

May 23, 2025 Maison des Mines et des Ponts, Paris, France



FABME

File-Level Placement Policy with Control Admission, Burst Prefetching, and Multi-Criteria Eviction for HPC Multi-Tier Storage

Hocine Mahni[†] Stéphane Rubini^{*} Sébastien Gougeaud[‡] Philippe Deniel[‡] Jalil Boukhobza[†]

[†]Lab-STICC, CNRS UMR 6285, ENSTA, Institut Polytechnique de Paris, 29806 Brest, France, [‡]CEA, Bruyères-le-Châtel, France ^{*}Univ. Brest, Lab-STICC, CNRS, UMR 6285, Brest, France



Outline

- Context
- Problem statement
- Background
- Motivations
- Challenges & Contributions
- Conclusion and future work



Context & Research problem

- The global data volume will reach 394 zettabytes in 2028
- Exascale computing may widen the gap between computation, main memory and storage
- Exploiting multi-tier and heterogeneous storage systems is a key to reach trade-off between performance, cost, and capacity (Tab 1, Fig 1)
- Data placement and migration between tiers is automated by hierarchical storage management (HSM)
- HSM tool like **Robinhood** manage data at **file granularity**, so **blockcaching** algorithms cannot be used
- MC-ARC file-level policy outperforms block caches strategies



Technology	Latency	Throughput (R/W)	IOPS	Capacity	Cost (\$/GB)
SSD (NVMe)	$\sim 20 \mu s$	$8.0/5.0{ m GB/s}$	1.2M	$< 32 \mathrm{TB}$	0.200
SSD	$\sim 100 \mu { m s}$	2.1/2.0 GB/s	0.8M	$< 8 \mathrm{TB}$	0.100
HDD	$\sim 10\mathrm{ms}$	250/240 MB/s	< 500	$< 14 \mathrm{TB}$	0.030
Tape	$> 20 \mathrm{s}$	315/315 MB/s	-	$< 15 \mathrm{TB}$	0.001



Fig. 1: Hierarchical Storage Systems (left), with HSM at the file granularity (right)

Research problem: How to design an extension to MC-ARC that integrates admission and prefetching mechanisms to optimize multi-tier storage management while preserving fairness among users?



Background

Multicriteria File-Level Placement Policy for HPC Storage

Hocine Mahni ¹, Stéphane Rubini³, Sébatien Gougeaud², Philippe Deniel², Jalil Boukhobza¹ ¹ ENSTA Bretagne, Lab-STICC, CNRS, UMR 6285, Brest, France ² CEA, Bruyères-le-Châtel, France ³ Univ. Brest, Lab-STICC, CNRS, UMR 6285, Brest, France {hocine.mahni,jalil.boukhobza}@ensta.fr,{sebastien.gougeaud,philippe.deniel}@cea.fr,{stephane.rubini}@univ-brest.fr

- MC-ARC designed to decide which files should reside in the high-performance tier of a multi-level system.
- MC-ARC evaluates each file according to three complementary scores:
- A1 Usage: to prioritize files with a large proportion of blocks accessed recently and frequently
- A2 Remaining lifetime : to prefer files predicted to stay relevant longer, so that long-lived files remain on the high-performance tier
- A3 User fairness : to penalize users who al-ready consume a disproportionate share of the SSD, achieving fair resource sharing
- MC-ARC can use either WSM or TOPSIS to merge the three scores



Motivations

Using MC-ARC simulator, replaying the YOMBO trace with configuration : (SSD size = 0.1-5% of workload) shows that:

- M1: 42% of files occupy 98% of SSD yet causing 93.5% of evictions \rightarrow cache pollution
- M2: 3.9 × SSD shuffled, taking 90% of execution time \rightarrow large-file thrashing
- M3: Bursty and periodic file profiles detected (see Fig. 2)



Parameter	Value		
Dominant frequency f _{dom}	0.02813 Hz		
Dominant period $T_{dom} = 1/f_{dom}$	35.55 s		
Peak amplitude A _{dom}	1551.77		
Peak z-score	7.19		
Threshold used	z > 3 (3-sigma rule)		

Fig. 2: Period detection on the YOMBO workload using FFT and Z-score



9th Per3S, Paris, France

Hocine Mahni & al.

Challenges & Contributions

Challenges : Summarize the motivations (M1, M2, M3)

- C1: Cache pollution caused by very large or rarely used files
- C2 Alternating large files overflows SSD tier, triggering thrashing
- C3 Files characterized by short, periodic I/O bursts tend to persist in the upper-tier cache well beyond the decay of their temporal locality, resulting in sub-optimal utilization of high-performance capacity

Applications



 \blacksquare Admission control: on each miss, the policy decides to admit the file to the SSD or confine it to HDD, according to its admission rules \rightarrow C1 C2

2 Burst prefetch: a lightweight detector anticipates bursts and prefetches the file just before the burst starts \rightarrow C3

§ MC-ARC updated: To manage evictions from the highperformance tier, we enhance MC-ARC to delay the eviction of files with bursty access patterns until their bursts are completed $\rightarrow C3$





FABME: FILE-LEVEL PLACEMENT POLICY WITH CONTROL ADMISSION, BURST PREFETCHING, AND MULTI-CRITERIA EVICTION FOR HPC MULTI-TIER STORAGE

Hocine Mahni[†], Stéphane Rubini[‡], Sebastien Gougeaud^{*}, Philippe Deniel^{*}, Jalil Boukhobza[†]

 $^\dagger Lab-STICC,$ CNRS UMR 6285 , ENSTA, Institut Polytechnique de Paris, 29806 Brest, France, $^\pm Univ.$ Brest, Lab-STICC, CNRS, UMR 6285, Brest, France, *CEA, Bruyères-le-Châtel, France

$1-\mathrm{HPC}$ Data Placement on Heterogeneous and Multilevel Storage System



Thank You !

I invite you to discuss it In front of my poster

Contacts:



{sebastien.gougeaud,philippe.deniel}@cea.fr

[‡]{stephane.rubini}@univ-brest.fr[‡]

