# Geriatrix

## Aging what you see, and what you don't see

Saurabh Kadekodi[+], Vaishnavh Nagarajan[+], Gregory R. Ganger[+] and Garth A. Gibson[+*]

[+]Parallel Data Laboratory, [*]Vector Institute
Carnegie Mellon University

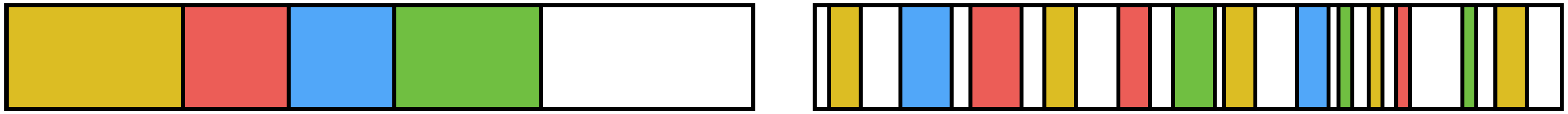
# In a nutshell

- **File system aging *__still__* matters**
  - Recreated published experiments w/ aging
  - Aging is even more important on SSDs

- ***Geriatrix* — a file system aging suite**
  - Induces adequate file & free space fragmentation
  - Profile driven with 8 built-in aging profiles

# Why study file system aging?

- FS performance can deteriorate with prolonged usage, mainly due to fragmentation

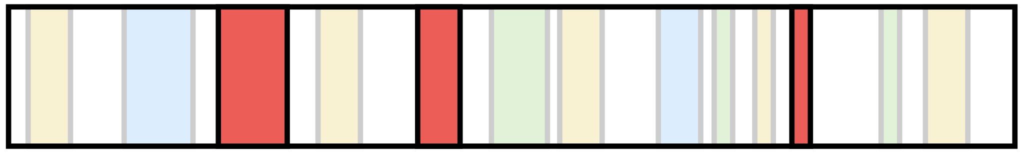


**Fresh or Defragged** **Aged**

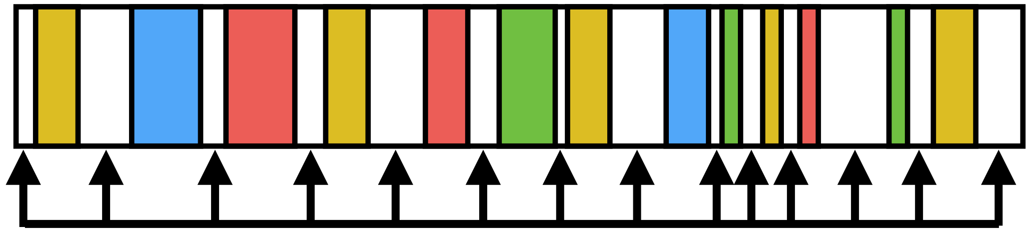
- Responsible FS benchmarking ***must*** include aging
  - Shown by Keith Smith & Margo Seltzer in 1997
- Despite evidence, aging and its effects are largely ignored
  - **13 of 20** file system papers fail to mention aging

# Fragmentation

- Aging produces two kinds of fragmentation:
  - File fragmentation



    - File readback causes long seeks
  - Free space fragmentation



    - Writing file causes long seeks
    - Leads to file fragmentation
- Current aging tools only focus on file fragmentation

Carnegie Mellon
Parallel Data Laboratory

# Part 1

## File System aging ***still*** matters

SSDs can perform worse than HDDs after aging

# Recreated experiments

- Recreated experiments from three publications:
- **Btrfs ACM TOS 2013 (HDD and SSD)**
  - HDD: 100GB aged image with 80% fullness
  - SSD: 60GB aged image with 70% fullness
- **F2fs USENIX FAST 2015 (SSD)**
- **NOVA USENIX FAST 2016 (NVM)**

# Recreated experiments

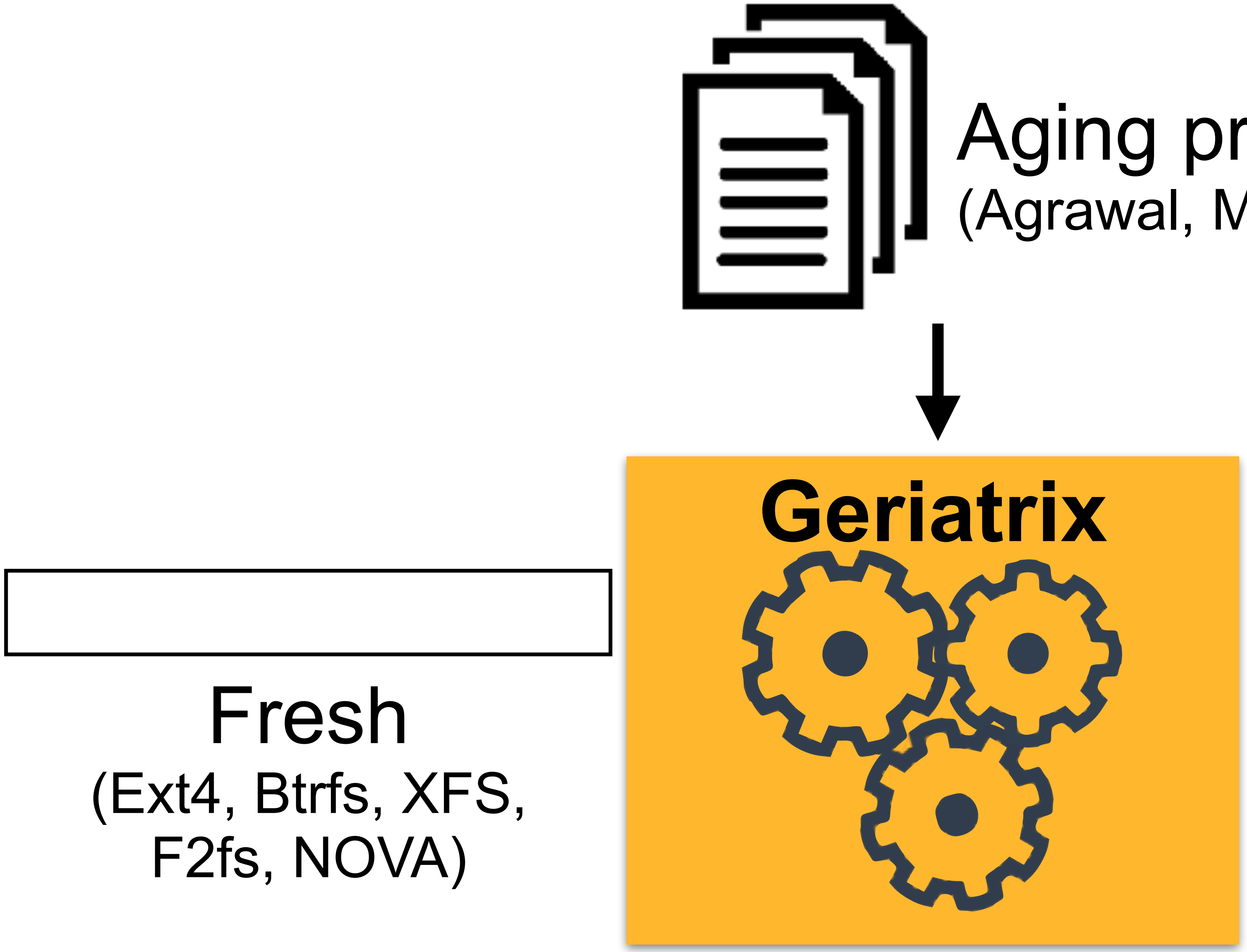
- Recreated experiments from three publications:

- **Btrfs ACM TOS 2013 (HDD and SSD)**
  - HDD: 100GB aged image with 80% fullness
  - SSD: 60GB aged image with 70% fullness

- **F2fs USENIX FAST 2015 (SSD)**

- **NOVA USENIX FAST 2016 (NVM)**

# Benchmarking with Geriatrix



Aging profile

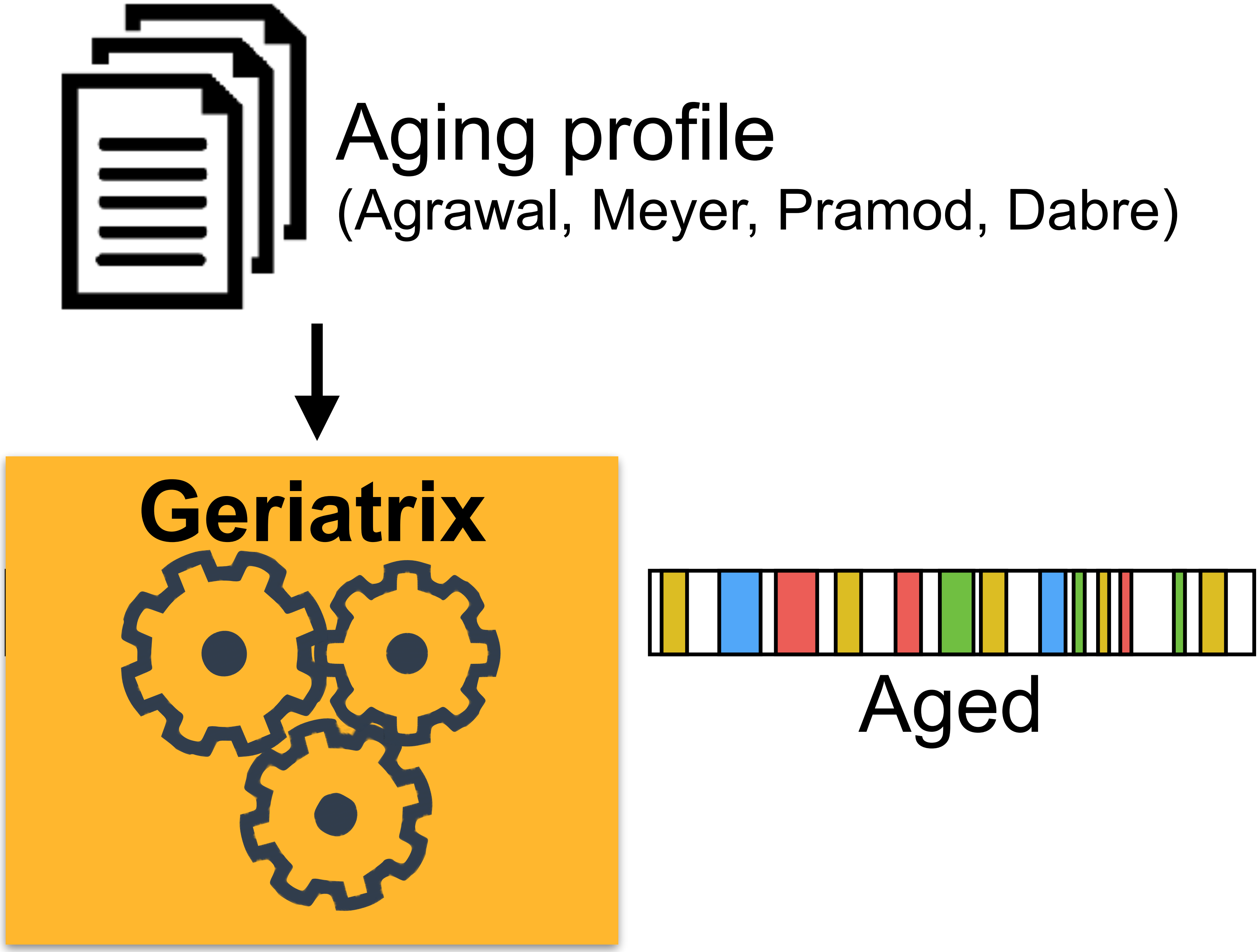
(Agrawal, Meyer, Pramod, Dabre)

Geriatrix

Fresh

(Ext4, Btrfs, XFS, F2fs, NOVA)

# Benchmarking with Geriatrix



Aging profile

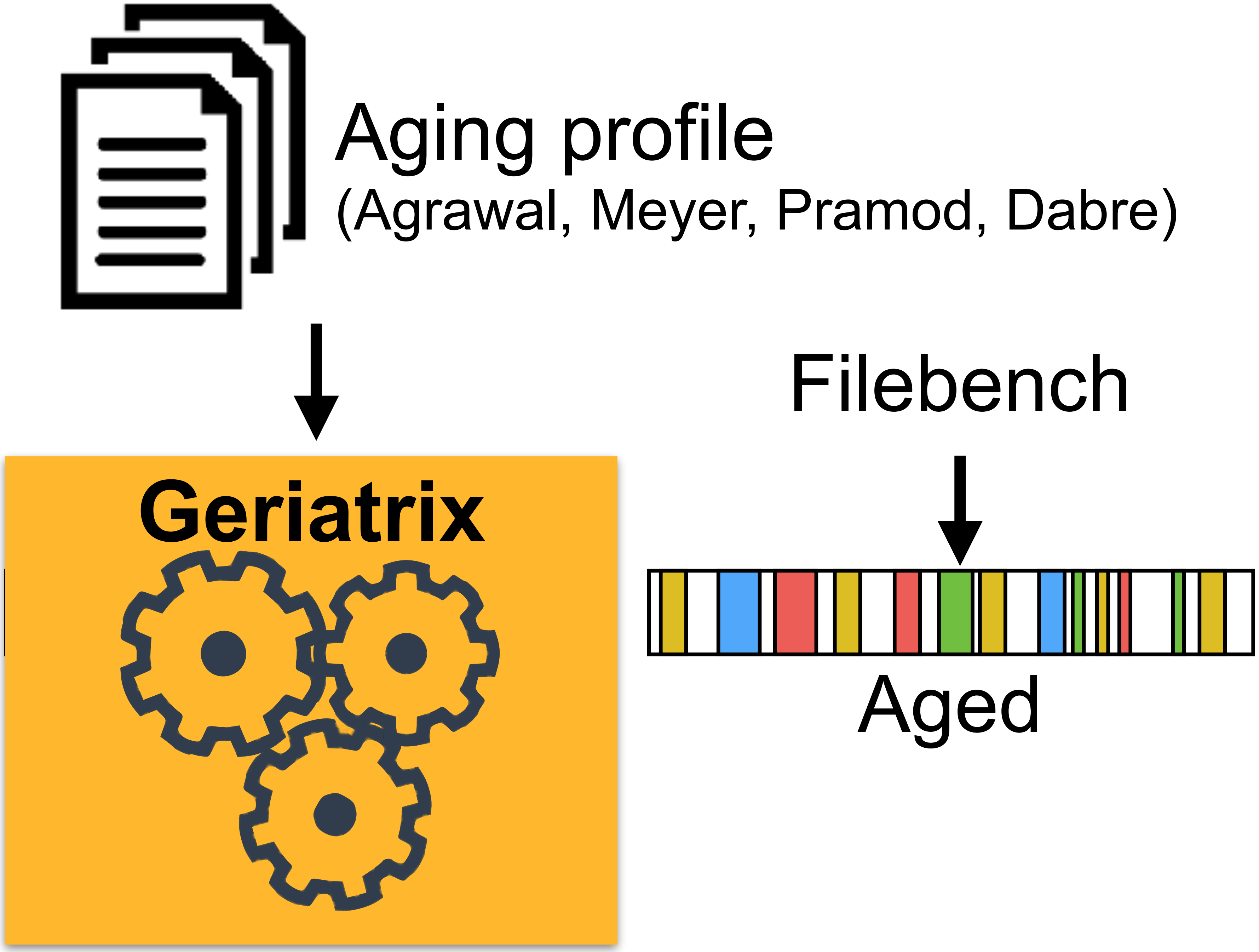
(Agrawal, Meyer, Pramod, Dabre)

Geriatrix

Aged

controlled sequence
of file creates / deletes

# Benchmarking with Geriatrix



Aging profile

(Agrawal, Meyer, Pramod, Dabre)

Geriatrix

Filebench

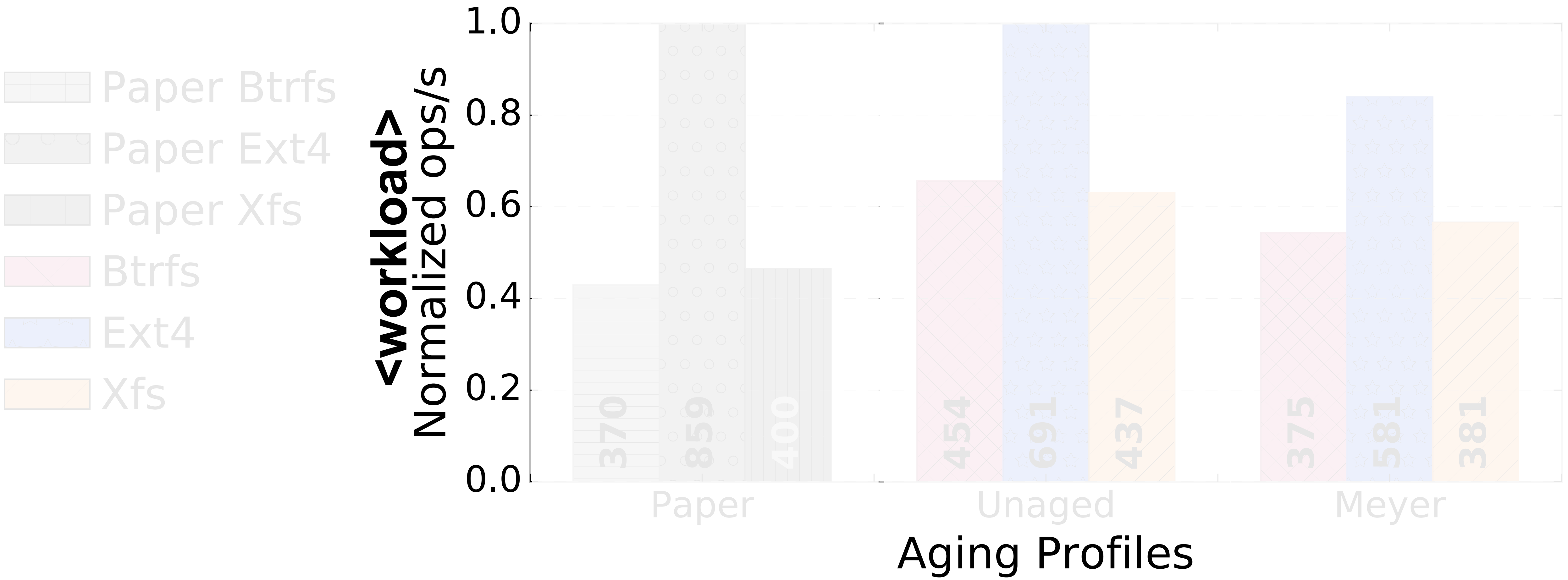
Aged

controlled sequence
of file creates / deletes

# Benchmarking configuration

- All recreations are Filebench benchmark reruns
- Each benchmark run lasted 10 min
- Three runs of each benchmark for variance
  - Error bars not shown since RMSE < 0.01%
- Throughput in **ops / s** as shown by Filebench
- All results are normalized to Ext4 performance

# Evaluation Format

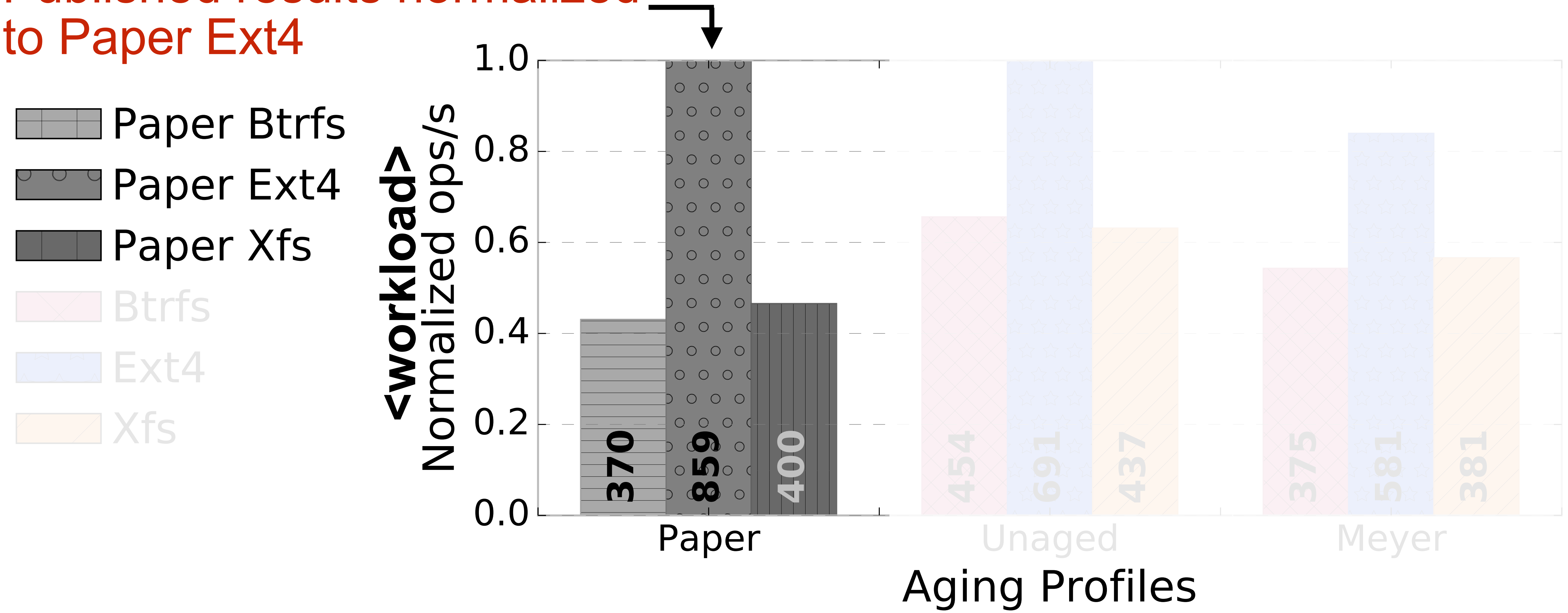


- The filebench workload used to benchmark the FS

Carnegie Mellon
Parallel Data Laboratory

# Evaluation Format

Published results normalized to Paper Ext4

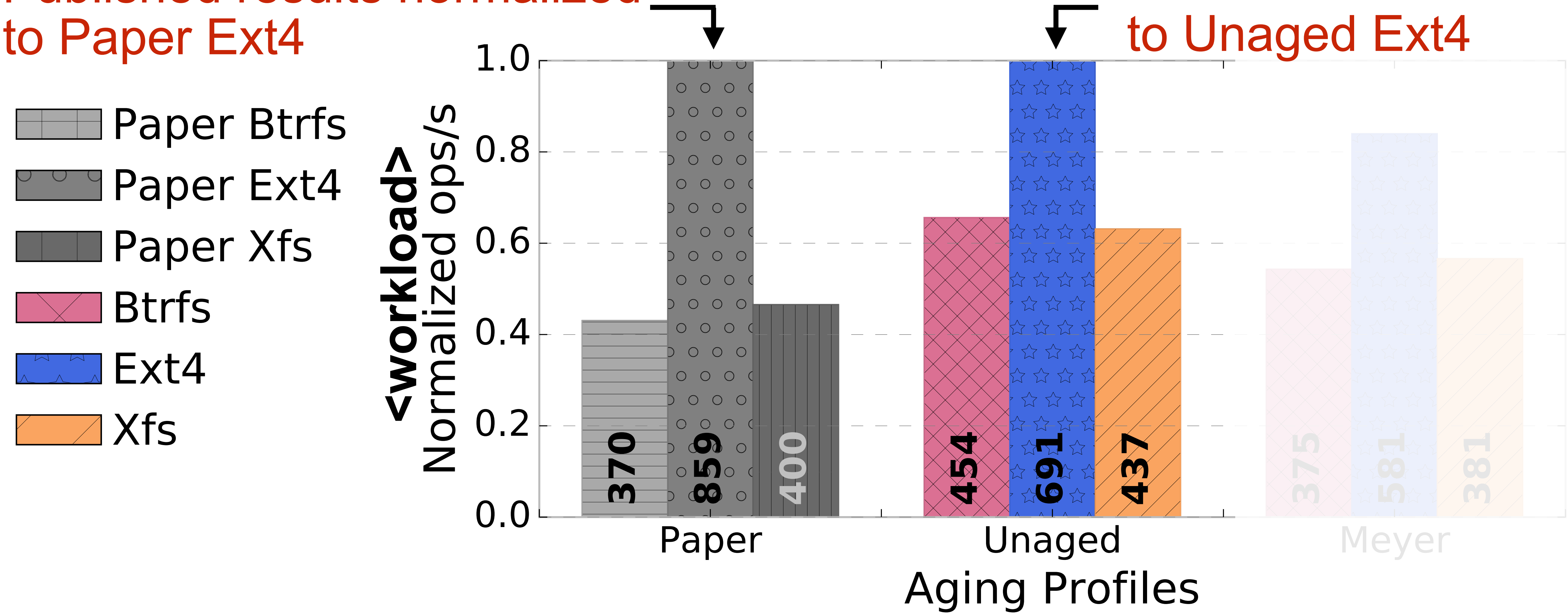


- The filebench workload used to benchmark the FS
- Published results (with raw performance numbers on the bar)

# Evaluation Format

Published results normalized to Paper Ext4

Our results normalized to Unaged Ext4

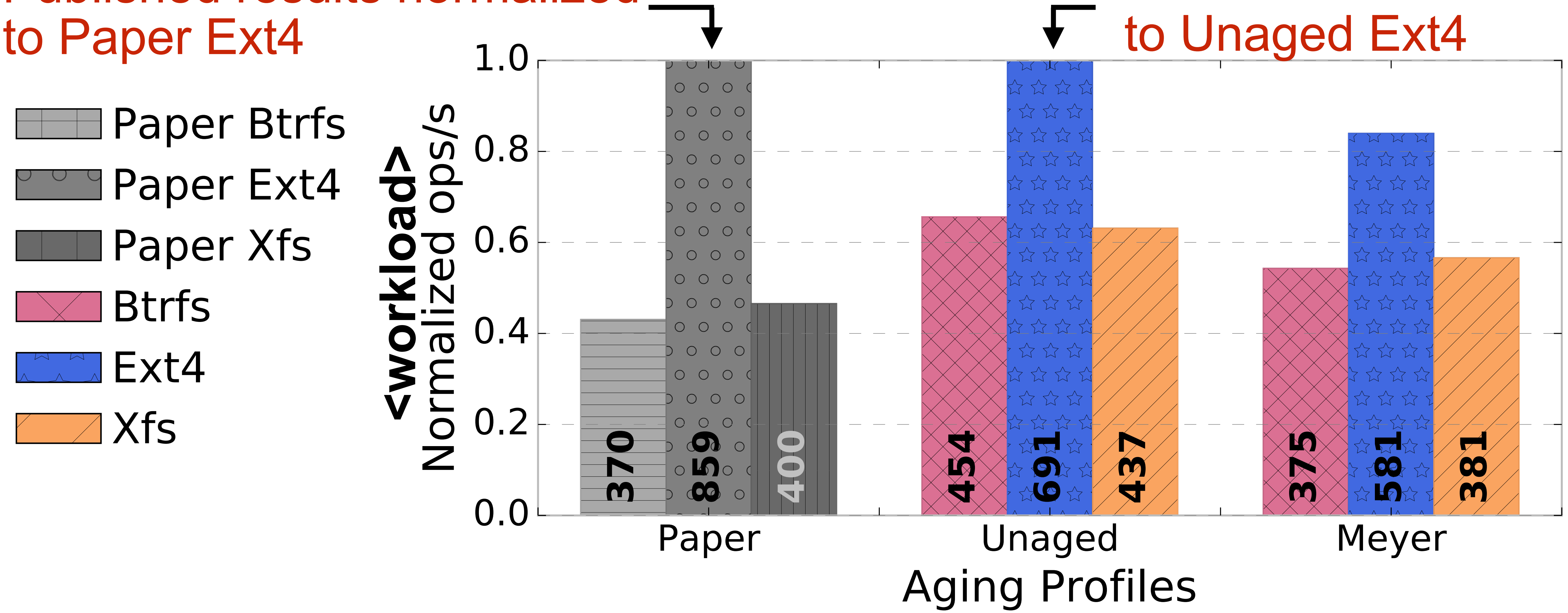


- The filebench workload used to benchmark the FS
- Published results (with raw performance numbers on the bar)
- Performance of unaged FS on our h/w using the publication config

# Evaluation Format

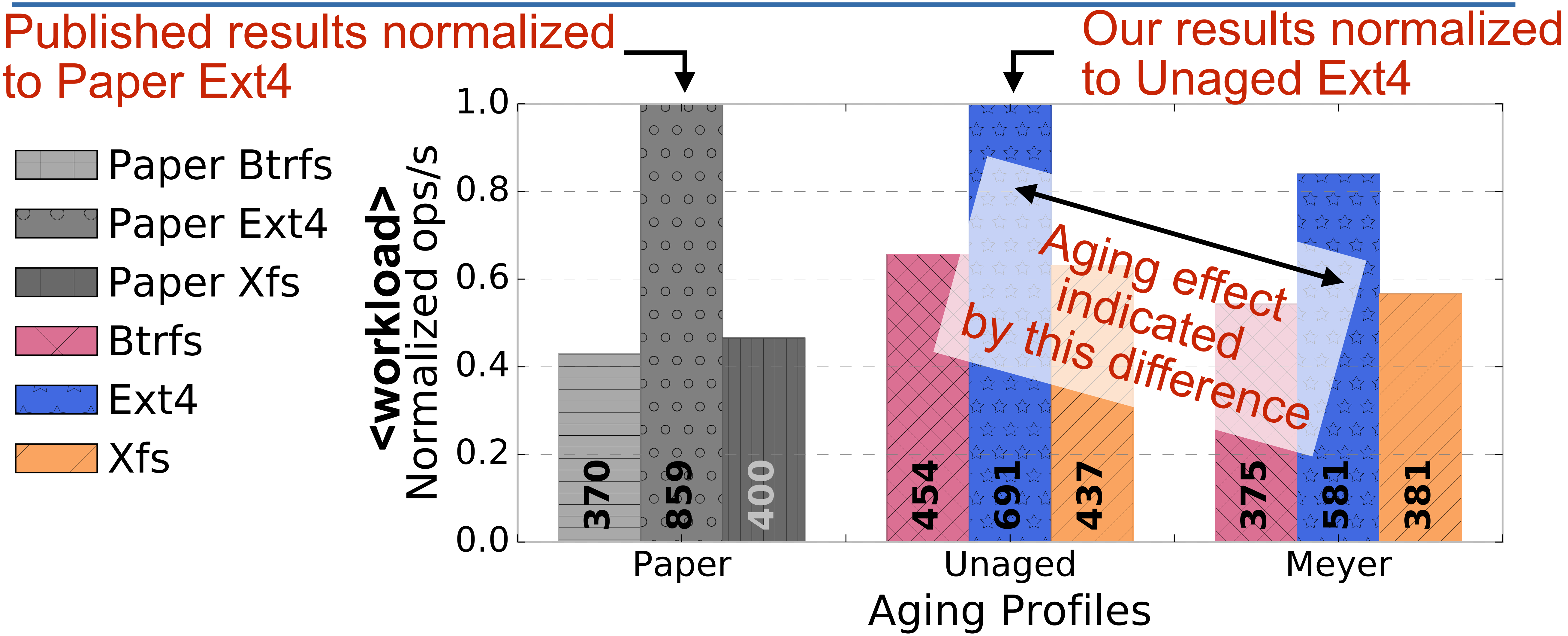
Published results normalized to Paper Ext4

Our results normalized to Unaged Ext4



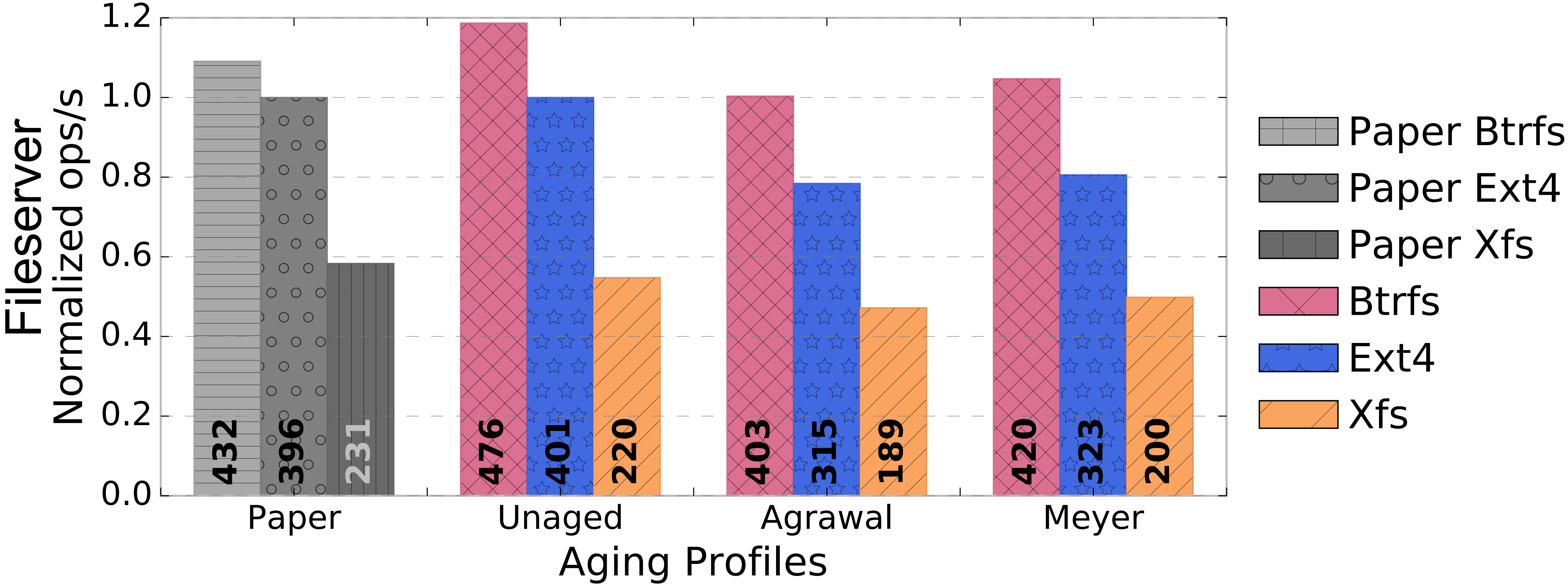
- The filebench workload used to benchmark the FS
- Published results (with raw performance numbers on the bar)
- Performance of unaged FS on our h/w using the publication config
- Performance of FS aged using indicated aging profile on our h/w

# Evaluation Format



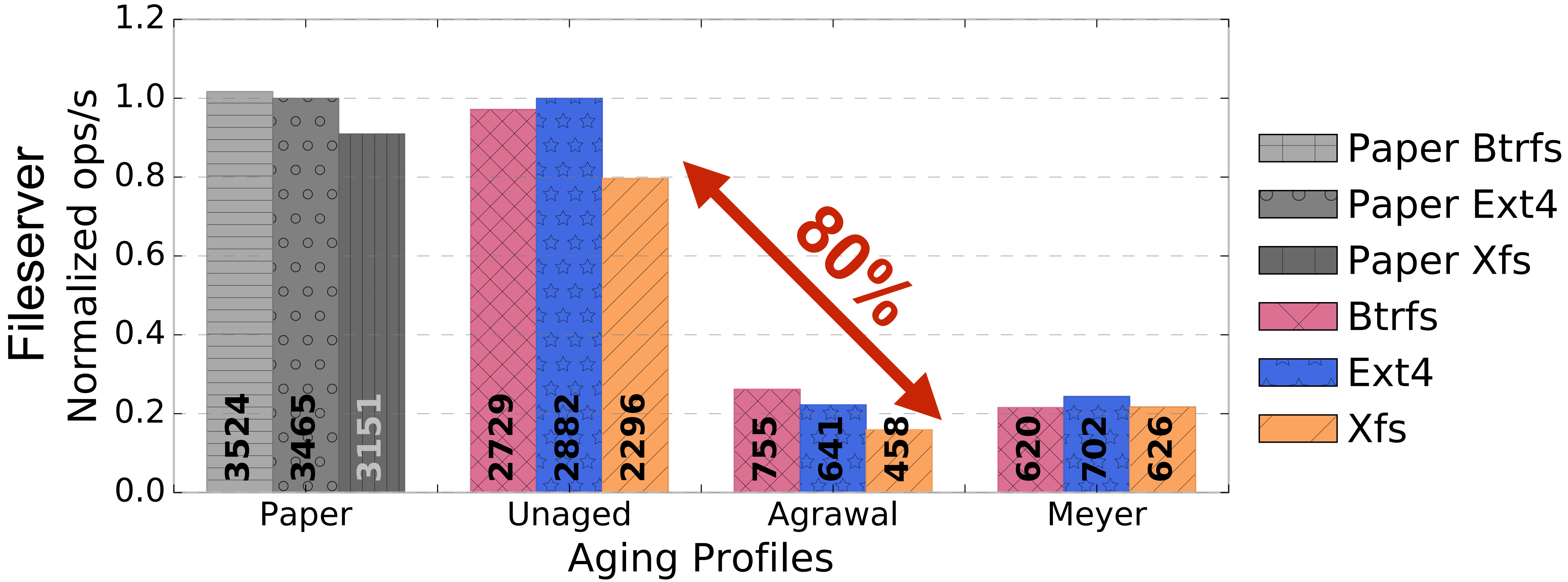
- The filebench workload used to benchmark the FS
- Published results (with raw performance numbers on the bar)
- Performance of unaged FS on our h/w using the publication config
- Performance of FS aged using indicated aging profile on our h/w

# Btrfs 2013 HDD Recreation



- 16-22% difference before and after aging
- Geriatrix also acts as stress tester
- As Smith and Seltzer said — we must pay attention to FS aging

# Btrfs 2013 SSD Recreation



- Rank ordering completely different from publication
- Different aging profiles result in different performance ranking
- SSD ages along with file system — exaggerated by free space fragmentation

Carnegie Mellon
Parallel Data Laboratory

# Other experiments (F2fs, NOVA)

- **F2fs USENIX FAST 2015 (SSD)**
  - Different SSDs — both across and within classes age very differently
- **NOVA USENIX FAST 2016 (NVM)**
  - Aged NOVA shows little throughput reduction (upto 6%)
  - Aged tail latencies are much more affected than throughput
  - For very low-latency FSes, tail latency slowdown is commentary on FS design
- Both recreations show different rank ordering of FSes compared to publication

# Part 2

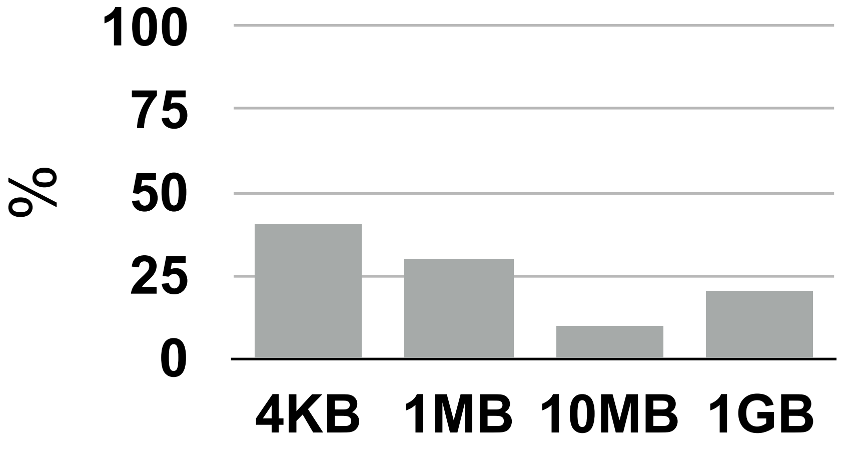
## *Geriatrix* — The aging suite

# Geriatrix aging process

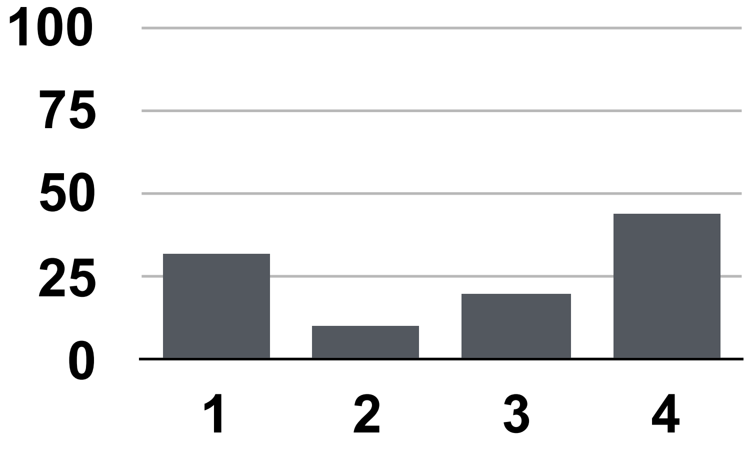
Aging Profile

1. File System Fullness (bytes, %)

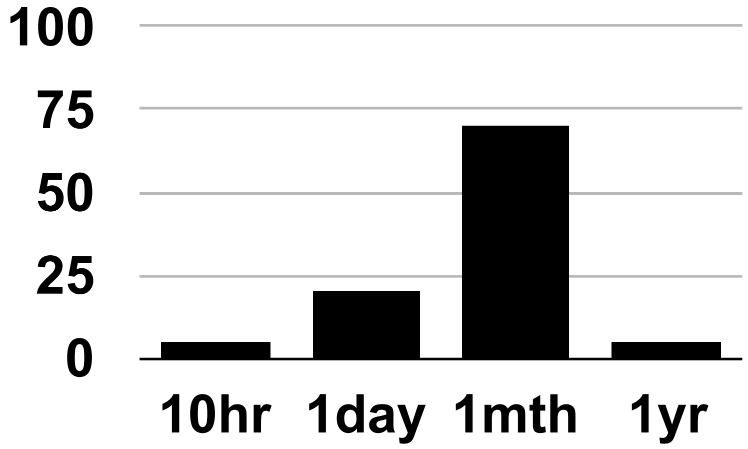
2. File Size Distr.



3. Dir Depth Distr.

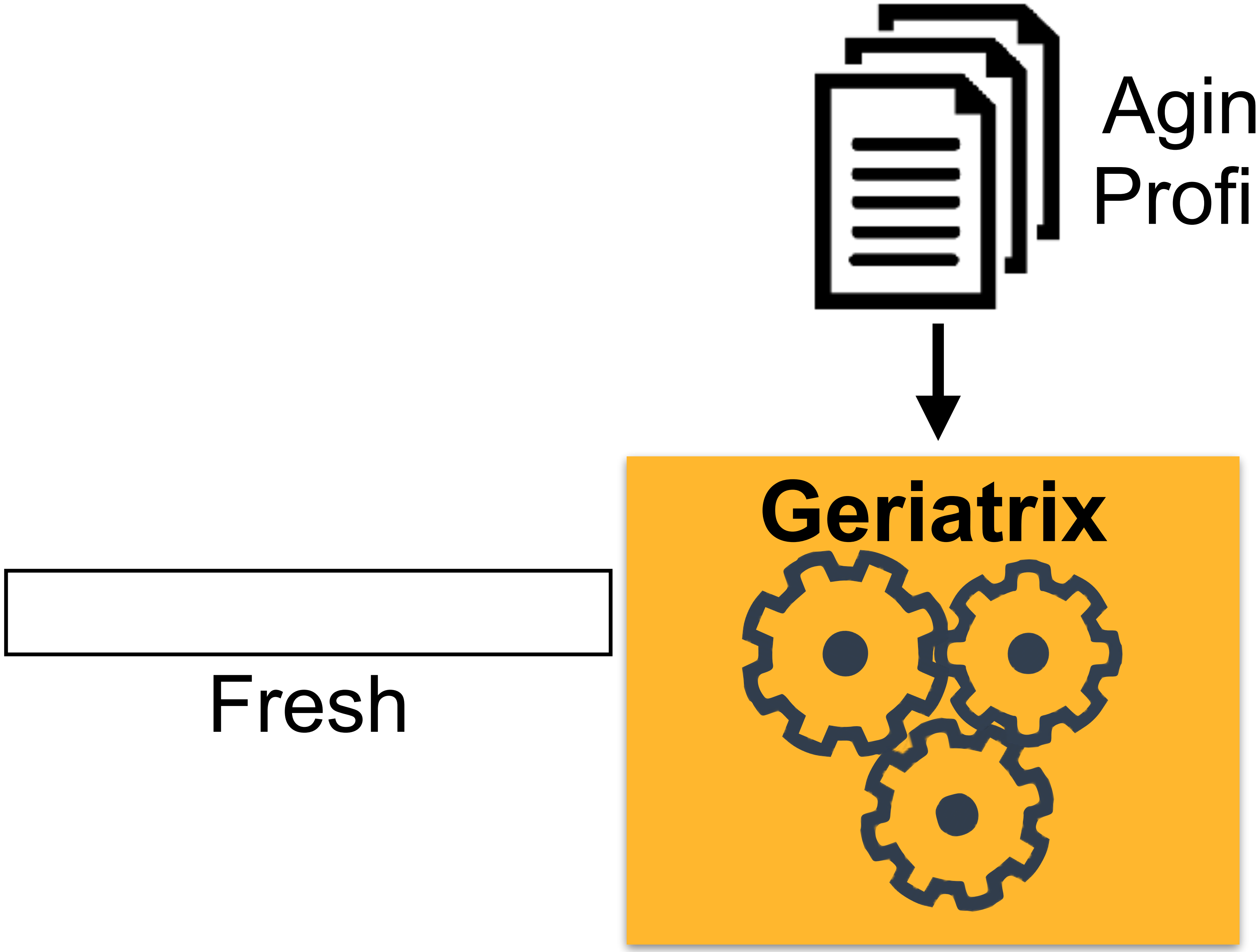


4. Relative Age Distr.



Aging profiles easily measured by a simple FS tree walk

# Geriatrix aging process

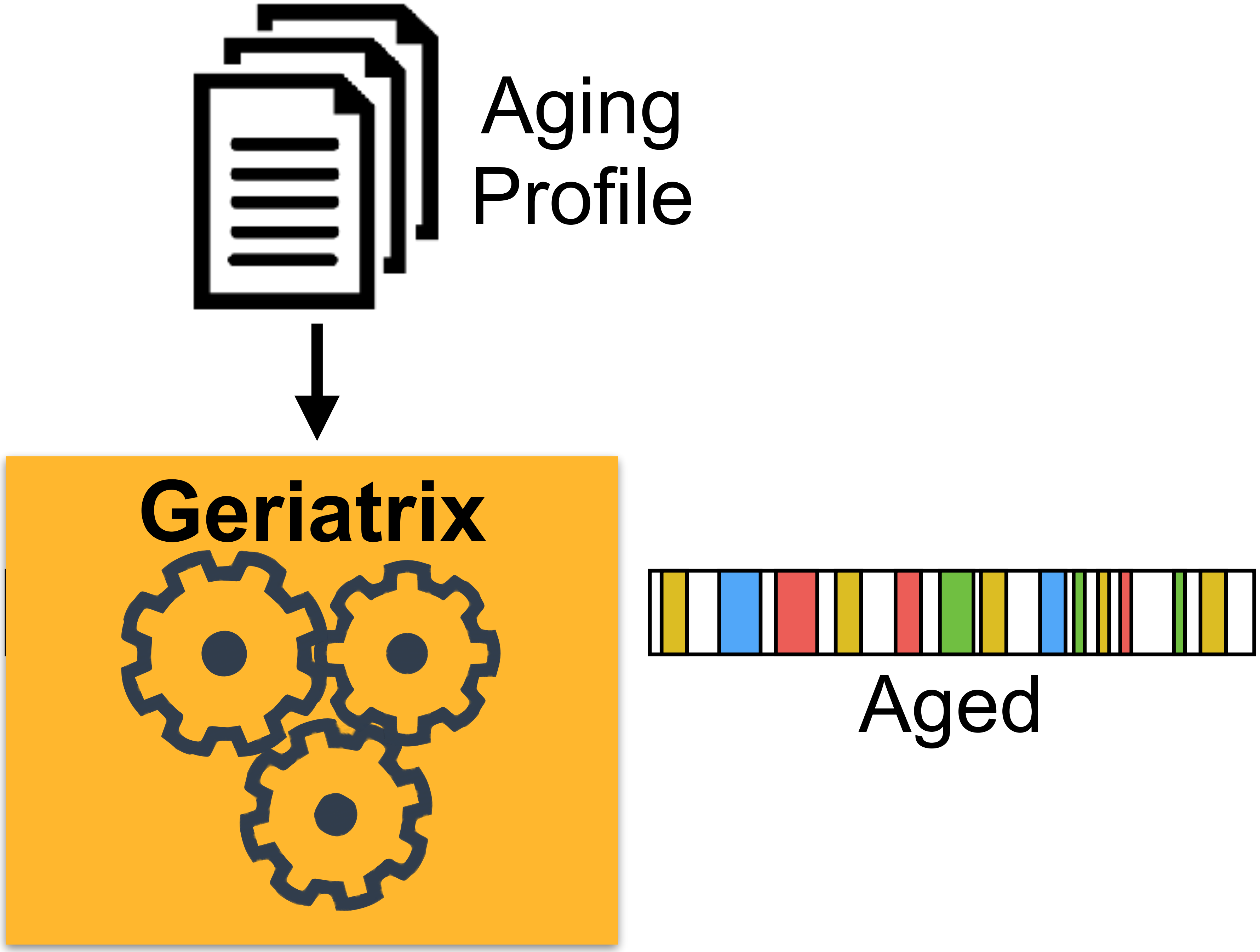


Aging Profile

Geriatrix

Fresh

# Geriatrix aging process



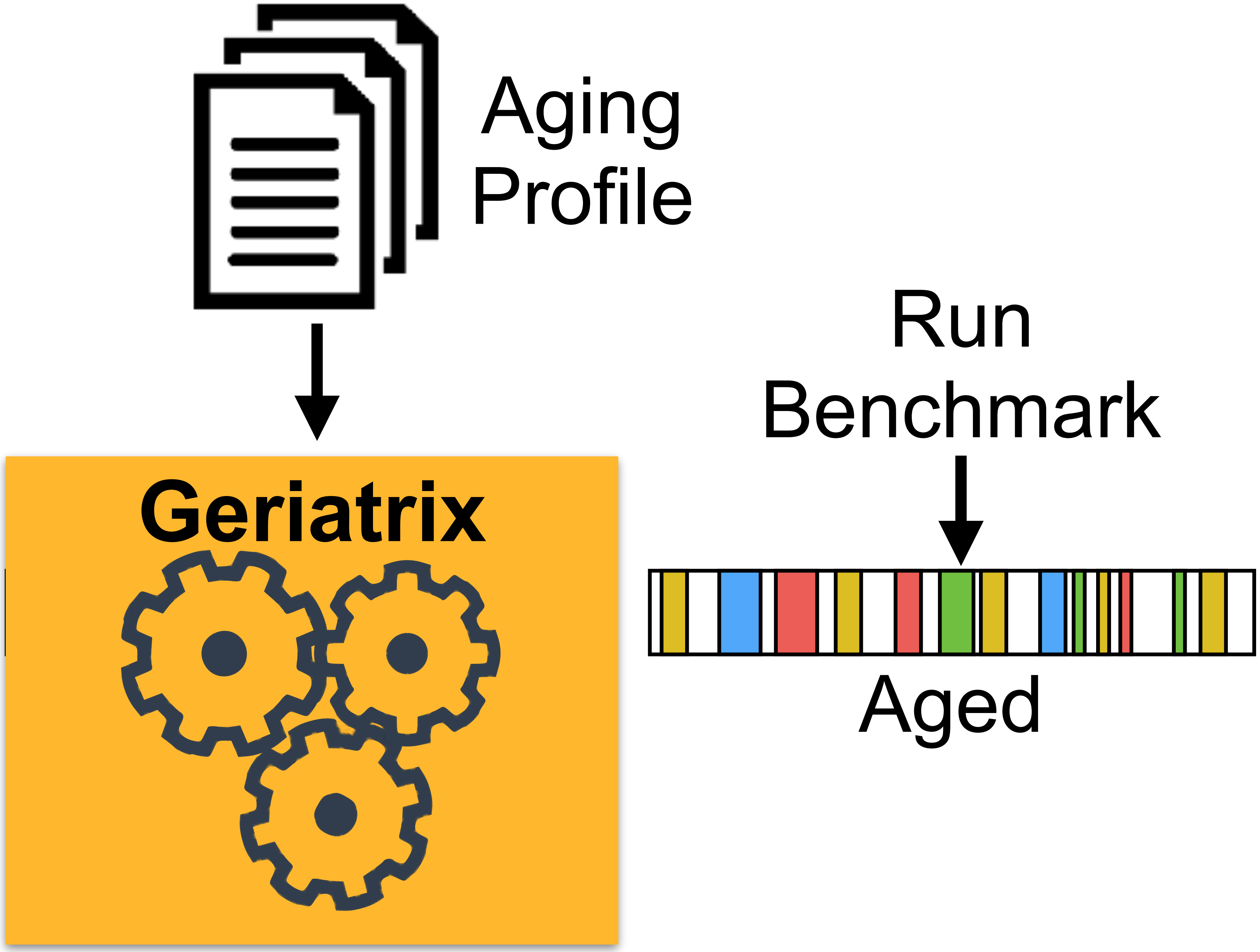
Aging Profile

Geriatrix

Aged

controlled sequence of file creates / deletes

# Geriatrix aging process



controlled sequence of file creates / deletes

# Geriatrix aging methodology

1. **Rapid aging**
   - Only file creates (aim is to achieve fullness %)
   - Continuously maintaining size & dir depth distrs.
2. **Stable aging**
   - File creates and deletes w/ fair coin tosses
     - to maintain fullness %
   - Continuously maintaining size & dir depth distrs.
   - Aim is to achieve relative age distribution

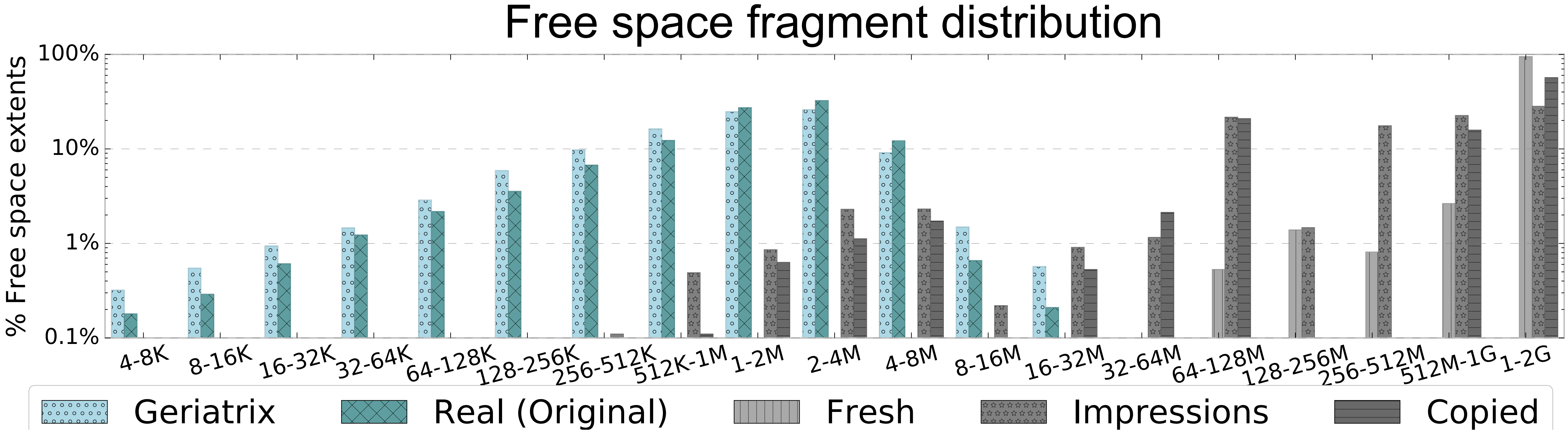
# Is Geriatrix accurate?

# Is Geriatrix accurate?

- Captured an aging profile from a colleague's HDD

# Is Geriatrix accurate?

- Captured an aging profile from a colleague’s HDD
  - Grundman aging profile included with Geriatrix



- Copying is similar to freshly fragmented
- Impressions has only large free space extents
- Geriatrix mimics original free space fragmentation

# How costly is Geriatrix?

- 50GB XFS image aged in memory w/ Geriatrix using 32 threads

| Aging profile | Age (yrs) | Workload (TB) | Duration (hrs) |
|---|---|---|---|
| Meyer | 2 | 7.8 | 1.3 |
| Wang-LANL | 11 | 1.4 | 2.4 |
| Agrawal | 14 | 12 | 7.8 |
| Wang-OS | 22 | 1.7 | 3.9 |

# How costly is Geriatrix?

- 50GB XFS image aged in memory w/ Geriatrix using 32 threads

| Aging profile | Age (yrs) | Workload (TB) | Duration (hrs) |
|---|---|---|---|
| Meyer | 2 | 7.8 | 1.3 |
| Wang-OS | 22 | 1.7 | 3.9 |

In-memory aging done in hrs

# Geriatrix — The tool

- Developed in C++
- Has 8 built-in aging profiles to standardize aging
    - 3 regular usage laptop workloads,
    - 2 desktop workloads,
    - 1 deduplication workload,
    - 1 OS archive (CMU datacenter)
    - 1 HPC workload
- Multi-threaded for faster aging
- Uses `fallocate` to avoid writing data

# Conclusion

- Responsible FS benchmarking ***must*** include aging
  - FS aging exists and continues to be ignored
  - Aging effects sometimes more dramatic on SSDs
- ***Geriatrix*** - an efficient, profile driven and reproducible aging suite that simplifies FS aging
  - Induces adequate file and free space fragmentation

bit.ly/geriatrix-code

Contributions encouraged

saukad@cs.cmu.edu