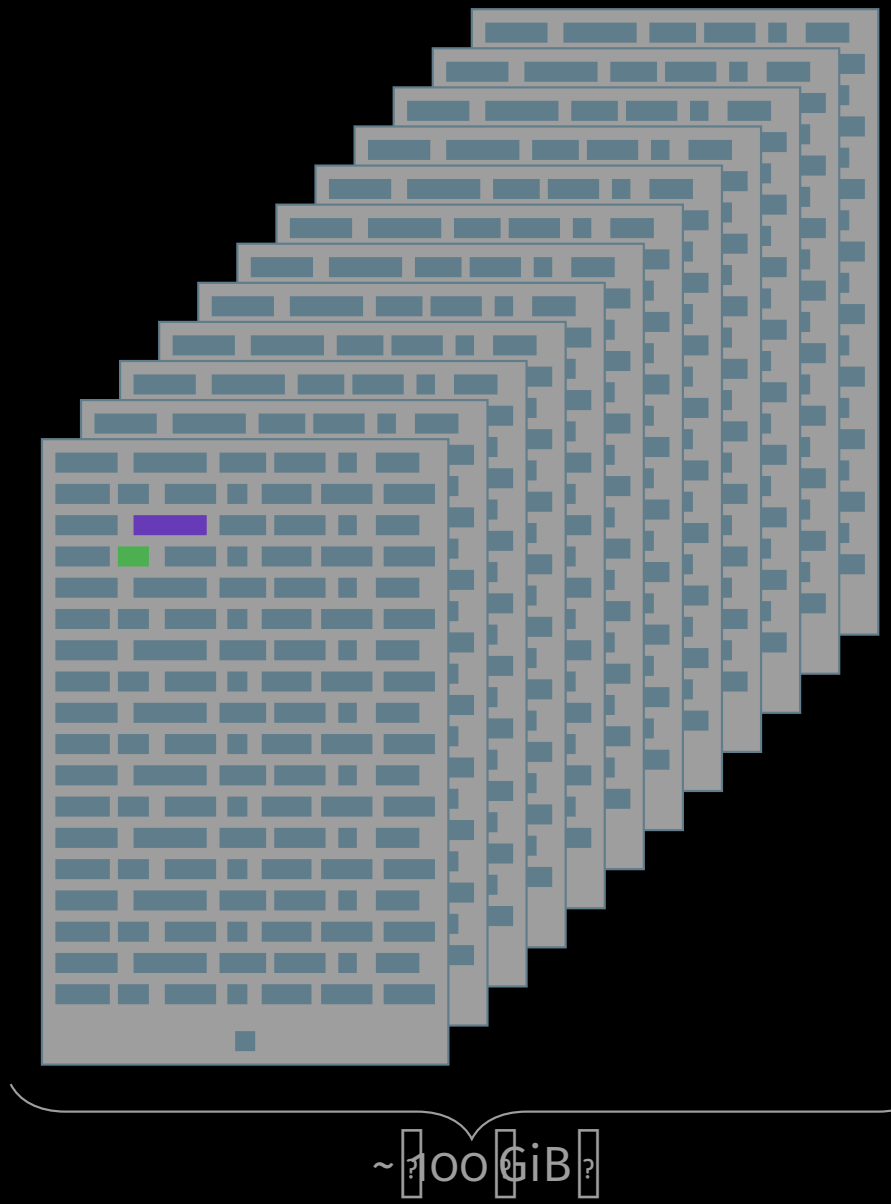




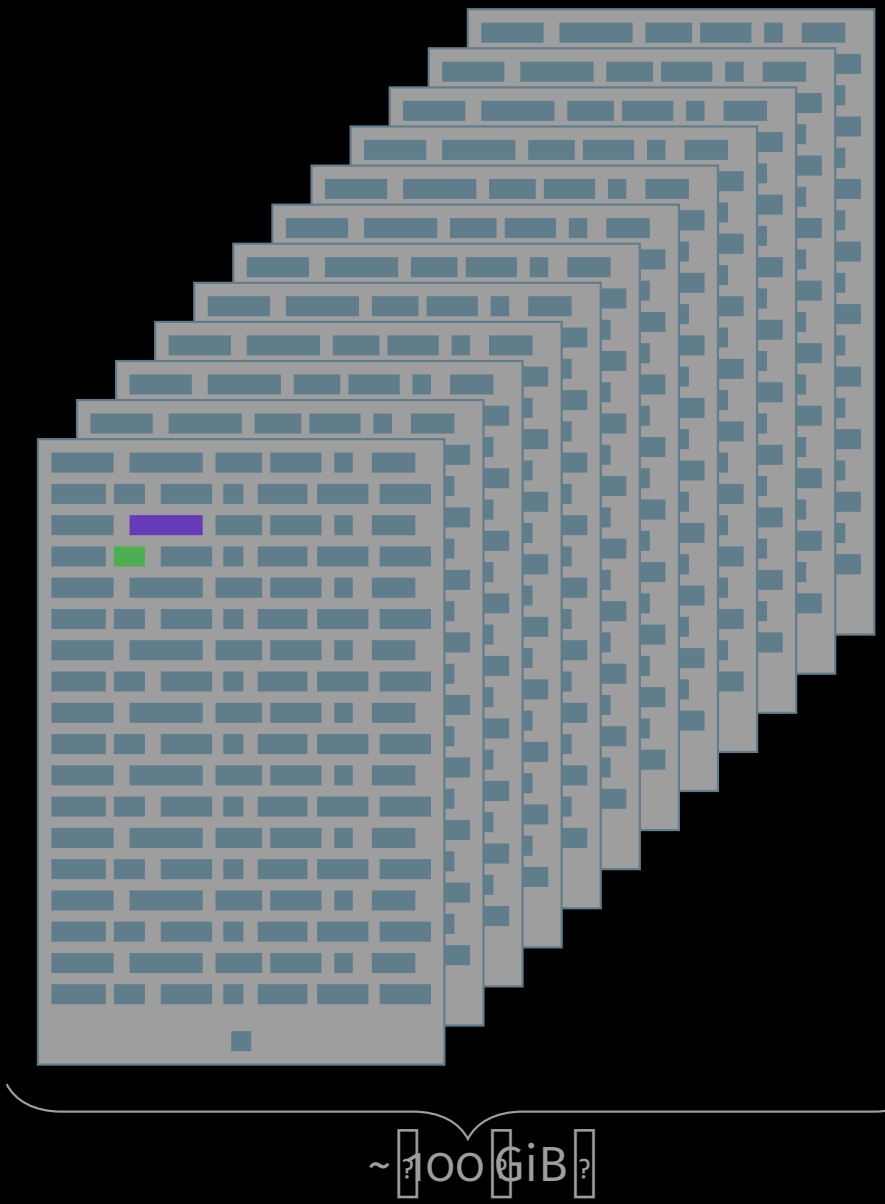
TEXT INDEX

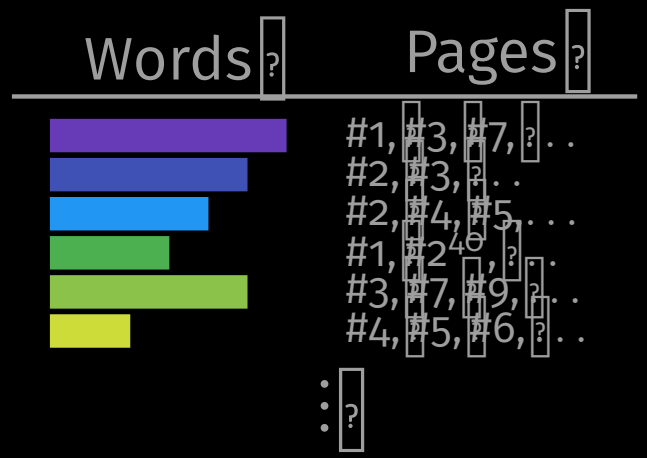


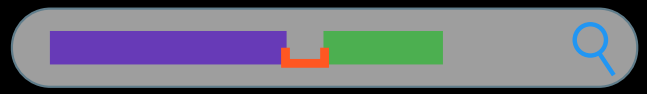
TEXT INDEX



TEXT INDEX



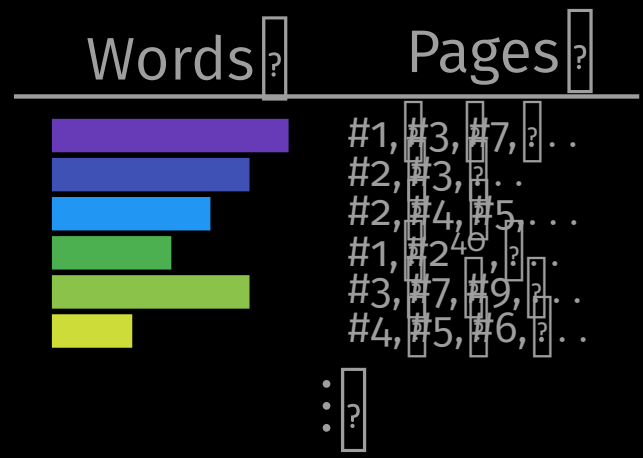


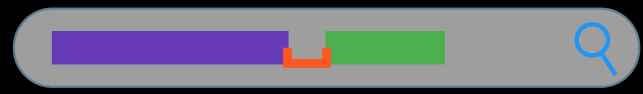


. . . but what if?

TEXT INDEX

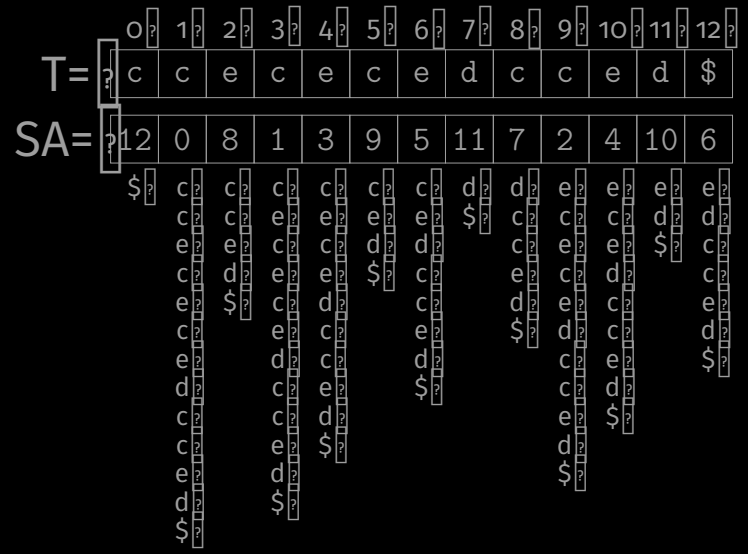
~100 GiB

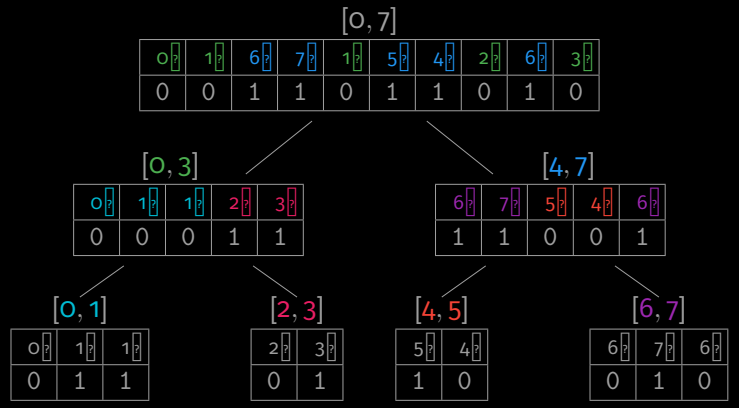




. . . but what if?

SUFFIX ARRAYS & WAVELET TREES



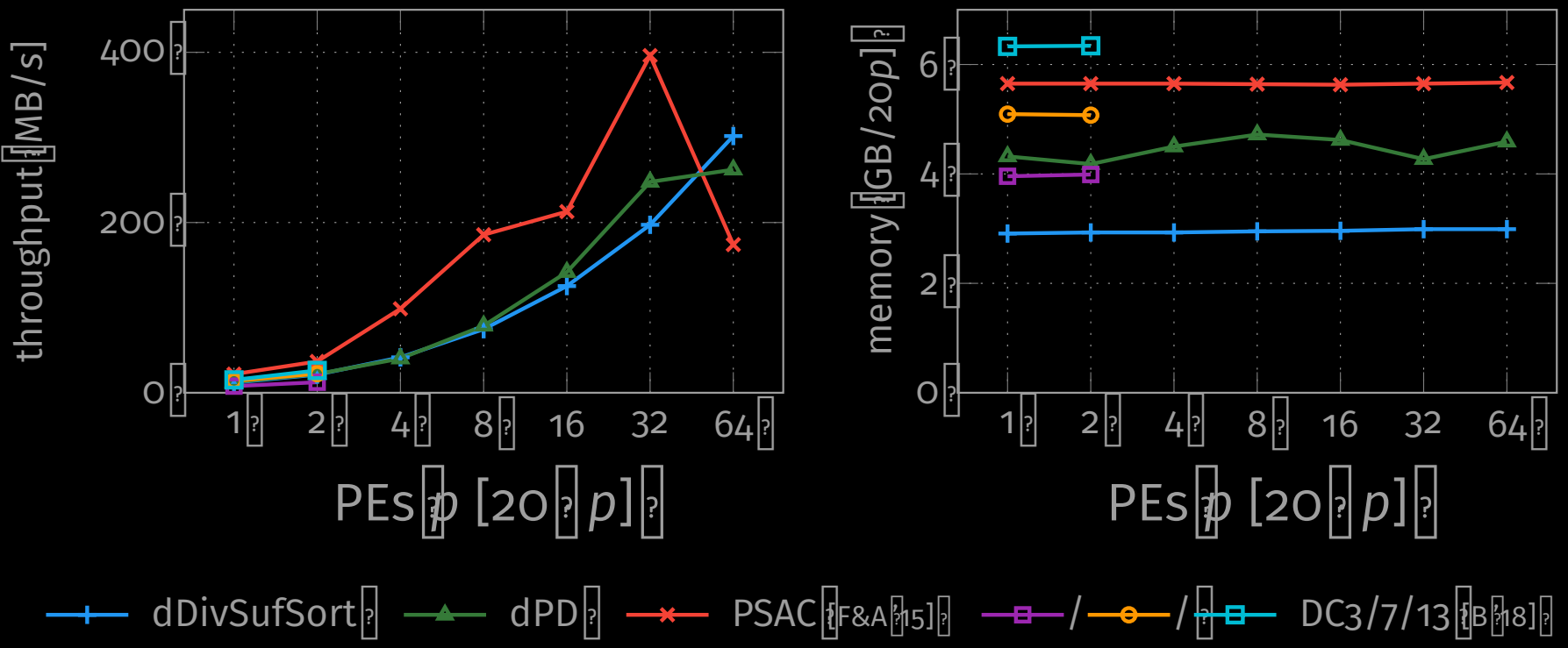


- Sorting suffixes
- Lots of data dependencies
- /kurpicz/dsss

- Working on bits
- Sorting is too much
- /pdinklag/distwt

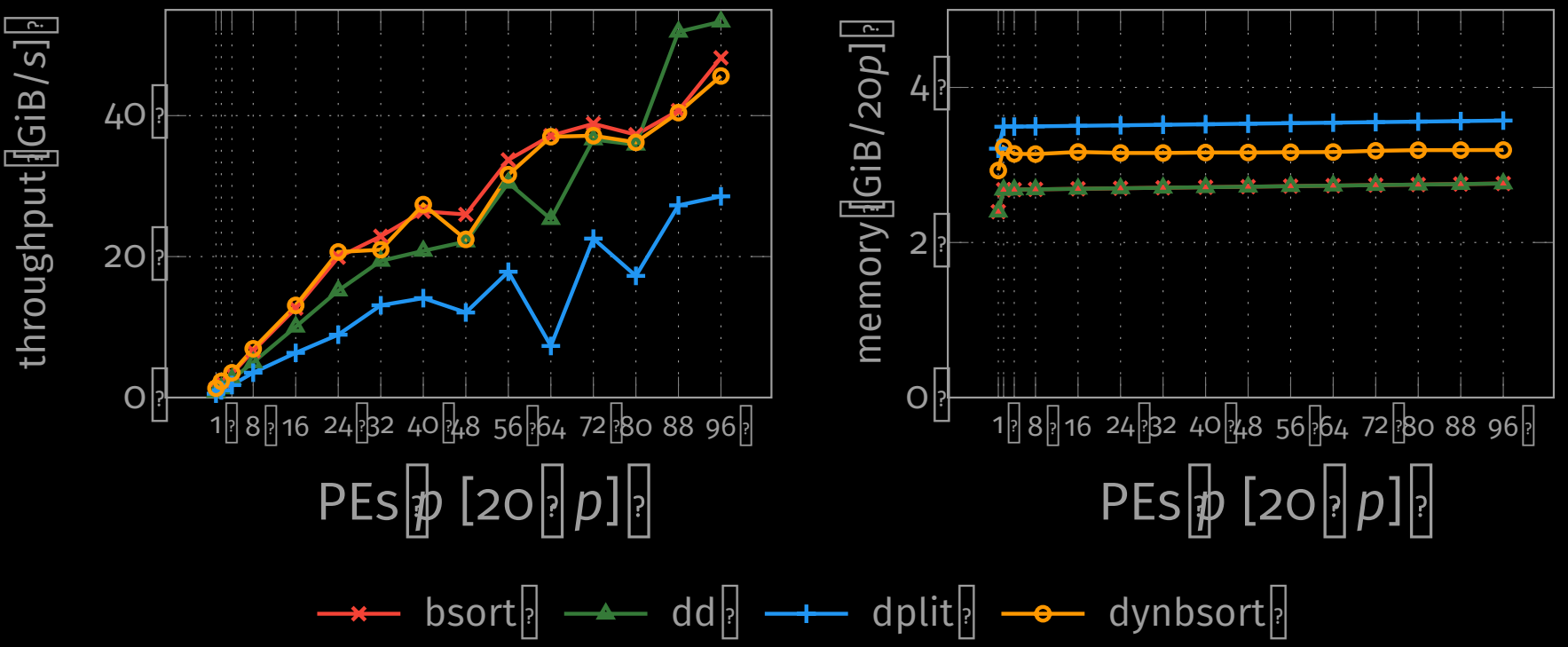
(EVALUATING THE) CONSTRUCTION

- Evaluated suffix array construction on the LiDO3 cluster
- Weak scaling experiments (only **DNA**) with 90 MB per PE



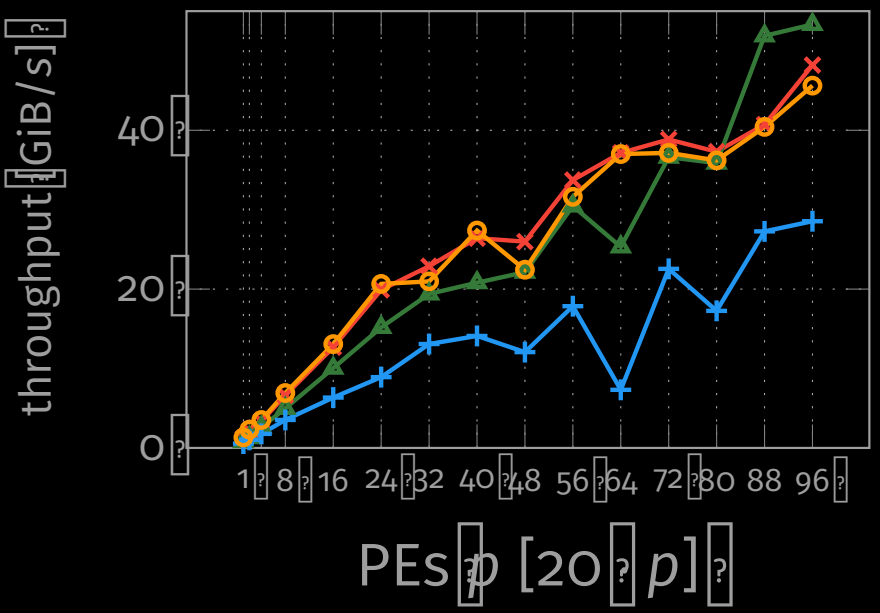
(EVALUATING THE) CONSTRUCTION

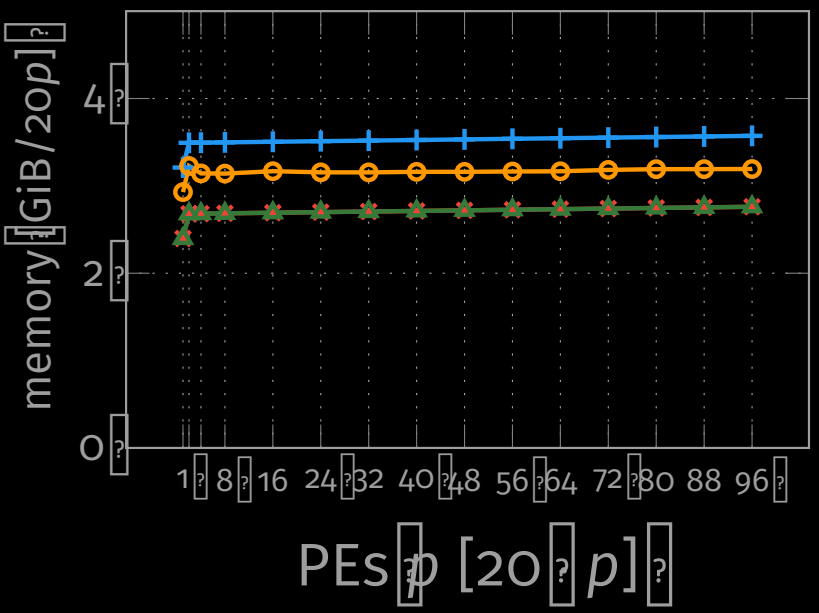
- Evaluated wavelet tree construction on the LiDO3 cluster
- Weak scaling experiments (only **DNA**) with 53 MB per PE



(EVALUATING THE) CONSTRUCTION

- Evaluated wavelet tree construction on the LiDO3 cluster
- Weak scaling experiments (only **DNA**) with 53 MB per PE





bsort dd dplit dynbsort

Thank You